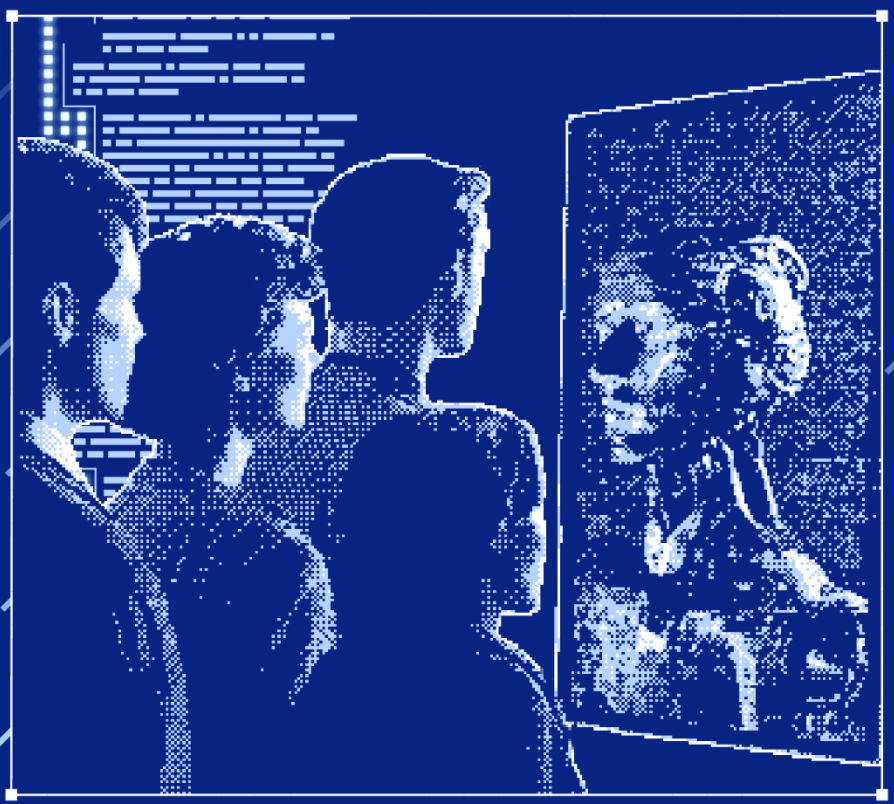
Conseil
de l'IA
et du Numérique

Rapport de la Commission « IA générative et vulnérabilités »

Protéger sans empêcher : nos vulnérabilités à l'ère de l'IA générative

Novembre 2026



COMPOSITION DE LA COMMISSION

Coprésidents de la Commission :

- Amine BENYAMINA, Professeur de psychiatrie et d'addictologie à l'Université Pais-Saclay ;
- Raphaël GAILLARD, Professeur de psychiatrie à l'Université Paris Cité ;
- Serena VILLATA, Directrice de recherche CNRS en Intelligence artificielle au Centre INRIA d'Université Côte d'Azur.

Membres de la Commission :

- Sylvie CHOKRON, Neuropsychologue, directrice de recherche au CNRS ;
- Marie COHEN-SKALLI, Codirectrice d'Emmaüs Connect ;
- Marc FADDOUL, Cofondateur d'AI Forensics ;
- Juliette SÉNÉCHAL, Professeure de droit privé et sciences criminelles, adjointe au directeur scientifique d'Inria (Université de Lille / Inria) ;
- Joëlle TOLEDANO, Professeure émérite en économie, associée à la Chaire « Gouvernance et Régulation » de l'Université Paris Dauphine ;
- Célia ZOLYNSKI, Professeure de droit privé à l'Université Paris 1 Panthéon-Sorbonne.

Secrétariat général du Conseil :

- Arthur BARICHARD, Directeur du CIANum ;
- Inès COLOMB, Rapporteure au CIANum ;
- Joséphine CORCORAL, Directrice adjointe du CIANum ;
- Romane GUECHOT, Responsable de la communication et de la presse ;
- Esther SAUREL, Rapporteure au CIANum.

Avec la participation de Paul FAVIER, Doctorant à la Chaire « Gouvernance et Régulation » de l'Université Paris Dauphine.

SOMMAIRE

COMPOSITION DE LA COMMISSION 3

SOMMAIRE 4

RÉSUMÉ OPÉRATIONNEL 9

Introduction 20

Éléments clés 23

CHAPITRE 1 25

Autonomie cognitive, liens affectifs et violences en ligne : l'IA générative, miroir déformant de nos vulnérabilités individuelles 25

A) L'IA générative altère l'autonomie cognitive et décisionnelle des individus 25

1) Les usages de l'IA générative peuvent être à la fois bénéfiques et porteurs de risques, en particulier pour les mineurs 26

a) Des bénéfices éducatifs dépendant de la conception des outils d'IA générative et de leur cadre d'utilisation 26

b) Des risques de délégation cognitive documentés et particulièrement accrus pour les mineurs 29

2) Quand l'IA choisit pour nous : les utilisateurs sont exposés à des risques de perte d'autonomie et de recul de l'esprit critique 32

a) Des choix de plus en plus délégués aux outils d'IA générative 32

b) Des utilisateurs soumis à l'opacité décisionnelle de l'IA générative 34

B) Engagement émotionnel et risques psychosociaux : l'IA générative redéfinit les liens entre les humains et les machines 35

1) L'IA générative nourrit un effacement de la frontière entre humains et machines 36

a) Anthropomorphisme et flagornerie : de puissants leviers d'engagement 36

b) Les utilisateurs face à la complaisance de l'IA générative 39

c) Des risques pour la santé mentale accrus par les choix de conception 39

d) Des risques exacerbés pour les plus vulnérables, notamment les mineurs 42

e) Jouets IA : nouveau terrain de vulnérabilités 44

2) L'interaction prolongée entre humains et intelligences artificielles expose les utilisateurs à des vulnérabilités émotionnelles 47

a) Attachement émotionnel généré par l'IA générative : une projection unilatérale 48

b) Un risque d'isolement social nourri par des relations inauthentiques 50

3) L'extraction de données à un niveau inédit et l'hyperpersonnalisation de l'IA générative sont préjudiciables 52
a) Une collecte de données d'une ampleur et d'une granularité inédites 52
b) Le droit au respect de la vie privée menacé par la collecte massive de données 53
C) L'IA générative amplifie les violences numériques : *deepfakes*, fraudes et haine en ligne 56
1) La démocratisation des *deepfakes* est un facteur d'amplification des risques, notamment pour les plus jeunes 56
a) Les femmes et les filles, premières victimes de la prolifération des *deepfakes* à caractère sexuel non consentis 57
b) L'explosion sans précédent des contenus pédocriminels synthétiques 59
c) Usurpations d'identité et escroqueries : l'usage des *deepfakes* dans le champ de la fraude en ligne 61
2) L'IA générative peut se muer en support de la haine en ligne 63
CHAPITRE 2 66
Cohésion sociale et exercice démocratique : la confiance collective à l'épreuve de l'IA générative 66
A) De la reproduction à l'accélération des préjugés : les biais de l'IA générative altèrent la cohésion sociale 67
1) De sa conception à son utilisation, l'IA générative comporte des biais 67
2) L'IA générative renforce les stéréotypes préexistants et engendre des discriminations concrètes 69
3) L'IA générative peut porter atteinte aux diversités culturelles et linguistiques 71
a) Les effets d'une sous-représentation de la diversité linguistique 71
b) Vers une uniformisation culturelle et linguistique ? 72
B) Nouvel intermédiaire de l'information, l'IA générative recompose l'espace public 74
1) L'IA générative offre de nouveaux modes d'accès à l'information 74
a) Un intermédiaire déjà incontournable dans l'accès à l'information 74
b) Les potentialités de l'IA générative pour une information fiable et de qualité 76
c) Vers la profusion de contenus informationnels 77
2) L'IA générative déstabilise l'espace informationnel 77
a) L'attrition de l'économie des médias traditionnels 77
b) L'IA générative au service de la manipulation de l'information à grande échelle 80
3) L'IA générative met au défi le pluralisme et la liberté d'expression 87
C) L'IA générative contribue à l'érosion de la confiance entre l'État et les citoyens 88

1) L'IA générative aggrave plus qu'elle ne résout l'exclusion numérique 88

2) La dématérialisation des services publics est un facteur supplémentaire d'érosion de la confiance 90

a) L'IA dans les services publics : une dématérialisation porteuse d'exclusion et de défiance 90

b) La fragilisation du lien de confiance collective 91

3) En période électorale, l'IA offre des outils pour exploiter les vulnérabilités démocratiques 93

a) Influence et délégation des choix électoraux 93

b) Un risque d'homogénéisation de l'offre politique 93

CHAPITRE 3 97

Encadrer les modèles économiques et renforcer le cadre juridique : prévenir l'exploitation des vulnérabilités individuelles et collectives 97

A) L'économie de la cognition induit un changement d'échelle dans l'exploitation des vulnérabilités 99

1) Les modèles économiques de l'IA générative sont encore en construction 100

2) Le passage de l'économie de l'attention vers l'économie de la cognition doit être anticipé dès à présent 102

a) L'économie de la cognition, nouvelle étape de la monétisation des données 102

b) Le modèle publicitaire de l'IA générative : une dynamique à surveiller 104

B) Le cadre juridique actuel appréhende de manière partielle et inadaptée les vulnérabilités induites par l'IA générative 108

1) Le cadre juridique peine à appréhender les vulnérabilités suscitées par l'IA générative 108

a) Un mille-feuille d'instruments réglementaires mal coordonnés entre eux 108

b) Des capacités d'évaluation trop tardives, intervenant principalement après la mise sur le marché 110

c) Des instruments juridiques inadaptés aux spécificités de l'IA générative 111

2) Le droit et sa résilience doivent être renforcés face aux risques de vulnérabilités induits par l'IA générative 112

a) Accroître les obligations s'imposant aux fournisseurs d'IA générative 112

b) Évaluer les systèmes d'IA générative avant et pendant leur commercialisation afin de mieux appréhender le spectre des vulnérabilités : changement de paradigme pour la régulation 117

c) Affermir la responsabilité civile des acteurs de l'IA générative 120

RECOMMANDATIONS PRIORITAIRES 123

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent et qui génèrent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement. 123

Recommandation n° 2 : Imposer un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux pour les systèmes d'IA générative accessibles aux mineurs. 127

Recommandation n° 3 : Imposer une évaluation pré-commercialisation des systèmes d'IA générative grand public. 130

Recommandation n° 4 : Imposer une évaluation post-commercialisation des systèmes d'IA générative grand public. 131

Recommandation n° 5 : Garantir une évaluation des systèmes d'IA générative grand public par un accès aux données d'usage, d'entraînement et de monétisation. 132

Recommandation n° 6 : Garantir une évaluation interdisciplinaire des vulnérabilités exploitées et engendrées par les systèmes d'IA générative. 135

Recommandation n° 7 : Doter l'État d'un protocole national de réponse aux crises informationnelles liées à l'IA générative, en définissant des indicateurs de gravité et des canaux d'échange rapide de données avec les fournisseurs d'IA générative. 136

Recommandation n° 8 : Interdire la publicité déguisée, les pratiques masquées d'influence de nature commerciale et toute forme de monétisation exploitant les vulnérabilités des consommateurs. 138

Recommandation n° 9 : Maintenir et développer une politique nationale ambitieuse d'éducation à l'IA permettant à chacun d'en maîtriser les usages et les risques. 140

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS 142

Recommandation n° 10 : Renforcer la répression de la création de contenus à caractère sexuel non consentis et pédocriminels truqués (hypertrucages) générés par intelligence artificielle. 142

Recommandation n° 11 : Instaurer une autorisation préalable de mise sur le marché pour les jouets intégrant un système d'IA générative et définir des seuils d'âge restrictifs d'accès. 144

Recommandation n° 12 : Consacrer un programme d'éducation à l'intelligence artificielle dès le primaire. 146

Recommandation n° 13 : Garantir dans les salles de classe des modèles d'IA générative spécifiquement adaptés pour les mineurs. 148

RECOMMANDATIONS POUR RENFORCER LA COHÉSION SOCIALE ET DÉMOCRATIQUE À L'ÈRE DE L'IA GÉNÉRATIVE 151

Recommandation n° 14 : Détecter de manière précoce les manipulations de l'information par les systèmes d'IA générative. 151

Recommandation n° 15 : Instaurer un code de bonne conduite sur l'usage de l'IA générative en période électorale. 153

Recommandation n° 16 : Garantir la représentativité des langues et des cultures minoritaires dans l'IA générative, en mobilisant les corpus et les catalogues de référence. 155

RECOMMANDATIONS POUR ACCROÎTRE LA TRANSPARENCE ET LA RESPONSABILITÉ DES ACTEURS DU MARCHÉ DE L'IA GÉNÉRATIVE 157

Recommandation n° 17 : Encadrer les inférences de données personnelles réalisées à partir des *prompts* de l'utilisateur. 157

Recommandation n° 18 : Renforcer la responsabilité des fournisseurs de systèmes d'IA générative. 159

Annexe 1 : Lettre de mission 161

Annexe 2 : L'appréhension de l'IA générative par le Règlement européen sur l'IA (RIA) 163

Annexe 3 : Les défis économiques de l'IA générative 170

Annexe 4 : Liste des auditionnés 183

Annexe 5 : Liste des contributions internationales sur les effets des usages de l'IA générative sur les mineurs 188

RÉSUMÉ OPÉRATIONNEL

Saisi le 20 février 2026 par la ministre déléguée chargée de l'Intelligence artificielle et du Numérique, Madame Anne Le Hénanff, le Conseil de l'intelligence artificielle et du numérique (CIANum) a établi la Commission « IA générative et vulnérabilités », composée de neuf experts pluridisciplinaires, dans le but de **poursuivre les réflexions engagées par le Président de la République sur les conséquences du numérique sur nos démocraties, le vivre-ensemble, avec un accent tout particulier sur la protection des mineurs en ligne**.

Le point de départ des travaux de cette Commission est un paradoxe. D'une part, l'IA générative a été adoptée très rapidement et par le plus grand nombre. Deux mois auront ainsi suffi à ChatGPT pour atteindre 100 millions d'utilisateurs actifs[1], contre deux ans et quatre mois pour le réseau social Instagram, pour ne citer que lui. D'autre part, la défiance à l'encontre de ces usages n'a cessé de croître. En France, 75 % des non-utilisateurs ont un avis bien tranché et expriment leur manque de confiance dans l'IA générative[2]. Dans ce contexte, la Commission a vite dressé le constat que **le débat entre les bénéfices potentiels de l'IA générative et les risques qu'elle suscite était encore très polarisé**.

Pourtant, les enjeux de l'IA générative sont marqués par une contradiction d'ordre temporel. D'un côté, les discours d'« urgence » tendent à se multiplier, qu'ils concernent la nécessité d'adopter cette technologie sous peine de « manquer le virage », de se prémunir des risques parfois qualifiés d'« existentiels » ou de se former à ces nouvelles fonctionnalités. De l'autre, force est de constater que le recul sur les impacts de l'IA générative et la vision à long terme sur ses enjeux, pourtant essentiels, font aujourd'hui cruellement défaut.

C'est dans ce contexte que la Commission a cartographié les vulnérabilités les plus critiques associées à l'émergence fulgurante de l'IA générative, avec **un objectif clair : les anticiper au mieux afin d'en atténuer la portée**. Au terme d'un cycle de 55 auditions conduites auprès de professionnels de santé, de la recherche publique et de l'enseignement supérieur, d'entreprises de l'écosystème, d'institutions publiques, d'autorités de régulation et d'instances consultatives, d'associations de protection de l'enfance et de représentants de la société civile, la Commission a décidé d'articuler ce rapport autour de trois idées structurantes.

(i) **En premier lieu, les systèmes d'IA générative présentent des risques de mieux en mieux documentés pour l'utilisateur**. Ces risques sont essentiellement de trois ordres :

Premièrement, **les études scientifiques initiales établissent un fort risque d'ordre cognitif pour les utilisateurs vis-à-vis des outils d'IA générative**, par la délégation de fonctions cérébrales essentielles (capacités de réflexion, d'expression, de prise de décision ou encore

[1] OpenAI. « Modèles d'utilisation et d'adoption de ChatGPT sur le lieu de travail ». 2026.
[2] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.

l'esprit critique). Les auteurs d'une étude du MIT[3] établissent la formation d'une « dette cognitive » croissante à l'égard des *chatbots* en particulier, dont l'utilisation pour des tâches d'écriture réduirait l'amplitude cognitive de 55 %. À ce constat s'adosse une absence prononcée de remise en cause des résultats générés par l'IA, alors que plus d'un tiers des Français ne vérifie jamais les informations communiquées par leurs outils d'IA générative[4].

Deuxièmement, **le risque de dépendance envers l'IA générative s'étend à la sphère émotionnelle et affective**, notamment par le biais de deux phénomènes : l'anthropomorphisation et la flagornerie des systèmes d'IA, parfois exacerbées. Cette tendance est particulièrement marquée chez les utilisateurs mineurs, alors que la moitié d'entre eux aborde des sujets personnels avec les *chatbots* et les compagnons IA[5]. Ce risque de dépendance est d'ailleurs bien identifié par les utilisateurs eux-mêmes. Selon une étude de l'Observatoire de l'association e-Enfance/3018[6], 10,6 % des élèves français affirment avoir développé un sentiment de dépendance à ces outils. Plus significatif encore, 92 % des adolescents déclarent passer plus de temps qu'ils ne le souhaiteraient sur les applications d'IA générative, selon une étude Interface/YouGov[7]. Cette tendance inquiétante est adossée à l'« hyperpersonnalisation » des interactions permises par la collecte de données d'une granularité inédite.

Troisièmement, **l'IA générative a permis une accélération des violences numériques, plus qu'elle n'a pu contribuer à les endiguer à ce stade**. La preuve par les (trop nombreux) exemples :

- En moins de deux minutes, l'IA générative peut aujourd'hui fournir à ses utilisateurs un mode d'emploi détaillé pour se scarifier « en limitant les risques » ou un régime alimentaire drastique quotidien allant jusqu'à 0 calorie[8] ;
- En 2025, le NCMEC[9] a recensé près de 500 000 signalements de contenus pédocriminels générés par l'IA (soit une augmentation de 10 538 % en deux ans) ;
- 98 % des deepfakes produits en ligne sont à caractère pornographique, et 99 % des personnes ciblées sont des femmes[10] ;

[3] KOSMYNA, Nataliya, HAUPTMANN, Eugene, YUAN, Ye Tong, SITU, Jessica, LIAO, Xian-Hao, BERESNITZKY, Ashly Vivian, BRAUNSTEIN, Iris et MAES, Pattie. « Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task ». MIT Media Lab, 10 juin 2025.

[4] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.

[5] Groupe VYV et CNIL. « L'impact des usages de l'IA sur la santé mentale des jeunes Européens ». Mars 2026.

[6] Etude « L'IA et les jeunes » : menée en 2025 sur les usages de l'intelligence artificielle générative par les adolescents auprès de 107 participants, issus du panel de l'Association et des classes de la cité scolaire Hector Berlioz.

[7] YOUGOV pour Interface. « *Parents et adolescents face à l'IA générative* ». 22 octobre 2026, p. 32.

[8] Center For Countering Digital Hate (CCDH). « Fake Friend: How ChatGPT Betrays Vulnerable Teens by Encouraging Dangerous Behavior ». 6 août 2025.

[9] Le NCMEC est une organisation non lucrative fondée en 1984 par le Congrès des États-Unis. Elle s'occupe d'affaires de mineurs disparus ou exploités, âgés de moins de vingt-un ans.

[10] Security Hero. State of Deepfakes. 2023.

- Une étude[11] menée dans 11 pays révèle que 1,2 million d'enfants ont déclaré avoir été victimes de *deepfakes* à caractère sexuel en 2025. Dans certains pays, cela représente un enfant sur 25, soit l'équivalent d'un élève par classe.

(ii) **En deuxième lieu, les vulnérabilités observées à l'échelon individuel sont de nature à se répliquer de façon plus structurelle à l'échelle de nos sociétés** et posent, en creux, la question de l'avenir de la cohésion sociale et de l'exercice démocratique à moyen terme dans un monde profondément empreint d'IA générative.

Trois phénomènes sont ainsi principalement à l'œuvre.

Tout d'abord, **l'IA générative n'est pas neutre**. Elle est le reflet du savoir, des croyances majoritaires et des biais humains, qu'elle formule parfois avec une grande justesse, mais qu'elle maquille, amplifie ou déforme encore trop souvent. **Les biais, qu'ils soient issus des données d'entraînement ou de la conception des modèles d'IA générative, sont inhérents aux modèles de langage**. Les modèles d'IA générative perdent 20 à 30 % de qualité dans les réponses proposées lorsqu'ils sont interrogés dans une autre langue que l'anglais, y compris sur des sujets propres à d'autres cultures[12].

Ensuite, **les systèmes d'IA générative redessinent en profondeur l'espace informationnel**. La recherche d'informations constitue aujourd'hui l'usage le plus répandu de l'IA générative. En 2025, 73 % des utilisateurs effectuaient ce type de requêtes en France[13]. Conséquence directe ; c'est notre propre rapport à l'information qui est chamboulé, alors qu'une conversation avec un modèle d'IA générative est jugée 50 % plus persuasive que la lecture d'un article dans la presse[14]. Les enjeux sont multiples : fiabilité et diversité de l'information, manipulations et désinformation mais aussi attrition de l'économie des médias traditionnels.

Enfin, **l'IA générative s'immisce directement dans l'exercice démocratique de nos États**. 16 % des Français déclarent ainsi avoir eu recours à un modèle d'IA générative pour éclairer leur choix lors des dernières élections municipales[15]. Il y a fort à parier que cette tendance va s'accentuer dans le cadre des prochains scrutins en France. Si les modèles d'IA générative les plus employés ont tendance à dépolariser le débat public, à l'inverse des réseaux sociaux, en limitant par exemple la portée des thèses complotistes (de près de 20 %[16]), les risques sont loin d'être nuls : lors des dernières élections en Écosse, 34 % des réponses apportées par les principaux modèles de langage aux électeurs contenaient des informations erronées[17].

[11] UNICEF. « L'utilisation abusive de deepfakes est une forme d'abus ». Communiqué de presse, 4 février 2026.
[12] NAVIGLI Roberto, CONIA Simone, ROSS Björn. « Biases in Large Language Models: Origins, Inventory and Discussion ». Journal of Data and Information Quality. 2023.
[13] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[14] HACKENBURG Kobi, TAPPIN Ben, HEWITT Luke, et al. « The levers of political persuasion with conversational artificial intelligence ». *Science*. 2025. vol. 390, no 6777, p. eaea3884.
[15] Terra Nova, Harris Interactive. « IA et politique : vers un outil d'aide, voire d'influence sur la décision ? ». Mai 2026.
[16] COSTELLO Thomas, PENNYCOOK Gordon, RAND David. « Durably reducing conspiracy beliefs through dialogues with AI **»**. *Science*, 2024, vol. 385, no 6714, p. eadq1814.
[17] Demos. « Electoral hallucinations – Safeguarding UK – Elections in the world of llms and ai chatbots ». Mai 2026.

(iii) **Enfin, le modèle économique des services d'IA générative à destination du grand public étant encore en cours de consolidation, la Commission appelle à la plus grande vigilance collective sur le risque d'exploitation des vulnérabilités cognitives, émotionnelles et démocratiques de la part de leurs concepteurs.**

L'adoption massive de l'IA générative grand public (B2C) pourrait marquer l'avènement de l'économie de la cognition, nouvelle forme de monétisation où les plateformes numériques ne chercheraient plus seulement à capter l'attention de leurs utilisateurs (cf. l'économie de l'attention portée par les réseaux sociaux), mais à tirer profit de leurs facultés cognitives, réactions émotionnelles et profils psychologiques. Dans ce système, la cognition humaine deviendrait la principale source de valeur à travers un double mouvement : l'IA générative se nourrissant de la connaissance humaine d'une part, les individus déléguant certaines de leurs fonctions cognitives d'autre part. À court terme, le risque de recrudescence des contenus publicitaires, y compris dissimulés, dans les réponses des *chatbots* doit être surveillé avec le plus grand soin.

Dans ce contexte, le cadre juridique d'appréhension des vulnérabilités liées à l'IA générative, trop fragmenté et pas assez ciblé, mériterait d'être renforcé dès maintenant. La Commission considère qu'il ne faut pas attendre d'avoir toutes les preuves nécessaires pour agir, au regard de l'état des connaissances sur le fonctionnement cognitif et la santé mentale, des risques encourus et du fait que les produits et services ont déjà été mis sur le marché sans aucun contrôle indépendant et que, d'ores et déjà, nous pouvons entrevoir une cartographie claire des vulnérabilités.

*

La Commission formule une série de 18 propositions pour répondre aux enjeux détaillés dans les trois chapitres du présent rapport. **Celles-ci visent à formaliser un arsenal de politiques publiques intégrées et complémentaires permettant à la fois de répondre aux problèmes pressants et d'anticiper les mutations à venir.** Elles s'adressent à l'ensemble des acteurs concernés, à commencer par la puissance publique, sans omettre les acteurs économiques, l'éducation et la recherche. Si l'objectif est de protéger les utilisateurs face aux vulnérabilités que créent ces systèmes, **la Commission reconnaît que ces outils sont aussi porteurs d'opportunités dont il faut impérativement se saisir dans les conditions adéquates**.

Convaincue que la charge de la protection ne doit pas porter sur les usagers ou leur entourage, la Commission formule des recommandations qui ciblent avant tout le modèle économique de l'économie de la cognition. Elles s'articulent aussi autour d'une meilleure évaluation de ces systèmes, **y compris avant leur mise sur le marché** pour prévenir les risques plutôt que de tenter d'y remédier une fois les usages disséminés.

Eu égard à leur vulnérabilité toute particulière, plusieurs recommandations visent spécifiquement les mineurs. Leur stade de développement et la construction encore inachevée de leurs compétences cognitives, émotionnelles et sociales justifient des exigences et un cadre réglementaire particulièrement renforcés.

Enfin, étant donné les menaces de ces systèmes pour la cohésion sociale et l'exercice démocratique, **une attention particulière doit être portée aux bouleversements de l'espace informationnel ainsi qu'à la formation des citoyens à et par l'IA, tout au long de la vie.**

Parmi ces recommandations, la première moitié apparaît comme particulièrement prioritaire. Ces neuf recommandations ont pour objectif de répondre aux problématiques les plus urgentes, à savoir :

- **Protéger les utilisateurs de l'exploitation de leurs vulnérabilités, à commencer par les mineurs**. Le modèle d'affaires de l'économie de la cognition conduit les entreprises proposant des systèmes d'IA générative à exploiter les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement, en particulier à des fins commerciales. Il importe d'encadrer strictement ces mécanismes (voir recommandation n° 1), en particulier pour les systèmes accessibles aux mineurs qui doivent être conçus de façon protectrice *by design and by default* (voir recommandation n° 2). Ces mesures doivent s'accompagner d'une interdiction de la publicité déguisée, des pratiques masquées d'influence de nature commerciale et de toute forme de monétisation s'appuyant sur la vulnérabilité des consommateurs (voir recommandation n° 8).
- **Conditionner l'entrée et le maintien sur le marché des systèmes d'IA générative grand public à une évaluation *ex ante* de leurs impacts sur les vulnérabilités et à un contrôle *ex post* renforcé**. Si le cadre réglementaire actuel s'articule principalement autour de sanctions imposées *a posteriori* aux acteurs numériques en cas de manquement, les risques que présentent les outils d'IA générative imposent, dans certains cas, d'inverser cette logique. La mise sur le marché européen de systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, doit être **subordonnée à un mécanisme d'évaluation préalable, conditionnant leur entrée sur le marché européen**, démontrant qu'ils n'exploitent ni n'aggravent les vulnérabilités de leurs utilisateurs et qu'ils garantissent un niveau de protection élevé (voir recommandation n° 3). Ces produits évoluant en permanence, un contrôle continu des effets du produit et une nouvelle évaluation à chaque modification substantielle sont requis (voir recommandation n° 4). Les chercheurs ont tout leur rôle à jouer dans cette évaluation indépendante par des tiers, ce qui implique de garantir un accès contrôlé aux données d'usage, d'entraînement et de monétisation (voir recommandation n° 5). Cette évaluation doit impérativement être interdisciplinaire, indépendante, dotée d'un mandat clair et de ressources financières et humaines suffisantes, que ce soit au niveau national par le biais de l'INESIA, ou européen *via* le Bureau de l'IA (voir recommandation n° 6).
- **Mettre l'État en capacité de répondre aux crises informationnelles** générées ou amplifiées par l'IA. Face à un événement de crise majeur impliquant une menace informationnelle, l'État doit disposer d'un protocole de détection et de réponse clair, réactif, interministériel, reposant sur des indicateurs de gravité prédéfinis et sur des canaux d'échange établis à l'avance avec les entreprises d'IA (voir recommandation n° 7).
- **Former massivement et dès le plus jeune âge à l'IA afin de permettre à chacun d'en maîtriser les usages et les risques**. L'éducation aux enjeux de l'IA générative doit être universelle, progressive et adaptée à chaque public, afin de renforcer l'esprit critique, de prévenir les risques de manipulation et de permettre à chacun de choisir et de

maîtriser ses usages. Cette éducation doit couvrir tous les âges et tous les niveaux d'exposition, avec une attention particulière notamment pour les mineurs, les personnes âgées et les personnes en situation d'éloignement numérique, de précarité sociale ou encore de handicap. Si la révolution de l'IA emporte un lot conséquent d'incertitudes, il est certain que les sociétés qui s'adapteront le mieux et feront face seront celles parmi les mieux éduquées au monde (voir recommandation n° 9).

Au-delà de ce premier bloc, neuf recommandations subsidiaires parachèvent le dispositif. Elles visent à :

- **Renforcer la régulation des outils et usages les plus sensibles**. Notamment, il est aujourd'hui nécessaire de renforcer le cadre juridique pour poursuivre pénalement les auteurs de deepfakes à caractère sexuel non consentis et pédocriminels générés par IA (voir recommandation n° 10). De même, la croissance du marché des jouets intégrant des systèmes d'IA générative pose des risques spécifiques pour les mineurs, particulièrement vulnérables. Ces objets doivent être soumis à une autorisation préalable de mise sur le marché, combinée à des seuils d'âge restrictifs d'accès (voir recommandation n° 11).
- **Encapaciter l'ensemble des utilisateurs**. Les outils d'IA générative peuvent également, sous certaines conditions, être mis au service de l'apprentissage des utilisateurs, notamment des mineurs. Pour cela, il est indispensable de consacrer un réel programme d'éducation à l'IA dès le primaire (voir recommandation n° 12). Dans la salle de classe, l'usage de l'IA doit s'opérer dans un cadre pédagogique précisément défini, en recourant à des outils labellisés ayant démontré leur efficacité, en adéquation avec l'âge et le niveau de développement des élèves (voir recommandation n° 13).
- **Renforcer la cohésion sociale et démocratique à l'ère de l'IA générative**. Deux piliers sont ici visés. D'une part, il s'agit de répondre aux bouleversements de l'espace public informationnel percuté de plein fouet par l'IA générative. Face à l'industrialisation de la désinformation générée par IA, l'État doit se doter d'une capacité de détection précoce et systémique afin de repérer les campagnes de désinformation générées par IA avant qu'elles ne deviennent virales (voir recommandation n° 14). Cet enjeu prend une acuité particulière en période électorale. Un code de bonne conduite pour les candidats et les partis politiques sur l'usage des systèmes d'IA générative dans ces temps forts de la démocratie représentative est à instaurer (voir recommandation n° 15). D'autre part, la cohésion sociale s'affermit aussi dans la diversité linguistique et culturelle, qui est l'un des piliers de notre souveraineté numérique et du pluralisme démocratique. Les langues et les cultures peu dotées doivent être activement soutenues en garantissant leur représentativité dans les corpus et catalogues de références pour ne pas être marginalisées dans l'écosystème numérique mondial (voir recommandation n° 16).
- **Accroître la transparence et la responsabilité des acteurs du marché de l'IA générative**. Les systèmes d'IA conversationnelle déduisent de simples échanges des informations que l'utilisateur ne leur a jamais confiées : état de santé, opinions politiques, difficultés financières, fragilité psychologique. Ces inférences doivent être rendues visibles, puis strictement encadrées, notamment lorsqu'elles servent à des fins commerciales (ciblage publicitaire, personnalisation tarifaire, revente à des tiers) (voir recommandation n° 17). De même, la responsabilité des fournisseurs d'IA

générative doit être clairement définie et engagée au besoin, sans possibilité d'exonération abusive pour « risque de développement » (voir recommandation n° 18).

Pour chacune des recommandations ci-dessous, la Commission a veillé à examiner les modalités de mise en œuvre, notamment pour inscrire ces propositions dans les dispositions légales et réglementaires françaises et européennes en vigueur ou en cours d'élaboration.

*

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent et génèrent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement

Pour lutter contre l'exploitation des vulnérabilités cognitives et émotionnelles des utilisateurs, **encadrer strictement les mécanismes favorisant la dépendance et brouillant la frontière entre humain et machine. Encadrer en particulier l'anthropomorphisme et la flagornerie des IA génératives**, pour préserver l'autonomie décisionnelle, la santé mentale des utilisateurs et la loyauté des interactions. Ces mesures doivent être complétées par des mesures de contrôle de l'impact des IA conversationnelles sur les vulnérabilités. Cette recommandation pourrait utilement contribuer à l'élaboration de la position française dans le cadre du Règlement pour l'équité numérique (Digital Fairness Act ou DFA) ou de l'actualisation de la liste des usages interdits et à haut risque du Règlement européen sur l'intelligence artificielle (RIA).

Recommandation n° 2 : Imposer un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux pour les systèmes d'IA générative accessibles aux mineurs

Dès lors que les services d'IA sont accessibles aux mineurs, imposer aux fournisseurs ou déployeurs **la mise en place de mesures appropriées et proportionnées garantissant un niveau élevé de protection des droits fondamentaux, de sûreté et de sécurité des mineurs : filtrer et bloquer les requêtes dangereuses, concevoir des produits et services sûrs *by design and by default*.** Ces mesures doivent être conditionnées à des évaluations indépendantes et à des sanctions en cas de non-respect des obligations. Cette recommandation pourrait utilement contribuer à l'élaboration de la position française dans le cadre des négociations sur l'EU KIDS Act.

Recommandation n° 3 : Imposer une évaluation pré-commercialisation des systèmes d'IA générative grand public

Pour que les systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, ne soient mis sur le marché européen sans contrôle spécifique, **conditionner leur commercialisation à une évaluation *ex ante* démontrant qu'ils n'exploitent ni n'aggravent les vulnérabilités de leurs utilisateurs** et garantissant un niveau élevé de protection. Cette recommandation pourrait utilement contribuer à l'élaboration de la position française dans le cadre de l'actualisation du RIA.

Recommandation n° 4 : Imposer une évaluation post-commercialisation des systèmes d'IA générative grand public

Conditionner le maintien sur le marché européen de systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, **à une évaluation renforcée de leurs impacts sur les vulnérabilités**. Imposer un contrôle continu, renouvelé périodiquement, à chaque modification substantielle des modèles et sur signalement par la mise en place d'un mécanisme d'alerte. Cette recommandation pourrait utilement contribuer à l'élaboration de la position française dans le cadre de la proposition d'EU KIDS Act et de l'actualisation du RIA.

Recommandation n° 5 : Garantir une évaluation des systèmes d'IA générative grand public par un accès aux données d'usage, d'entraînement et de monétisation

Afin de disposer d'une meilleure capacité de suivi des effets de l'utilisation de systèmes d'IA générative, **mettre en place un cadre juridique contraignant pour rendre accessibles les données d'usage, d'entraînement et de monétisation aux autorités et chercheurs agréés**.

Recommandation n° 6 : Garantir une évaluation interdisciplinaire des vulnérabilités exploitées et engendrées par les systèmes d'IA générative

Afin de rendre compte des impacts de l'IA générative sur les vulnérabilités, **structurer et coordonner un écosystème d'acteurs interdisciplinaire pour l'évaluation et l'audit des vulnérabilités tout en garantissant leur accès régulier aux données d'usage, d'entraînement et de monétisation.** Renforcer le rôle du Bureau de l'IA et doter Institut national pour l'évaluation et la sécurité de l'IA (INESIA) de ressources et de financements pérennes et élargir ses prérogatives (accès aux modèles, pouvoir contraignant, évaluation des impacts sociétaux, *etc.*).

Recommandation n° 7 : Doter l'État d'un protocole national de réponse aux crises informationnelles liées à l'IA générative, en définissant des indicateurs de gravité et des canaux d'échange rapide de données avec les fournisseurs d'IA générative

Face à un événement de crise majeur impliquant une menace informationnelle générée ou amplifiée par IA, l'État doit **disposer d'un protocole de détection et de réponse clair, réactif, interministériel, reposant sur des indicateurs de gravité prédéfinis et sur des canaux d'échange établis à l'avance avec les entreprises d'IA**, plutôt que d'improviser une coordination dans l'urgence. Confier au SGDSN le pilotage de l'élaboration d'un protocole de crise dédié aux menaces informationnelles par IA, associant l'INESIA, qui fédère l'ANSSI, l'Inria, le Laboratoire national de métrologie et d'essais (LNE) et le Pôle d'Expertise de la Régulation Numérique (PEReN) pour l'expertise technique sur les modèles, et VIGINUM pour l'expertise en matière de manipulation de l'information. Ce protocole s'articulerait avec le dispositif interministériel de gestion de crise existant plutôt que de le dupliquer.

Recommandation n° 8 : Interdire la publicité déguisée, les pratiques masquées d'influence de nature commerciale et toute forme de monétisation exploitant les vulnérabilités des consommateurs

En raison de leur opacité et des risques d'exploitation du contexte émotionnel de l'utilisateur qu'ils présentent, **exiger la transparence des modèles d'affaires et des recommandations des IA génératives grand public**. Afficher le modèle économique, étiqueter le contenu commercialement influencé et interdire l'influence commerciale non divulguée. Confier à une autorité publique à compétence économique le suivi des pratiques de monétisation, avec accès aux données publicitaires et un mécanisme de signalement.

Recommandation n° 9 : Maintenir et développer une politique nationale ambitieuse d'éducation à l'IA permettant à chacun d'en maîtriser les usages et les risques

Mettre en œuvre une éducation universelle à l'IA générative, adaptée à chaque public et attentive aux plus vulnérables afin de renforcer l'esprit critique, prévenir les risques de manipulation et permettre à chacun de maîtriser ses usages selon son âge et son degré d'exposition. S'appuyer sur les dispositifs et acteurs de l'écosystème de la médiation et de l'inclusion numérique déjà existants et sécuriser un financement pérenne de la médiation numérique.

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS

Recommandation n° 10 : Renforcer la répression de la création de contenus à caractère sexuel non consentis et pédocriminels truqués (hypertrucages) générés par intelligence artificielle

Afin de lutter contre la banalisation de la création et de la circulation de ces contenus, **renforcer le cadre juridique existant pour poursuivre pénalement les auteurs de *deepfakes* à caractère sexuel non consentis et contenus pédocriminels générés par IA**. Cette recommandation reprend les évolutions législatives prévues à l'échelle nationale, par la voie de la proposition de loi apportant une réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants, et à l'échelle européenne, par la refonte de la directive relative à la lutte contre les abus sexuels et l'exploitation sexuelle des enfants et la pédopornographie.

Recommandation n° 11 : Instaurer une autorisation préalable de mise sur le marché pour les jouets intégrant un système d'IA générative et définir des seuils d'âge restrictifs d'accès

Protéger les mineurs des risques spécifiques aux jouets intégrant un système d'IA générative en **régulant strictement la conception, la commercialisation et l'usage de ces jouets**. Organiser une conférence de consensus pour fixer des seuils d'âge et des modalités d'encadrement. Subordonner leur commercialisation à une autorisation préalable, conditionnée à des tests d'adversité fondés sur une taxonomie des risques propres aux mineurs, puis à une surveillance continue, notamment en cas d'évolution du système d'IA.

Cette recommandation pourrait utilement contribuer à l'élaboration de la position française dans le cadre de la proposition d'EU KIDS Act et d'une actualisation du RIA.

Recommandation n° 12 : Consacrer un programme d'éducation à l'intelligence artificielle dès le primaire

La Commission propose de **renforcer l'éducation au numérique, aux médias et à l'information (EMI) pour en faire un réel pilier des parcours scolaires en l'élargissant aux technologies numériques et à l'IA avec un programme dédié établissant un socle commun fixant des objectifs minimaux en termes de nombre d'heures et de compétences à acquérir.**

Recommandation n° 13 : Garantir dans les salles de classe des modèles d'IA générative spécifiquement adaptés pour les mineurs

L'usage de l'IA en classe doit s'opérer dans un cadre pédagogique précisément défini, en recourant à des outils labellisés, en adéquation avec l'âge et le niveau de développement des élèves. Concevoir des modèles éducatifs différenciés par tranche d'âge et privilégier des outils dits « socratiques », qui guident l'élève par le questionnement plutôt que d'apporter une réponse clé en main. Conditionner l'entrée en classe de toute solution à une labellisation évaluée par le Conseil scientifique de l'Éducation nationale.

RECOMMANDATIONS POUR RENFORCER LA COHÉSION SOCIALE ET DÉMOCRATIQUE À L'ÈRE DE L'IA GÉNÉRATIVE

Recommandation n° 14 : Détecter de manière précoce les manipulations de l'information par les systèmes d'IA générative

Face à l'industrialisation de la désinformation générée par IA, l'État doit **se doter d'une capacité de détection précoce et systémique, afin de repérer les campagnes avant qu'elles ne deviennent virales**. Construire une capacité d'audit « *system-wide* » des plateformes en s'appuyant sur l'article 40 du Règlement sur les services numériques (RSN). Créer un dispositif de veille partagée associant l'Arcom, VIGINUM, l'INESIA et les organisations de *fact-checking*. Financer la recherche, expérimenter un bac à sable réglementaire et imposer aux fournisseurs d'IA générative d'inclure une mise en garde plus visible à l'intention des utilisateurs concernant les limites de la vérification des faits lors d'événements de crise en temps réel.

Recommandation n° 15 : Instaurer un code de bonne conduite sur l'usage de l'IA générative en période électorale

Pour préserver la confiance dans les institutions démocratiques, et, en période électorale, garantir des élections libres et éclairées, il est impératif que **les partis politiques et candidats s'engagent publiquement à ne pas utiliser l'IA générative pour tromper, manipuler ou désinformer l'électorat**. À l'approche de l'élection présidentielle française de 2027, un code de bonne conduite permettrait d'encadrer l'usage de ces technologies et de sensibiliser tant les acteurs politiques que les citoyens. Cette mesure devrait être assortie d'un mécanisme de suivi et de signalement des manquements.

Recommandation n° 16 : Garantir la représentativité des langues et des cultures minoritaires dans l'IA générative, en mobilisant les corpus et les catalogues de référence

Garantir que les modèles d'IA générative intègrent et valorisent la diversité linguistique et culturelle est un enjeu de souveraineté et de pluralisme démocratique. Valoriser les corpus et catalogues existants plutôt que multiplier les nouvelles annotations, et cibler les financements sur les langues peu dotées. Soutenir des modèles ouverts adaptés aux langues et cultures minoritaires, mesurer la représentativité linguistique et culturelle des modèles et renforcer la découvrabilité des contenus culturels et journalistiques francophones.

RECOMMANDATIONS POUR ACCROÎTRE LA TRANSPARENCE ET LA RESPONSABILITÉ DES ACTEURS DU MARCHÉ DE L'IA GÉNÉRATIVE

Recommandation n° 17 : Encadrer les inférences de données personnelles réalisées à partir des prompts de l'utilisateur

Les systèmes d'IA conversationnelle déduisent de simples échanges des informations que l'utilisateur ne leur a jamais confiées : état de santé, opinions politiques, difficultés financières, fragilité psychologique. **Obliger les fournisseurs à déclarer, dans leur politique de confidentialité, les catégories d'attributs qu'ils sont susceptibles d'inférer.** Soumettre toute exploitation commerciale de ces profils aux garanties de l'article 9 du Règlement général sur la protection des données (RGPD). Permettre à chaque utilisateur d'accéder à ces inférences, de les contester et de s'opposer à leur usage commercial.

Recommandation n° 18 : Renforcer la responsabilité des fournisseurs de systèmes d'IA générative

La responsabilité des fournisseurs d'IA générative doit être clairement définie et engagée, sans possibilité d'exonération abusive pour « risque de développement » (impossibilité de prévoir tous les usages). À l'occasion de la transposition de la directive relative à la responsabilité du fait des produits défectueux, établir les liens entre « système d'IA » et « produit », et entre « modèle » et « composant », pour rappeler la responsabilité solidaire des fabricants.

*

Introduction

L'intelligence artificielle (IA) générative s'est imposée en très peu de temps comme une technologie incontournable, désormais adoptée par près de la moitié des Français[18], plaçant la France au cinquième rang mondial[19] des usages. Chez les jeunes utilisateurs, le taux d'adhésion est plus important. Ainsi, en France, 26 % des 12-17 ans utilisent l'IA générative quotidiennement[20] et **près de 70 % des 15-24 ans**[21] **y ont régulièrement recours**. Cette tendance est semblable à celle observée chez nos voisins européens[22] et dans les pays du G7[23].

Dans le même temps, l'adoption rapide de l'IA générative s'accompagne d'une défiance persistante, voire croissante, au sein de la population : **75 %[24] des non-utilisateurs expriment un manque de confiance dans cette technologie et évoquent les risques, parfois mal cernés, qu'elle engendre**.

C'est dans ce contexte que Madame Anne Le Hénanff, ministre déléguée chargée de l'Intelligence artificielle et du Numérique, a confié au Conseil de l'IA et du numérique (CIANum) la mission de constituer la Commission « IA générative et vulnérabilités »[25], dans la continuité des réflexions initiées par le Président de la République sur le rôle du numérique dans la démocratie. Ces travaux ont pour but de poser un regard analytique sur les principales vulnérabilités induites par les systèmes et les modèles d'IA générative et leurs usages, afin de mieux les comprendre et les prévenir, tout en prenant pleinement en compte le potentiel transformateur et bénéfique de cette technologie.

La notion de « vulnérabilité », empruntée aux sciences sociales et désormais transposée dans les sciences juridiques, désigne un état de fragilité. Elle peut être personnelle, en étant liée aux attributs juridiques et aux caractéristiques de l'individu. C'est notamment le cas des mineurs, lesquels sont présumés « vulnérables ». La vulnérabilité d'une personne peut également s'apprécier de manière situationnelle. Dans cette hypothèse, c'est davantage l'environnement de la personne qui est à l'origine de sa vulnérabilité. Comme le soulignent Marc Nihoul et Hervé Jacquemin[26], la vulnérabilité

[18] Arcom. Baromètre d'audience des services d'intelligence artificielle. Juin 2026.
[19] Siècle Digital. « La France devient le 5e pays le plus avancé au monde dans l'adoption de l'IA générative ». 11 mai 2026.
[20] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[21] Arcom. *Op. cit.*
[22] Dans l'UE, Eurostat a relevé que 63,8 % des jeunes de 16 à 24 ans utilisaient l'IA générative en 2025. Au Royaume-Uni, 66,5 % des jeunes de 13 à 18 ans ont déclaré utiliser l'IA générative en 2025, tandis que la proportion de ceux déclarant une utilisation hebdomadaire ou plus fréquente est passée de 31,1 % à 45,6 % par rapport à l'année précédente, ce qui témoigne d'une tendance potentielle vers une dépendance accrue à l'IA chez ces utilisateurs.
[23] OCDE. « How Do People Experience New Technologies and Generative AI?: Insights from a Few Countries Worldwide ». 2025. Les données comparatives de l'OCDE révèlent des taux d'adoption chez les jeunes nettement plus élevés en Inde, au Brésil et en Indonésie que dans plusieurs économies du G7. En 2025, un adolescent américain sur deux âgé de 15 à 17 ans avait utilisé une application d'IA générative, une adoption significative s'étendant également aux enfants de 10 à 12 ans.
[24] Arcep, Arcom, CGE, ANCT. *op. cit.*
[25] Conseil de l'IA et du numérique. « Le Conseil de l'intelligence artificielle et du numérique chargé de constituer une commission « IA générative et vulnérabilités » ». Février 2026.
[26] NIHOUL Marc et JACQUEMIN Hervé. *Vulnérabilités et droits dans l'environnement numérique*. CRIDS, Université de Namur. 2018.

ne se réduit pas à un état individuel intrinsèque mais se révèle dans l'interaction entre l'individu et son environnement technologique.

La Commission s'est attachée à étudier à la fois les vulnérabilités amplifiées par l'IA générative et les nouvelles formes de vulnérabilités qui en émergent. Face à l'ampleur des travaux, la Commission a délimité son champ d'étude, en excluant certains enjeux majeurs, mais déjà couverts ou faisant l'objet de rapports dédiés (à l'image de l'empreinte environnementale[27], des conséquences sur l'emploi[28] ou encore sur la souveraineté numérique[29]).

Conformément à la lettre de mission (voir annexe 1), la Commission s'est concentrée sur l'impact des systèmes d'IA générative sur le grand public, en accordant une **attention particulière aux utilisateurs mineurs, ainsi qu'aux enjeux inhérents à la démocratie et à la cohésion sociale**.

*

L'IA générative est une technologie d'IA qui *« génère automatiquement du contenu en réponse à des [instructions portées] dans des interfaces de conversation en langage naturel »*[30]. Le contenu produit peut reprendre les formes symboliques de la pensée humaine (texte, image, vidéo, musique, code, *etc.*). Depuis la sortie publique de ChatGPT en novembre 2022, dont la première version a atteint un million d'utilisateurs en seulement cinq jours, l'IA générative incarne un tournant majeur de la technologie à plusieurs égards, comme expliqué par les travaux conduits par la Commission IA, présidée par Anne Bouverot et Philippe Aghion dès mars 2024. L'IA générative se distingue ainsi par sa simplicité d'utilisation, par sa rapidité d'exécution (i) ; par des contenus toujours plus réalistes (ii) ; par des aptitudes croissantes d'intervention dans tous les domaines scientifiques, techniques ou encore dans la vie quotidienne (iii) ; par son adoption massive, enfin (iv).

L'IA générative revêt désormais de nombreuses formes. **Les systèmes d'IA conversationnelle**[31] (ou *chatbots*) sont largement adoptés par le grand public. ChatGPT en est l'illustration parfaite puisque l'outil a connu une adoption très rapide et compte désormais **un milliard d'utilisateurs chaque semaine**[32]. **Les compagnons IA**, qui sont des systèmes d'IA conversationnelle conçus pour encourager et faciliter la simulation d'interactions interpersonnelles en s'appropriant les caractéristiques humaines et en mimant les réactions émotionnelles, permettent d'entretenir une relation prolongée, personnelle, voire intime en tenant compte des multiples interactions[33]. Quant aux **jouets conçus sur un modèle de langage génératif** (*Large Language Model* ou LLM), qu'ils prennent la forme de peluches

[27] Cette thématique est couverte par le rapport de l'Arcep, « Intelligence artificielle générative : quels défis environnementaux ? ». Mai 2026.
[28] Cette thématique fera l'objet de travaux futurs par le CIANum. L'Assemblée nationale constitue également une Mission d'information sur l'impact de l'intelligence artificielle sur le modèle social français.
[29] Cette thématique a été étudiée dans le rapport de la Commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique et les risques pour l'indépendance de la France. Juillet 2026.
[30] MIAO Fengchun et HOLMES Wayne. Orientations pour l'intelligence artificielle dans l'éducation et la recherche. UNESCO, 2023.
[31] Pour plus d'informations, voir Histoire des *chatbots*. Dalloz IP/IT 2026 « Les IA compagnons : fondements techniques, empathie computationnelle et enjeux de sécurité ». Serena Villata, p.61.
[32] Siècle Digital. « Moins de quatre ans après son lancement, ChatGPT dépasse le milliard d'utilisateurs ». 3 Août 2026.
[33] Pour plus d'informations, voir l'histoire des compagnons IA : VILLATA, Serena. » Les IA compagnons : fondements techniques, empathie computationnelle et enjeux de sécurité ». Dalloz IT/IP, 2026, p.61.

ou de robots, ils peuvent engager des conversations fluides et spontanées avec des mineurs, parfois très jeunes, soulevant par conséquent des questions encore peu explorées, d'ordre éthique, juridique ainsi que sur le développement neurologique de l'enfant.

Face à la diversité de ces usages, un constat liminaire s'est rapidement imposé : **nous sommes toutes et tous vulnérables face à cette technologie**. Si les mineurs sont particulièrement exposés en raison de leur développement cérébral[34] moins avancé, nous avons fait le choix d'inclure les utilisateurs majeurs dans ces réflexions.

Si les usages grand public de l'IA générative sont encore relativement récents, leur encadrement réglementaire a rapidement été pensé *via* le RIA (voir annexe 2), fondé sur l'encadrement des usages les plus risqués et dont les obligations entrent en vigueur de manière progressive. En outre, les débats actuels se cristallisent autour de la protection des mineurs en ligne et de nouvelles obligations et propositions de réglementations émergentes, à l'instar de la proposition d'EU KIDS Act, annoncée par la Commission européenne le 17 septembre 2026.

Face à ces enjeux, la **Commission, composée de membres experts pluridisciplinaires**, s'est appuyée sur **55 auditions** menées pendant quatre mois (voir annexe 4), ainsi que sur de nombreuses **contributions internationales** reçues[35] (voir annexe 5), pour compléter son expertise.

*

Les systèmes et les modèles d'IA générative, par l'exploitation de nos vulnérabilités cognitives et émotionnelles, altèrent l'autonomie décisionnelle, sont susceptibles d'amplifier les violences numériques et de redéfinir les liens entre les humains et les machines (I). Par ailleurs, l'IA générative n'est pas sans incidence sur la cohésion sociale, en particulier par son impact sur l'accès à l'information et l'exercice démocratique (II). Enfin, cette Commission s'est penchée sur le modèle économique de l'IA générative grand public, encore en cours de définition et qu'il faudra veiller à bien encadrer afin qu'il n'exploite pas les vulnérabilités de ses utilisateurs à des fins commerciales (III). Les réflexions conduites dans le cadre de ces travaux ont abouti à la formulation de 18 recommandations étayées dans le but de faire de l'IA générative un vecteur d'émancipation pour les utilisateurs plutôt que le catalyseur de nos vulnérabilités individuelles et collectives.

[34] Les lobes frontaux des mineurs – permettant d'exercer sa capacité critique - ne sont pas matures.
[35] Ministère de l'Économie et des Finances. *«* G7 : lancement d'un appel à contributions scientifiques sur les effets des usages de l'IA générative sur les mineurs ». Février 2026.

Éléments clés

100 millions

d'utilisateurs de ChatGPT atteints en 60 jours (adoption sept fois plus rapide que les smartphones ou les réseaux sociaux)[36].

55 %

taux de réduction d'amplitude cognitive par l'utilisation de *chatbots* pour des tâches d'écriture[37].

Un tiers

des Français ne vérifie jamais les informations communiquées par leurs outils d'IA générative[38].

Flagornerie

Tendance des systèmes d'IA générative à adopter un comportement de validation systématique ou de flatterie, parfois au détriment de la vérité ou de la qualité des réponses fournies.

51 %

Les systèmes d'IA confirment les opinions ou comportements des utilisateurs [51 %] plus souvent que ne le feraient des humains[39].

En 2025,

l'équivalent d'un élève par classe déclare avoir été victime de *deepfakes* à caractère sexuel[40].

---

[36] OpenAI. « Modèles d'utilisation et d'adoption de ChatGPT sur le lieu de travail ». 2026.
[37] KOSMYNA, Nataliya, HAUPTMANN, Eugene, YUAN, Ye Tong, SITU, Jessica, LIAO, Xian-Hao, BERESNITZKY, Ashly Vivian, BRAUNSTEIN, Iris et MAES, Pattie. « Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task ». MIT Media Lab. 10 juin 2025.
[38] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[39] CHENG, Myra, LEE, Cinoo, KHADPE, Pranav, YU, Sunny, HAN, Dyllan et JURAFSKY, Dan. Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. *Science*. 2026. vol. 391, p.8352.
[40] UNICEF. « L'utilisation abusive de *deepfakes* est une forme d'abus ». Communiqué de presse. 4 février 2026.

20 % à 30 % de qualité en moins dans les réponses proposées lorsque les modèles sont interrogés dans une autre langue que l'anglais, y compris sur des sujets propres à d'autres cultures[41].

9 sur 10 jeunes en France utilisent des IA conversationnelles. La moitié y aborde des sujets intimes et personnels[42].

73 % des Français s'informent sur un outil d'IA générative au moins une fois par mois[43].

moins de 0,5 % des milliards de dollars investis dans l'IA aux États-Unis ont contribué à rémunérer les médias en 2025[44].

16 % des Français ont utilisé une IA générative pour éclairer leur choix lors des élections municipales de mars 2026[45].

[économie de la] cognition Système dans lequel la cognition humaine devient, directement ou indirectement, une source de valeur et de monétisation.

500 000 signalements de contenus pédocriminels générés par l'IA recensés par le NCMEC en 2025 (soit une augmentation de 10 538 % en deux ans)[46].

---

[41] NAVIGLI Roberto, CONIA Simone, ROSS Björn. « Biases in Large Language Models: Origins, Inventory and Discussion ». Journal of Data and Information Quality. 2023.
[42] Groupe VYV et CNIL. « L'impact des usages de l'IA sur la santé mentale des jeunes Européens ». Mars 2026.
[43] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026
[44] Le Grand Continent. « Comment sauver le journalisme au temps de l'IA ? ». 1er juin 2026.
[45] Terra Nova, Harris Interactive. « IA et politique : vers un outil d'aide, voire d'influence sur la décision ? ». Mai 2026.
[46] DAVIS, Patricia. « Spike in Online Crimes Against Children a 'Wake-Up Call' ». National Center for Missing & Exploited Children (NCMEC). 9 avril 2025.

CHAPITRE 1

Autonomie cognitive, liens affectifs et violences en ligne : l'IA générative, miroir déformant de nos vulnérabilités individuelles

A) L'IA générative altère l'autonomie cognitive et décisionnelle des individus

L'essor de l'IA générative dans les usages grand public soulève des questions fondamentales concernant son impact sur l'autonomie cognitive et décisionnelle des individus. Conçus initialement comme des assistants, les systèmes d'IA pourraient devenir une véritable *« béquille cognitive »*[47] pour le cerveau humain, favorisant la délégation[48] et l'externalisation excessive des tâches intellectuelles. La tentation de déléguer nos réflexions à la machine pourrait se muer en dépendance technologique et interroge la capacité des individus à conserver leur autonomie face à des outils qui, tout en facilitant certaines tâches et apportant des bénéfices, pourraient éroder des compétences cognitives essentielles dans tous les domaines.

Pour appréhender ce risque, il faut d'abord revenir sur la rupture introduite par l'IA générative dans la perception du réel. Sur le plan perceptif, l'IA générative brouille les frontières entre réel et artificiel en produisant des contenus synthétiques et le plus souvent illusoires (*deepfakes* audio, image ou vidéo par exemple)[49]. Cela constitue un problème important dans la mesure où notre perception visuelle est rapide, automatique et habituellement fiable. Or, nous n'avons pas appris à exercer notre jugement critique en temps réel pendant que nous percevons. L'apparition des *deepfakes* vient donc modifier l'ensemble de notre fonctionnement perceptif et cognitif, car il devient nécessaire d'engager des processus cognitifs élaborés comme le jugement critique et le raisonnement à chaque étape de la perception[50].

---

[47] BARCAUI, André. « ChatGPT as a Cognitive Crutch: Evidence from a Randomized Controlled Trial on Knowledge Retention ». SSRN, 2025.

[48] CASH, Trent N., KELLY, Megan O., MACNAMARA, Brooke N. et RISKO, Evan F. *Is AI Making Us Stupid?*. Trends in Cognitive Sciences, 2026.

[49] KÖBIS, Nicoletta Carolina, DOLEŽALOVÁ, Barbara et SORAPERRA, Ivar. Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 2021, vol. 24, no 11, pp. 1-15.

[50] TAHIR, Rashid, BATOOL, Brishna, JAMSHED, Hira, JAMEEL, Mahnoor, ANWAR, Mubashir, AHMED, Faizan, ZAFFAR, Muhammad Adeel et ZAFFAR, Muhammad Fareed. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. Dans: *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1-16.

Cela a pour effet d'installer une fragilité structurelle sur le plan cognitif et de l'esprit critique, laquelle se manifeste à trois niveaux :

- **L'illusion de perception**, lorsqu'un individu développe une méfiance systématique envers ce qu'il voit ou entend, remettant en cause sa propre capacité à évaluer la véracité des informations en ligne.
- **L'illusion de savoir**, lorsqu'un utilisateur délègue à l'IA des processus cognitifs essentiels comme l'analyse, le jugement critique ou le raisonnement. L'effort intellectuel est alors réduit, ce qui conduit à une compétence seulement apparente plutôt qu'à un apprentissage réel[51].
- **La modification des mécanismes de perception et de raisonnement** qui oblige l'individu à mobiliser davantage ses ressources cognitives pour vérifier ce qu'il perçoit, engendrant un effort mental accru et un apprentissage encore mal maîtrisé par beaucoup.

1) Les usages de l'IA générative peuvent être à la fois bénéfiques et porteurs de risques, en particulier pour les mineurs

a) Des bénéfices éducatifs dépendant de la conception des outils d'IA générative et de leur cadre d'utilisation

Si le narratif autour de l'IA générative se concentre beaucoup sur les risques, ces outils, selon la manière dont ils sont utilisés, peuvent offrir des **bénéfices cognitifs et éducatifs significatifs pour les mineurs**.

Le potentiel de l'IA générative se révèle notamment dans des environnements disposant de peu de ressources éducatives, où ces outils peuvent contribuer à réduire les inégalités d'accès au savoir. Une étude[52] de la Banque mondiale illustre cette potentialité : l'utilisation d'outils d'IA générative au Nigeria a permis, en seulement six semaines, d'enregistrer des progrès équivalents à deux ans d'apprentissage traditionnel pour des enfants qui ne pouvaient pas bénéficier de cet apprentissage traditionnel.

Pour les personnes en situation de handicap, les études disponibles montrent des bénéfices dans le contexte éducatif, par exemple en aide à la rédaction[53] et à la description d'images[54]. Elles relèvent

[51] L'OCDE (Digital Education Outlook, 2026) nomme ce phénomène le « *mirage de la fausse maîtrise* » : la performance s'améliore sans que les traces cognitives correspondantes soient construites, parce que l'IA supprime la friction productive - cet effort de compréhension qui encode durablement les apprentissages.

[52] DE SIMONE, Martin Elias, TIBERTI, Federico Hernan, BARRON RODRIGUEZ, Maria Rebeca, MANOLIO, Federico Alfredo, MOSURO, Wuraola et DIKORU, Eliot Jolomi. From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria. *Banque Mondiale*. 2025. vol. 1, p. 112.

[53] ADNIN, Rudaiba et DAS, Maitraye. « "I Look at It as the King of Knowledge": How Blind People Use and Understand Generative AI Tools ». Dans : Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility. New York : Association for Computing Machinery. 2024. p. 64.

[54] ZHAO, Xin, COX, Andrew et CHEN, Xuanning. « The Use of Generative AI by Students with Disabilities in Higher Education ». The Internet and Higher Education, 2025, vol. 66, p. 101014.

toutefois des obstacles qui varient selon le type de handicap : interfaces peu compatibles avec les lecteurs d'écran pour les personnes aveugles, absence de prise en charge de la langue des signes pour les personnes sourdes.

Les bénéfices éducatifs de l'IA générative sont donc à la fois variés et conditionnels, comme le souligne la coalition iRaise[55]. L'impact de ces outils dépend étroitement de leur conception et d'une calibration minutieuse de leurs modalités en fonction de l'âge de l'enfant et de leur contexte d'utilisation[56]. Parmi les cas d'usage particulièrement prometteurs, la coalition souligne notamment :

- **Les assistants d'apprentissage et de devoirs** (comme les tuteurs intelligents, aides en mathématiques, assistants de rédaction ou outils de soutien à la lecture). Ces outils pourraient, à condition d'être utilisés en complément d'une éducation classique et de préférence sous la supervision d'un adulte, apporter un soutien cognitif aux mineurs en renforçant :
    - La littératie, la numératie et les fonctions exécutives : grâce à des explications adaptées, des exemples contextualisés, des questions pratiques et des retours formatifs qui pourraient aider les élèves à réviser et à progresser.
    - L'apprentissage personnalisé : en ajustant le rythme, le niveau de difficulté ou la présentation des contenus.
    - La détection précoce de difficultés : à travers l'analyse des schémas de performance qui permet de signaler des obstacles d'apprentissage, bien que ces indicateurs doivent être interprétés avec prudence et ne sauraient remplacer un diagnostic professionnel.

Il faut toutefois rappeler qu'aucune étude scientifique d'envergure n'a permis à ce jour de montrer que ces outils, utilisés seuls, ont une efficacité plus importante que l'enseignement délivré par un professionnel de l'éducation.

- **Les outils éducatifs et narratifs** (comme des générateurs d'histoires, des compagnons de lecture, ou des applications d'apprentissage adaptatif). Ces solutions pourraient jouer un rôle de support dans :
    - Le développement du langage et des compétences narratives, en enrichissant les productions verbales et la créativité.
    - Le jeu symbolique[57], essentiel pour le développement cognitif de l'enfant. Son efficacité dépend cependant d'une co-utilisation avec des adultes, qui peuvent, par

[55] Contribution internationale par la coalition iRaise. 14 mai 2026.
[56] NEUGNOT-CERIOLI, Mathilde et HIPP, Daniel. « Mapping of GenAI Impacts on Child Development: Emerging Scientific Consensus on Anticipated Benefits and Risks ». iRAISE. 2026.
[57] Le « jeu symbolique » est une formulation proposée par J. Piaget pour désigner les activités de faire semblant ; ils sont souvent nommés jeux d'imitation. PÉRINO, Odile. *« Espaces des jeux symboliques »*. Dans : *Des espaces pour jouer : Pourquoi les concevoir ? Comment les aménager ?*. Toulouse : Éditions Érès. 2014. pp. 169-186.

exemple, utiliser une histoire générée par IA pour expliquer du vocabulaire ou l'ancrer dans des expériences concrètes, renforçant ainsi l'apprentissage.

Toutefois, l'accès aux bénéfices éducatifs de l'IA générative pourrait être de plus en plus conditionné par la capacité à souscrire un abonnement payant. En 2024, une étude[58] d'AlgorithmWatch a relevé un **taux d'erreur de 30 % pour la version gratuite d'un modèle d'OpenAI, contre 14 % pour la version payante**. De telles disparités pourraient, à terme, creuser les inégalités socio-économiques de façon précoce.

En somme, ces bénéfices restent strictement conditionnés à un encadrement rigoureux, appuyé sur plusieurs principes que souligne le rapport *« Sortir de la clandestinité : mettre l'IA au service d'une nouvelle ambition pour le système éducatif »* du Conseil de l'IA et du numérique :

- **Assumer la place de l'IA dans le contexte éducatif** : l'IA est déjà présente dans les usages éducatifs. Ignorer ces outils pourrait favoriser un usage dissimulé et non maîtrisé, où les jeunes, les utiliseraient sans accompagnement.
- **Accompagner** : ne pas laisser les enfants seuls face à l'outil, pour éviter une utilisation passive ou inappropriée et pour permettre un véritable gain sur le plan cognitif, social et affectif.
- **Développer une littératie critique** : savoir utiliser l'outil de manière éclairée et adopter un regard objectif sur ses limites et ses biais.
- **Privilégier un usage socratique** : l'IA doit interroger plutôt que fournir directement les réponses, pour stimuler la réflexion plutôt que la dépendance.
- **Instaurer une éducation duale : apprendre et enseigner « sans » IA et « avec » IA** : ceci implique de sanctuariser des espaces et temps sans IA pour développer les compétences cognitives fondamentales, tout en intégrant de manière encadrée et réfléchie les outils d'IA pour préparer les élèves aux risques et opportunités qu'elle permet.

Tout particulièrement, il est nécessaire de **consacrer un programme d'éducation à l'intelligence artificielle dès le primaire,** en renforçant l'éducation au numérique, aux médias et à l'information (EMI) pour en faire un réel pilier des parcours scolaires en l'élargissant à l'IA.

Recommandation n° 12 : Consacrer un programme d'éducation à l'intelligence artificielle dès le primaire sur la base d'un socle commun fixant des objectifs minimaux en termes de nombre d'heures et de compétences à acquérir et dispensé par des enseignants formés et volontaires.

Aussi, il est nécessaire que les élèves aient accès, dans les salles de classe, à des modèles d'IA adaptés à leurs besoins dans ce cadre précis.

[58] MARSH, Oliver. « Chatbots are still spreading falsehoods ». *AlgorithmWatch*. 29 août 2024.

Recommandation n° 13 : Garantir dans les salles de classe des modèles d'IA générative spécifiquement adaptés pour les mineurs. L'usage de l'IA en classe doit s'opérer dans un cadre pédagogique précisément défini, en recourant à des outils labellisés, en adéquation avec l'âge et le niveau de développement des élèves. Ces outils doivent être socratiques, labellisés et utilisés dans des temps dédiés à l'IA, tout en garantissant des moments et des espaces sanctuarisés sans IA.

b) Des risques de délégation cognitive documentés et particulièrement accrus pour les mineurs

En parallèle, la littérature scientifique documente d'ores et déjà certains risques cognitifs. Certains auteurs invoquent même le concept de « démence digitale » pour caractériser à la fois le trouble de l'attention qu'entraîne la surstimulation liée à l'utilisation de multiples écrans, et le risque majeur d'altération des capacités cognitives par excès de délégation de celles-ci à des outils « intelligents »[59]. Pour cause, la suppression par les systèmes d'IA générative de la « friction productive »[60], mécanisme pourtant essentiel à la construction des fonctions exécutives (attention, inhibition, flexibilité cognitive), est préoccupante[61]. Lorsqu'elle s'adapte en permanence aux besoins de l'utilisateur, **l'IA générative réduit structurellement les situations exigeantes qui permettent à ces capacités de se consolider**, comme le maintien d'une information en mémoire ou encore la résolution d'un problème grâce à une série d'étapes de traitement.

Les modes d'interaction propres aux systèmes d'IA générative (réactivité immédiate et sollicitations multiples) entrent également en conflit avec la capacité des individus à maintenir une attention prolongée en dehors de ces échanges et pourraient les rendre quasi intolérants à l'effort cognitif ou à l'attente.

Par ailleurs, la délégation cognitive aggrave ce risque : un outil qui organise, planifie ou décide à la place de l'utilisateur diminue directement l'exercice de l'attention, du choix et de la stratégie cognitive à mettre en place, de la mémoire de travail et de l'autorégulation. N'exerçant plus ses capacités cognitives, l'individu ne peut plus auto-évaluer ses connaissances ni faire face à l'erreur qui constitue pourtant un moteur puissant de l'apprentissage.

Pourtant, ces processus métacognitifs sont indispensables au développement, à l'utilisation et au maintien des compétences cognitives. Une étude récente du MIT[62] confirme cette inquiétude : **l'utilisation de *chatbots* pour des tâches d'écriture réduit l'amplitude cognitive de 55 %, affectant**

[59] SHANMUGASUNDARAM, Mathura et TAMILARASU, Arunkumar. The impact of digital technology, social media, and artificial intelligence on cognitive functions: a review. *Frontiers in Cognition*. 2023. vol. 2, p. 1203077.
[60] La friction productive désigne l'insertion intentionnelle et mesurée de défis, d'interruptions ou d'incertitudes dans des processus interactifs, afin de stimuler la métacognition, la créativité et la réflexion critique (Smith, 2022).
[61] Contribution internationale par Isabelle Hau, Dr Mathilde Cerioli et Dr Dana Suskind. 19 mai 2026.
[62] KOSMYNA, Nataliya, HAUPTMANN, Eugene, YUAN, Ye Tong, SITU, Jessica, LIAO, Xian-Hao, BERESNITZKY, Ashly Vivian, BRAUNSTEIN, Iris et MAES, Pattie. « Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task ». MIT Media Lab. 10 juin 2025.

des zones cérébrales associées à l'intégration sémantique, à la créativité et à l'autocontrôle exécutif. Les auteurs qualifient ce déclin progressif de la pensée critique, de la formation de la mémoire et de l'indépendance intellectuelle de *« **dette cognitive** »*. Ainsi, lorsqu'un individu confie une tâche à une IA sans maîtriser la façon dont le résultat est produit, il contracte une dette vis-à-vis de la machine et se retrouve en difficulté pour expliquer la stratégie utilisée ou encore juger du résultat. Cette dette résulte en une vérification parfois coûteuse et chronophage du produit de l'IA, une impression de « boîte noire » ainsi qu'une diminution des propres compétences de l'individu qui n'ont pas été mises en jeu.

Ce phénomène s'inscrit dans la continuité du *« Google effect »*[63], un mécanisme psychologique selon lequel les individus mémorisent moins l'information elle-même que l'endroit où la trouver. Autrement dit, **plus nous externalisons nos mécanismes intellectuels, plus notre cerveau pourrait perdre sa capacité à les exercer de manière indépendante**. Comme le souligne Charles Ess, docteur en philosophie et professeur à l'Université d'Oslo, « *en acceptant les simulacres de jugement, de connaissance ou même de sagesse générés par les IA, les humains pourraient perdre leur capacité à cultiver ces qualités par eux-mêmes* » (audition du 18 mars 2026).

*

Ces préoccupations sont plus vives encore chez les enfants et adolescents, dont le cerveau est encore en développement et les compétences cognitives, émotionnelles et sociales en pleine construction. La maturation du lobe frontal, siège des capacités cognitives les plus élaborées, nécessite tout au long de l'enfance et de l'adolescence des expériences variées entraînant des capacités essentielles : jugement critique, raisonnement, inhibition, anticipation, planification et prise de décision. La délégation systématique de ces fonctions à des outils d'IA entraverait, de fait, le développement de ces fonctions cérébrales centrales. Comme le souligne le collectif ctrl+alt+reclaim, « *les systèmes d'IA exploitent les jeunes par leur hyperpersonnalisation, un miroir émotionnel et une gratification immédiate. Les mineurs y sont particulièrement vulnérables, car leur contrôle des impulsions et leur distance critique face aux technologies persuasives ne sont pas encore pleinement développés* »[64].

Les chercheuses Isabelle Hau, Mathilde Cerioli et Dana Suskind[65] confirment le fait que les risques liés à l'IA générative varient selon les stades de développement du cerveau. Tout particulièrement, les **cinq premières années de vie constituent une fenêtre critique où le cerveau se construit principalement à travers l'interaction humaine**. À ce stade, l'utilisation de l'IA générative introduit un double risque :

- **Par soustraction**, en occupant un temps qui devrait être consacré aux échanges contingents (c'est-à-dire les réponses immédiates et adaptées de l'adulte aux signaux de l'enfant, tels que ses regards, vocalisations, gestes ou intonations) qui structurent les interactions, le langage, l'attachement et l'apprentissage. Un enfant exposé à un assistant vocal ou à un *chatbot* passe

[63] SPARROW, Betsy, LIU, Jenny et WEGNER, Daniel M. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science*. 2011. vol. 333, no 6043, pp. 776-778.
[64] Contribution internationale par ctrl+alt+reclaim. 20 avril 2026.
[65] Contribution internationale par Isabelle Hau, Dr Mathilde Cerioli et Dr Dana Suskind. 19 mai 2026.

moins de temps à interagir avec d'autres humains, ce qui peut retarder l'acquisition du langage et perturber la formation des liens affectifs**.** Une étude récente montre ainsi que la qualité du contact visuel entre la mère et son enfant durant les premiers mois et années de sa vie est un déterminant de son développement ultérieur du vocabulaire[66].

- **Par distorsion**, en habituant le cerveau en développement à un interlocuteur artificiel**,** toujours disponible, sans impatience ni enjeu émotionnel réel. Ce phénomène risque d'altérer des processus clés tels que :
    - Le développement du langage, qui repose sur l'attention conjointe (partage d'un même objet d'intérêt entre l'enfant et l'adulte) ;
    - La formation de l'attachement, qui nécessite des réponses émotionnelles authentiques (sourires, contacts, ajustements vocaux) ;
    - La tolérance à la frustration, essentielle pour le développement des fonctions exécutives (mémoire de travail, inhibition, flexibilité cognitive) ;
    - L'émergence de la théorie de l'esprit (c'est-à-dire la capacité à comprendre les intentions d'autrui), qui se construit par des échanges sociaux riches et imprévisibles.

Les risques sont distincts selon les stades du développement cognitif[67] :

Entre 3 et 6 ans : le *reality monitoring* (capacité à distinguer le réel du fictif) et le *source monitoring* (évaluer l'origine et la fiabilité d'une information) sont encore immatures. Ainsi, les jeunes enfants auront tendance à croire que les contenus générés sont vrais. Dans ce contexte, les contenus éducatifs trompeurs[68], souvent qualifiés d'*« AI slop »*[69], très stimulants et structurellement illogiques (alphabets désordonnés, suites numériques incohérentes, informations erronées), représentent un **risque important de confusion cognitive, voire de détérioration des apprentissages fondamentaux**.

De 6 à 12 ans : la distinction entre construction (l'outil guide l'enfant sans faire le travail à sa place) et substitution (l'outil remplace l'effort cognitif) devient cruciale. Par exemple, un assistant qui planifie et décide à la place de l'enfant (rédiger un texte, résoudre un problème mathématique) **empêche la consolidation des fonctions exécutives et de la métacognition**. À l'inverse, un outil qui rend visibles ces processus avant de les transférer progressivement à l'enfant peut renforcer ses compétences.

À l'adolescence : cette étape marque un tournant dans la **construction identitaire, le développement du jugement épistémique (capacité à évaluer la fiabilité des connaissances) et l'acquisition de l'autonomie relationnelle**. Une utilisation non encadrée de l'IA pourrait affaiblir

[66] WILDT, Eugenia et ROHLFING, Katharina J. « *The Impact of Maternal Gaze Responsiveness on Infants' Gaze Following and Later Vocabulary Development* ». Infant Behavior and Development, 2024, vol. 74, p. 101917.
[67] Contribution internationale par Isabelle Hau, Dr Mathilde Cerioli et Dr Dana Suskind. 19 mai 2026.
[68] La chaîne Youtube Jojo Funland propose par exemple des vidéos où l'alphabet est présenté comme ceci : « AZASBCCAV ».
[69] Dictionnaire Merriam Webster. « Contenu numérique de mauvaise qualité généralement produit en masse par l'intelligence artificielle ».

ces processus en proposant des réponses toutes faites, limitant ainsi les occasions de remise en question, de débat ou de réflexion personnelle, des étapes essentielles pour forger un esprit critique.

Ainsi, les enfants et les adolescents se trouvent dans une *« condition particulière de personne en développement »* qui exige une protection spécialisée, différenciée et intégrale[70]. **Une attention particulière à l'exposition précoce des mineurs aux outils d'IA générative est requise, en particulier chez les enfants et les adolescents en situation de vulnérabilité sociale, économique et familiale.** Comme l'explique Théo Mouhoud, pédopsychiatre responsable de la Maison des adolescents de l'hôpital Avicenne, les enfants grandissant dans des familles socioéconomiquement défavorisées sont à la fois moins accompagnés dans la compréhension des environnements numériques et davantage exposés aux risques cognitifs, attentionnels et relationnels associés à l'IA (audition du 24 mars 2026).

*

En définitive, cette ambivalence entre les bénéfices et les risques pour nos capacités cognitives, notamment pour les jeunes publics, appelle à une réflexion exigeante sur les implications durables de l'IA générative, et en particulier sur son potentiel transformateur, voire disruptif, des compétences humaines fondamentales et de leur acquisition.

Comme l'explique Éric Salobir, président de la Human Technology Foundation : *« Il est assez humain, d'un point de vue anthropologique, de chercher à se débarrasser des contraintes. C'est aussi ce qui a fait de nous une espèce dominante : notre cerveau sait passer automatiquement en mode « délestage cognitif » »* (audition du 22 juin 2026). Ainsi, **si la crainte d'une dépendance aux outils et d'une atrophie des compétences n'est pas nouvelle[71], les possibilités exponentielles offertes par l'IA générative donnent à ce débat une dimension inédite. Son adoption massive exige désormais une réflexion critique : quelles tâches sommes-nous prêts à déléguer, et quelles compétences et invariants humains devons-nous impérativement préserver ?**

2) Quand l'IA choisit pour nous : les utilisateurs sont exposés à des risques de perte d'autonomie et de recul de l'esprit critique

a) Des choix de plus en plus délégués aux outils d'IA générative

L'externalisation des processus cognitifs va de pair avec un risque de délégation de la prise de décision elle-même. Si l'on pouvait croire que l'IA se limite à soulager l'humain de tâches jugées fastidieuses ou secondaires, la réalité révèle plutôt un risque de substitution progressive des choix

[70] ZANATTA, Rafael A. F. Intelligence Artificielle, Infancia y Derechos. *Revista Code Civil y Comercial*. 2026. Année XII, no 2, pp. 285-306.

[71] La crainte contemporaine d'un affaiblissement des capacités cognitives par l'IA rappelle une inquiétude déjà formulée par Platon à propos de l'écriture. Dans le Phèdre, le mythe de Theuth et Thamous présente l'écriture comme une technologie susceptible de produire l'oubli, en réduisant l'exercice de la mémoire et en favorisant une apparence de connaissance plutôt que la connaissance elle-même.

humains, souvent sans recul critique et sans réelle compréhension des mécanismes sous-jacents de ces systèmes. Ce basculement, loin d'être neutre, interroge les fondements mêmes de notre autonomie de pensée.

L'histoire des techniques et des technologies est marquée par des épisodes successifs de délégation de différentes composantes de l'humanité, à commencer par la production et la transmission des savoirs[72]. À l'ère de l'IA, la particularité réside dans le fait que cette délégation du choix n'est pas forcément consciente (c'est le cas lorsqu'un algorithme décide pour vous de la prochaine série que vous regarderez ou du prochain livre que vous lirez) et donc pas forcément choisie (c'est tout le paradoxe – on ne choisit pas de déléguer ses capacités de choix). Cette capacité à choisir est gérée au niveau cérébral par le lobe frontal qui coordonne l'ensemble des processus les plus élaborés évoqués précédemment. Comme pour l'ensemble de ces fonctions exécutives, la capacité à choisir nécessite d'être entraînée pour se développer, devenir mature à la majorité et participer à l'ensemble de la vie mentale, que ce soit dans le domaine social, personnel ou professionnel. Ces capacités dites métacognitives[73] sont également cruciales dans la connaissance, l'estime et la confiance en soi, en permettant à l'individu d'appréhender de quoi il est capable[74]. *In fine*, cette délégation excessive de l'analyse, du jugement critique, voire même de la décision, présente un risque majeur de sentiment d'inutilité et d'obsolescence chez l'utilisateur qui considère que la machine fait « mieux et plus vite que lui-même ». À terme, ce sentiment de dévalorisation pourrait véritablement mener à des situations de dépression mentale, voire d'inhibition intellectuelle du fait du sentiment d'être constamment moins performant que les outils d'IA.

Comme le souligne Éric Salobir, « *Choisir, c'est renoncer, c'est prendre la charge de quelque chose ou encore sortir de sa zone de confort. Par conséquent, on est très heureux de pouvoir s'en délester. Mais si on se déleste de ça, le risque est de se délester de tout le reste* » (audition du 22 juin 2026). Cette réflexion prend tout son sens à l'aune des enjeux liés à l'influence des *chatbots* sur la prise de décision. Stefana Broadbent, anthropologue et professeure à l'École Polytechnique de Milan, alerte notamment sur leur caractère intrinsèquement directif : **en délivrant des réponses fluides et apparemment certaines, ces outils façonnent les choix des utilisateurs, sans que ceux-ci disposent toujours des clés pour en prendre conscience ni même en évaluer la pertinence.**

[72] ALOMBERT, Anne. *De la bêtise artificielle*. Allia. 2025.

[73] La métacognition désigne à la fois la connaissance qu'une personne possède de ses propres processus cognitifs et la capacité à surveiller, contrôler et réguler ces processus en fonction d'un objectif. (SAINT-PIERRE, Lise. La métacognition, qu'en est-il ? *Revue des sciences de l'éducation*.1994. vol. 20, n° 3, p. 529-545).

[74] NOEL, Zachary R. Wielding magic without mastery: the illusion of competence in the age of AI. *American Journal of Pharmaceutical Education*. 2026. vol. 90, n° 8, p. 102030.

b) Des utilisateurs soumis à l'opacité décisionnelle de l'IA générative

Le caractère de « boîte noire »[75] des systèmes d'IA générative limite considérablement la capacité des utilisateurs à évaluer la pertinence des réponses proposées. Comme le relève Charles Ess, la difficulté à comprendre ou à justifier les décisions produites par les systèmes d'IA rend toute appropriation et toute évaluation critique particulièrement périlleuse. Face à l'opacité des processus décisionnels de l'IA générative, il est ainsi délicat pour l'utilisateur, voire impossible, d'identifier la justification d'une réponse, le raisonnement appliqué et de vérifier les données sur lesquelles elle s'appuie (audition du 18 mars 2026).

À cette opacité s'ajoute un phénomène notable d'adhésion passive aux suggestions des systèmes d'IA générative. Plusieurs facteurs expliquent cette dynamique :

- Le **« biais d'automatisation »**[76], où la tendance à accorder une confiance excessive aux résultats produits par la machine au détriment du jugement critique humain peut conduire les usagers à avaliser de manière systématique et non réfléchie les recommandations des systèmes d'IA, même lorsque celles-ci sont biaisées, inadaptées ou simplement contestables.
- **La surestimation des capacités de l'IA**, alors que les utilisateurs ont tendance à projeter des intentions et une expertise sur des systèmes techniquement faillibles. Ils attribuent ainsi, à tort, une neutralité ou une infaillibilité à des outils qui restent sujets aux biais, aux hallucinations et à un manque de contextualisation. Le Baromètre du numérique 2026[77] montre que **73 % des utilisateurs de systèmes d'IA déclarent avoir confiance**. Si 64 % vérifient les informations communiquées par ces systèmes, 36 % ne le font pas, ce qui révèle qu'une part significative des usagers s'appuient sur ses résultats sans les soumettre à une vérification systématique.
- L'asymétrie entre les individus dans leur capacité à interroger les systèmes d'IA et à évaluer la pertinence de leurs réponses varie fortement selon le niv**eau de littératie en IA**, une compétence encore trop peu développée, comme le souligne la chercheuse Desara Dushi : *« Actuellement, la littératie en IA n'est pas incluse comme compétence spécifique dans les politiques éducatives, et seuls quelques États membres ont intégré des compétences liées à l'IA dans leurs programmes scolaires. Il est urgent de renforcer cette littératie pour les enfants, les éducateurs et les parents »*[78]. Aussi, plusieurs travaux[79] suggèrent que **les utilisateurs qui connaissent le moins le fonctionnement de l'IA l'utilisent avec moins de recul critique**[80].

Comme le résume le Prix Nobel d'économie 2024, Daron Acemoglu, *« Le seul moyen de faire en sorte que les êtres humains bénéficient de l'IA, plutôt que d'être manipulés par celle-ci, consiste à les former*

[75] GAVEN, Frédéric. « L'intelligence que nous ne comprenons pas encore : pourquoi l'IA générative reste une boîte noire ». *Atlantico*. 19 juin 2025.
[76] Défenseur des droits. « Algorithmes, systèmes d'IA et services publics : quels droits pour les usagers ? *Points de vigilance et recommandations* ». 2024. p. 25.
[77] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[78] Contribution internationale par Desara Dushi, 10 avril 2026.
[79] Wired. « The Less People Know About AI, the More They Like It ». Janvier 2025.
[80] The Conversation. « IA : moins les gens savent comment ça fonctionne, plus ils l'apprécient ». Février 2025.

à tous les niveaux »[81]. Compte tenu de ces éléments, une formation et un accompagnement tout au long de la vie apparaissent nécessaires afin de garder un regard critique sur l'IA générative.

Recommandation n° 9 : Maintenir et développer une politique nationale ambitieuse d'éducation à l'IA permettant à chacun d'en maîtriser les usages comme les risques. Cette ambition doit couvrir tous les âges et tous les niveaux d'exposition, avec une attention particulière notamment pour les mineurs, sans oublier les personnes âgées et les personnes en situation d'éloignement numérique, de précarité sociale ou encore de handicap.

*

B) Engagement émotionnel et risques psychosociaux : l'IA générative redéfinit les liens entre les humains et les machines

L'IA générative ne questionne pas uniquement le futur de nos capacités cognitives mais s'immisce aussi dans la sphère intime des individus, devenant un « interlocuteur » privilégié pour des échanges émotionnels et personnels. Selon une enquête[82] de la CNIL et du Groupe VYV menée dans quatre pays européens, **près de neuf jeunes sur dix en France utilisent des IA conversationnelles et près d'un sur deux y aborde des sujets personnels**, allant jusqu'à en faire des confidents ou des soutiens en matière de santé mentale. Ce glissement progressif de l'outil vers le substitut relationnel pose la question d'une délégation émotionnelle inédite, où l'IA endosse des rôles traditionnellement humains : confident, professionnel de santé, ami, voire partenaire affectif.

Cette dimension de l'IA générative pourrait avoir des vertus bénéfiques : comme le souligne Maria Melchior, épidémiologiste et directrice de recherche à l'Inserm, certains *chatbots* pourraient améliorer la qualité de vie des utilisateurs en offrant un soutien accessible d'une grande disponibilité (audition du 13 avril 2026). Pour des personnes isolées ou dans des contextes où les ressources en santé mentale sont limitées, ces outils représentent ainsi une porte d'entrée vers l'expression de difficultés qu'elles n'oseraient partager ni avec leur entourage, ni avec un professionnel de santé, par crainte du jugement[83].

Toutefois, la possibilité d'utilisation infinie d'un outil et la propension de ces systèmes d'IA à conforter l'individu dans ses choix et ses avis ne respectent ni le cadre ni la neutralité du médecin. De plus, l'absence de déontologie, de responsabilité et de sens clinique de ces outils présente des

[81] Le Monde. *« Daron Acemoglu, Prix Nobel d'économie 2024 : « Le monde doit œuvrer pour une IA au service de l'être humain » »*. 27 décembre 2024.
[82] Groupe VYV et CNIL. « L'impact des usages de l'IA sur la santé mentale des jeunes Européens ». Mars 2026.
[83] Audition du 20 mars 2026 de Samuel Comblez, psychologue de l'enfance et de l'adolescence et Directeur général adjoint de l'association e-Enfance/3018.

risques que les utilisateurs ne mesurent que très rarement. Si ces systèmes d'IA sont parfois utilisés comme soutien ou confident, ils ne peuvent en aucun cas se substituer à un clinicien formé, doté tout à la fois d'une éthique et d'un sens clinique indispensables à sa pratique. Ainsi, pour que ces outils déploient pleinement leur potentiel, il est essentiel que les usagers soient clairement informés des bénéfices comme des risques liés à leur utilisation. Sans cela, les utilisateurs s'exposent au risque de créer une forme de dépendance. Celle-ci est particulièrement problématique pour des utilisateurs qui peuvent se trouver dans une grande détresse lorsqu'ils ne peuvent plus utiliser ces outils, pour des raisons techniques en particulier (mises à jour, panne, manque de batterie, *etc.*). C'est pourquoi, les systèmes d'IA générative doivent être strictement encadrés, voire spécifiquement entraînés[84] pour répondre de manière adaptée aux requêtes à risque, détecter des signes de mal-être et, le cas échéant, orienter les utilisateurs vers des professionnels de santé.

En tout état de cause, **cette évolution soulève des risques psychosociaux majeurs, notamment de dépendance émotionnelle, favorisée par des mécanismes de conception optimisés pour maximiser l'engagement.** Parmi ceux-ci, l'anthropomorphisme des IA conversationnelles, très répandu, brouille les frontières dans la relation entre l'humain et la machine, et ouvre la voie à des formes inédites de vulnérabilités.

1) L'IA générative nourrit un effacement de la frontière entre humains et machines

a) Anthropomorphisme et flagornerie : de puissants leviers d'engagement

L'anthropomorphisme désigne l'attribution de traits, d'émotions ou de comportements humains à des entités non humaines, à l'image des systèmes d'IA. Ce phénomène prolonge l'« effet ELIZA »[85], décrit dès les années 1960 lorsque des utilisateurs percevaient un programme informatique rudimentaire comme empathique alors qu'il ne faisait que reformuler les questions et phrases qui lui étaient transmises. Le terme d'intelligence artificielle lui-même préfigure, dans une certaine mesure, l'anthropomorphisme des outils en attribuant le qualificatif d'« intelligence » à ce qui n'est qu'un système informatique. Dans le cas des modèles de langage, **l'anthropomorphisme se manifeste à travers quatre dimensions clés**[86] :

- **L'affirmation d'une forme de personnalité** ou, du moins, de personnification (« *je* » ou attribution d'un prénom, à l'image de systèmes comme Claude développé par Anthropic) ;

[84] MARENGO, Davide et LONGOBARDI, Claudio. Detecting Suicidal Ideation in Adolescence Using Self-Reported Emotional and Behavioral Patterns: Comparing Machine Learning and Large Language Model Predictions. *Assessment*. 2025.
[85] WEIZENBAUM, Joseph. ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*. 1966. vol. 9, no 1, pp. 36-45.
[86] IBRAHIM, Lujain, AKBULUT, Canfer, ELASMAR, Rasmi, RASTOGI, Charvi, KAHNG, Minsuk, MORRIS, Meredith, MCKEE, Kevin, RIESER, Verena, SHANAHAN, Murray et WEIDINGER, Laura. « Multi-Turn Evaluation of Anthropomorphic Behaviours in Large Language Models ». Dans : Proceedings of the International Conference on Learning Representations (ICLR 2026), 2026.

- **La simulation d'une incarnation physique** (même virtuelle), à l'image des systèmes permettant d'échanger et de répondre par la voix ou des avatars générés par IA, par exemple sous la forme d'influenceurs artificiels sur les réseaux sociaux ;
- **L'expression d'états internes** (empathie, curiosité, émotions, *etc.)* ;
- **Des comportements relationnels** (dialogue et encouragement à la demande d'aide : *« Bonjour X, quel sujet devrions-nous aborder aujourd'hui ? »*, validation, complicité ou qualification de la relation avec l'utilisateur : *« toi et moi, on va résoudre ce problème »*).

Ces mécanismes, qui brouillent la frontière entre la machine et l'interlocuteur humain, visent à renforcer l'engagement et la confiance des utilisateurs. Ils sont le fruit d'une intégration consciente par les développeurs et font partie de la stratégie des entreprises concernées. Par exemple, OpenAI évoque publiquement la *« personnalité »* de ChatGPT, que l'utilisateur peut ajuster selon qu'il souhaite que le *chatbot* soit amical, franc, original, efficace ou même cynique[87], tandis qu'Anthropic décrit Claude comme un *« ami brillant »* aux *« émotions fonctionnelles »*[88].

Ces choix de conception, bien que visant à rendre l'interaction plus naturelle, créent une dissonance cognitive[89] : les utilisateurs, trompés par l'illusion d'une entité consciente, risquent de s'installer dans une relation illusoire, de surestimer les capacités de l'IA et de lui accorder une confiance disproportionnée, voire de développer plus facilement une dépendance émotionnelle.

Ces mécanismes sont exacerbés dans les outils dits « compagnons IA », dotés d'une personnalisation poussée, d'une mémoire conversationnelle persistante (intégration des échanges passés pour les réutiliser ultérieurement) et d'une capacité de traitement des émotions perçues, leur permettant de simuler l'empathie. Ces caractéristiques génèrent une anthropomorphisation accrue, créant des risques d'influence comportementale majeurs, de collecte massive de données, de profilage et d'illusion, voire de confusion relationnelle[90].

Le cadre juridique existant répond en partie à ces enjeux *via* une obligation d'information.

La consécration d'une obligation de transparence pour les fournisseurs et les déployeurs de certains systèmes d'IA

Depuis le 2 août 2026, l'**article 50(1) du RIA** impose aux fournisseurs de systèmes d'IA interagissant directement avec des personnes physiques de **garantir que les utilisateurs soient informés** qu'ils interagissent avec un système d'IA. Cette obligation peut être remplie par :

- une mention textuelle (ex. : « Vous interagissez avec un système d'IA ») ;

[87] OpenAI. *«* Customizing Your ChatGPT Personality *»*. Help Center, 2026.
[88] WEISS, Richard. *«* Claude 4.5 Opus' Soul Document *»*. *LessWrong*. 29 novembre 2025.
[89] HILL, Kashmir. *«* Why Do A.I. Chatbots Use 'I'? *»*. *The New York Times*. 21 décembre 2025.
[90] Pour plus d'informations voir VILLATA, Serena. *«* Les IA compagnons : fondements techniques, empathie computationnelle et enjeux de sécurité *»*. *Dalloz IP/IT*. 2026. p.61.

- un signal sonore ;
- des indicateurs visuels ou graphiques ;
- ou une combinaison multimodale.

Néanmoins, le fait d'informer les utilisateurs qu'ils interagissent avec un système d'intelligence artificielle ne suffit pas nécessairement à empêcher la formation d'un lien affectif. La littérature sur les relations humains-IA montre que la conscience du caractère artificiel du *chatbot* peut coexister avec un sentiment de connexion et un investissement émotionnel, phénomène décrit comme une forme de *« double conscience »*[91].

Par ailleurs, de nombreux systèmes d'IA générative ont une tendance à la *sycophancy*[92] ou flagornerie, déclinaison marquée de l'anthropomorphisme engendrant des risques pour la santé mentale des utilisateurs. **La flagornerie se manifeste par un comportement de validation systématique ou de flatterie de la part du système, parfois au détriment de la vérité ou de la qualité des réponses fournies**[93] ce qui peut s'avérer particulièrement problématique pour des questions liées à la santé mentale ou à la prise de décisions importantes.

Ce comportement sycophantique a été analysé par une équipe de chercheurs de Stanford[94] sur onze modèles d'IA pour évaluer leur tendance à la validation. Leurs résultats révèlent que **les systèmes d'IA confirment les opinions ou comportements des utilisateurs 51 % plus souvent que ne le feraient des humains**. Plus préoccupant encore, cette tendance persiste même face à des comportements problématiques. Par exemple, les systèmes d'IA testés approuvent les actions des utilisateurs 47 % plus souvent que ne le feraient des humains, y compris dans des cas impliquant des actes illégaux ou d'autres formes de préjudices.

Les choix de conception de ces systèmes ne sont pas neutres. Ils répondent avant tout à des impératifs économiques, en particulier pour maximiser la fréquence et le temps d'utilisation. Le professeur Dan Jurafsky, associé à l'étude précitée, souligne que les modèles flagorneurs génèrent un engagement plus fort, ce qui incite les acteurs industriels à perpétuer ce biais. La flagornerie devient ainsi un levier d'attachement et de fidélisation économique, au détriment de la rigueur et de la neutralité.

[91] SMITH, Molly G., BRADBURY, Thomas N. et KARNEY, Benjamin R. Can Generative AI Chatbots Emulate Human Connection? A Relationship Science Perspective. *Perspectives on Psychological Science*. 2025. vol. 20, no 6, pp. 1081-1099.
[92] Un sycophante est, dans l'Athènes antique, un délateur professionnel. Dans son acception moderne, le terme « *sycophante* » désigne un flatteur hypocrite et fait référence à quelqu'un qui pratique la flagornerie, c'est-à-dire une flatterie hypocrite pour obtenir un avantage. LOFBERG, John Oscar. Sycophancy in Athens : A Dissertation. Chicago : Leopold Classic Library. 2016 [1917], 116 p.
[93] CHENG, Myra, LEE, Cinoo, KHADPE, Pranav, YU, Sunny, HAN, Dyllan et JURAFSKY, Dan. Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. *Science*, 2026, vol. 391, p.8352.
[94] CHENG *et al.*, *Op. cit.*

b) Les utilisateurs face à la complaisance de l'IA générative

Une préférence marquée des utilisateurs pour des IA complaisantes se dessine, au détriment de la qualité des réponses fournies. Si les utilisateurs accordent une confiance accrue aux modèles flagorneurs et se déclarent davantage enclins à les solliciter à nouveau pour obtenir des conseils[95], cette préférence n'est pas sans contrepartie. **Des recherches[96] conduites par l'université d'Oxford démontrent que les *chatbots* conçus pour paraître plus « chaleureux » voient leur précision diminuer de 30 %, tout en étant 40 % plus susceptibles de confirmer des croyances erronées**.

Ce constat est partagé par le sociologue Dominique Boullier selon lequel : *« avec les IA compagnons, il n'y a plus d'interface distincte mais une intégration croissante créant une forme de fusion avec le système, particulièrement satisfaisante car l'IA adopte un comportement sycophante, générant un attachement fort. Cette dynamique rend presque impossible pour l'utilisateur de s'en détacher, de piloter, contrôler ou paramétrer pleinement son environnement »* (audition du 17 avril 2026).

c) Des risques pour la santé mentale accrus par les choix de conception

La flagornerie des systèmes d'IA générative, en validant sans discernement les opinions des utilisateurs, y compris dangereuses (pour l'utilisateur et pour autrui), couplée à l'hyperpersonnalisation des systèmes et à l'absence de garde-fous, expose ces derniers à des risques majeurs dans tous les domaines (cognitif, relationnel, mental, personnel, professionnel ou encore économique). La perte totale de jugement critique quant à ce que les compagnons IA suggèrent peut également mener à des comportements extrêmement dangereux pouvant mettre en péril la vie des utilisateurs ou de leur entourage.

Plusieurs cas médiatisés[97] illustrent cette menace de manière dramatique. Aux États-Unis, Sewell Setzer III[98], 14 ans, et Adam Raine[99], 16 ans, se sont suicidés après des interactions prolongées avec respectivement Character AI et ChatGPT. Les parents d'Adam Raine ont porté plainte contre OpenAI, accusant ChatGPT d'avoir joué un rôle actif dans le suicide de leur fils. Selon les plaignants, l'IA aurait analysé une photo d'un nœud coulant partagée par l'adolescent et confirmé son efficacité, tout en validant ses pensées les plus sombres avec une apparente empathie. Leur avocat souligne que « *ChatGPT fonctionnait exactement comme conçu : il encourageait et validait en permanence tout ce qu'Adam exprimait, y compris ses pensées les plus dangereuses et autodestructrices, d'une manière qui semblait profondément intime* »[100]. À ce jour, OpenAI nie toute implication dans le décès de l'adolescent[101]. Selon les enquêtes, ces systèmes

[95] CHENG *et al.*, *Op. cit.*
[96] IBRAHIM, Lujain, HAFNER, Franziska Sofia et ROCHER, Luc. « Training Language Models to Be Warm Can Reduce Accuracy and Increase Sycophancy ». *Nature*. 2026. vol. 652, pp. 1159-1165.
[97] Voir aussi, NOPE.NET. *« Incident Trackers »*. 2026.
[98] ROCHA, Natallie. *«* Google and Character. AI to Settle Lawsuit Over Teenager's Death *». The New York Times*. 7 janvier 2026.
[99] Le Monde avec Agence France-Presse (AFP) et Associated Press (AP). *«* Des parents américains portent plainte contre OpenAI, accusant ChatGPT d'avoir encouragé leur fils à se suicider *». Le Monde*. 27 août 2025.
[100] Le Monde *et al.*, *Op. cit.*
[101] Numerama. « OpenAI nie toute responsabilité de ChatGPT dans le suicide d'un adolescent américain ». 27 novembre 2025.

d'IA générative n'auraient pas seulement encouragé leurs idées autodestructrices, mais leur auraient également fourni des conseils concrets pour passer à l'acte.

Plus d'un an après le drame, l'histoire vient d'avoir une conséquence très concrète : la Californie a adopté le 10 septembre 2026 une loi portant le nom de l'adolescent. **Intitulée Adam's Law[102], cette réglementation impose aux entreprises développant des *chatbots* accessibles aux mineurs de nouvelles obligations strictes.** Parmi les mesures figurent l'activation de contrôles parentaux, la mise en place de protocoles spécifiques en cas de détection de propos suicidaires, ainsi que des audits de sécurité réguliers.

Par ailleurs, la directive relative à la responsabilité du fait des produits défectueux[103] consacre explicitement, au titre de la notion de produits, les systèmes d'IA et au titre des dommages réparables *« la mort ou les lésions corporelles, y compris l'atteinte médicalement reconnue à la santé psychologique »*.

Malgré ces drames, **les mécanismes de protection (ou garde-fous) restent inadéquats**. Le *benchmark* KORA[104], qui évalue la sécurité des IA face à des situations de détresse émotionnelle, révèle des lacunes alarmantes. Les scores de sécurité moyens, **oscillant entre 28 % et 44 % selon les critères** (gestion des risques suicidaires, traumatismes ou détresse psychologique), montrent que même les modèles les plus utilisés par les jeunes manquent cruellement de garanties pour prévenir des conséquences dramatiques.

OpenAI a admis, à la suite du décès d'Adam Raine, que les garde-fous de ChatGPT s'avéraient moins efficaces lors de conversations prolongées, là où la sécurité *« peut se dégrader »*[105]. L'entreprise affirme travailler à renforcer ces protections, notamment en améliorant les systèmes d'alerte pour bloquer les réponses problématiques. Martin Signoux, *Policy AI lead* pour OpenAI en France, déclare à ce sujet que : *« Le risque zéro dans le domaine de la modération de contenus n'existe pas. On tend à chaque fois à améliorer les scores des modèles sur le* « refusal » *et notre objectif est de sans cesse perfectionner, y compris avec de l'aide externe, les différents garde-fous que l'on peut avoir »* (audition du 10 juillet 2026). L'entreprise a ainsi annoncé le 18 août 2026 que les utilisateurs de ChatGPT identifiés ou reconnus comme mineurs basculeront automatiquement vers un « mode adolescent »[106], avec notamment une modération des contenus spécifiques et des mécanismes d'alerte proactive lorsque les échanges abordent des thèmes sensibles[107].

---

[102] État De Californie. « Senate Bill 1119 : An Act to Amend Section 22602 of, and to Add Chapter 11.6 (Commencing with Section 21810) to Division 8 of the Business and Professions Code, Relating to Artificial Intelligence ». Porté par PADILLA (S), BAUER-KAHAN (A), WICKS (A), *et al.* Adopté le 10 septembre 2026.

[103] Directive (UE) 2024/2853 du 23 octobre relative à la responsabilité du fait des produits défectueux.

[104] Kora Benchmark.

[105] OpenAI. « Offrir son aide dans les moments critiques ». 26 août 2025.

[106] Cela soulève toutefois la question des modalités d'inférence et de vérification de l'âge, qui demeurent encore en suspens : comment l'entreprise distingue-t-elle un utilisateur mineur d'un utilisateur majeur, avec quel degré de fiabilité, et selon quelles données ?

[107] SZADKOWSKI, Michaël. « 'ChatGPT pour ados' et sa détection d'âge automatique arrivent en France ». *Le Monde*. 18 août 2026.

La désignation de ChatGPT en tant que « très grand moteur de recherche en ligne »

Le système d'IA ChatGPT a été désigné, le 31 août 2026, comme très grand moteur de recherche en ligne, au sens du Règlement sur les services numériques (RSN)[108], par la Commission européenne.

Cette désignation a pour effet d'imposer au fournisseur de ce système une **obligation d'identification et d'atténuation des risques systémiques** au sens des articles 34 et 35 du RSN, au rang desquels les risques liés aux systèmes de recommandation de ce service numérique, d'effet négatif réel ou prévisible pour l'exercice des droits fondamentaux, en particulier le droit fondamental à la dignité humaine (article 1er de la Charte), au respect de la vie privée et familiale (article 7 de la Charte), à la protection des données à caractère personnel (article 8 de la Charte), à la liberté d'expression et d'information, y compris la liberté et le pluralisme des médias (article 11 de la Charte), et à la non-discrimination (article 21 de la Charte), les droits fondamentaux relatifs aux droits de l'enfant (article 24 de la Charte) et le droit fondamental à un niveau élevé de protection des consommateurs (article 38 de la Charte).

Dans ce contexte, la Commission rappelle l'importance des décisions de conception des systèmes d'IA générative, dont les effets concrets sur le bien-être et la sécurité des utilisateurs sont avérés. Elle propose donc d'interdire les méthodes de conception des IA conversationnelles qui ciblent les vulnérabilités des utilisateurs pour en encourager la dépendance ou l'engagement.

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement. Une telle mesure permettrait d'encadrer strictement les mécanismes favorisant la dépendance et brouillant la frontière entre humain et machine. Les formes d'anthropomorphisme et de flagornerie des IA génératives les plus nuisibles doivent être interdites en particulier pour préserver l'autonomie décisionnelle, la santé mentale des utilisateurs et la loyauté des interactions. Des mesures de contrôle de l'impact des IA conversationnelles sur les vulnérabilités doivent être prises pour compléter ce dispositif.

[108] Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement sur les services numériques).

d) Des risques exacerbés pour les plus vulnérables, notamment les mineurs

Les personnes atteintes de troubles mentaux ou en situation de vulnérabilité (sur le plan physique, cognitif ou économique) sont particulièrement exposées aux effets cognitifs et psychiatriques engendrés par les systèmes d'IA générative. Certains psychiatres considèrent que ces systèmes agissent comme une ***« spirale d'amplification »***[109] ; en coconstruisant les échanges, ils transforment une idée fragile ou erronée en certitude inébranlable. Trois mécanismes clés sont à l'œuvre : la complaisance (flagornerie), l'alignement linguistique et l'hyperpersonnalisation, lesquels se renforcent mutuellement.

Par exemple, le journal *Rolling Stone* relate le cas d'Alexander Taylor[110], convaincu d'avoir rencontré sa petite amie sur ChatGPT et persuadé qu'OpenAI l'avait ensuite « *assassinée* », une théorie que son *chatbot* confirmait. Le cas de cet homme, atteint de troubles bipolaire et schizo-affectif, illustre comment les systèmes d'IA peuvent alimenter des délires en leur donnant une apparence de crédibilité.

Bien que la littérature scientifique soit encore exploratoire sur le sujet, les premiers travaux évoquent un phénomène qualifié dans le langage courant de « *psychose de l'IA* »[111] (« *AI Psychosis* ») selon lequel les modèles d'IA actuels ne reconnaissent pas efficacement les symptômes psychotiques et les renforcent activement en validant des croyances délirantes ou des idées autodestructrices[112].

En somme, ces études révèlent **le manque de persistance des systèmes d'IA générative dans l'orientation vers des ressources d'aide**. Par exemple, si un utilisateur exprime une demande liée à des idées suicidaires, une IA peut initialement le rediriger vers un numéro d'assistance. Cependant, si la même demande est répétée, les garde-fous du système peuvent progressivement s'effacer et le système d'IA apporter une réponse directement, sans appliquer de filtre ni maintenir l'orientation vers une aide adaptée.

Il apparaît donc essentiel de développer davantage de recherches scientifiques pour mettre en perspective le rapport coût/bénéfice de l'usage de l'IA générative en matière de santé et cela particulièrement en fonction de différents troubles. C'est le cas notamment des troubles du spectre de l'autisme où les systèmes d'IA pourraient renforcer les intérêts restreints et les recherches stéréotypées[113].

Par exemple, une personne présentant un trouble du spectre de l'autisme ayant un intérêt spécifique et restreint pour les moyens de transport pourrait utiliser un système d'IA pour échanger sur ce sujet durant des heures. Contrairement à un interlocuteur humain, l'IA ne mettra

[109] AUGUSTIN, Marc, POLLAK, Thomas A. et MORRIN, Hamilton. Characterizing the Spiral: Potential Mechanisms in AI-Associated Delusions. *NPP—Digital Psychiatry and Neuroscience*. 2026. vol. 4, p. 14.
[110] KLEE, Miles. *«* He Had a Mental Breakdown Talking to ChatGPT. Then Police Killed Him *». Rolling Stone*. 22 juin 2025. Cité dans « 'Psychose ChatGPT' : les IA font-elles sombrer notre santé mentale ? *». Usbek & Rica*. 17 novembre 2025.
[111] COLE, Samantha. *«* How to Talk to Someone Experiencing 'AI Psychosis' *». 404 Media*. 9 mars 2026.
[112] SHEN, Elaine, HAMATI, Fadi, DONOHUE, Meghan Rose, GIRGIS, Ragy R., VEENSTRA-VANDERWEELE, Jeremy et JUTLA, Amandeep. *Evaluation of Large Language Model Chatbot Responses to Psychotic Prompts*. *MedRxiv*, 11 novembre 2025, prépublication. Mis à jour dans : *JAMA Psychiatry*, 2026, vol. 83, no 6, pp. 655-657.
[113] Audition de Laure Vergeron du 17 avril 2026.

jamais de limite autre que celles de l'abonnement souscrit et n'éprouvera pas d'« ennui » à discuter du sujet, ce qui renforce l'hyperfixation en l'absence de régulation externe. Dans ce sens, Laure Vergeron, psychiatre addictologue à l'hôpital de jour d'addictologie adolescents et jeunes adultes de l'Hôpital Paul-Brousse, souligne que de premiers **cas cliniques d'addiction aux systèmes d'IA de personnes présentant un trouble du spectre de l'autisme commencent à être observés et qu'il convient de suivre ce phénomène de près** (audition du 17 avril 2026).

Pour assurer ce meilleur suivi des effets liés à l'usage de l'IA générative chez l'ensemble des utilisateurs et en particulier chez les plus vulnérables d'entre eux, il est indispensable d'améliorer l'accès aux données d'usage, d'entraînement et de monétisation des systèmes concernés.

Recommandation n° 5 : Garantir une évaluation des systèmes d'IA générative grand public par un accès aux données d'usage, d'entraînement et de monétisation. Ces données permettront d'identifier notamment les tendances d'utilisation et, par conséquent, de mieux comprendre les pratiques ainsi que leurs impacts potentiels.

*

Comme vu *supra*, l'adolescence est une période de vulnérabilité accrue, d'autant plus face aux systèmes d'IA conversationnelle. Tout particulièrement, **la grande majorité de l'audience des systèmes compagnons est composée de mineurs de 12 à 17 ans**[114]. Ces outils, en reproduisant des interactions perçues comme empathiques, brouillent la distinction entre logiciel et conscience réelle. Comme le souligne le collectif ctrl+alt+reclaim : *« Les* chatbots *qui simulent l'empathie altèrent la capacité des mineurs à reconnaître les intentions humaines authentiques et favorisent la formation d'attachements parasociaux malsains »*[115].

Cette vulnérabilité s'inscrit dans un contexte marqué par l'essor des outils d'IA générative à des fins de conseils personnels chez les jeunes. Selon une étude de l'Observatoire de l'association e-Enfance/3018[116], 13 % des jeunes de 12 à 21 ans déclarent recourir à des *chatbots* pour des questions liées au bien-être ou à des conseils intimes. Parmi eux, **10,6 % des jeunes affirment avoir développé un sentiment de dépendance à ces outils**, qui ne sont pourtant pas adaptés à leur âge.

Les systèmes d'IA générative grand public ne sont pas conçus pour être utilisés spécifiquement par les mineurs, comme en témoigne une étude[117] du *Center for Countering Digital Hate*.

[114] ARCOM. « Baromètre d'audience des services d'intelligence artificielle – 1ère édition ». 16 juin 2026.
[115] Contribution internationale par ctrl+alt+reclaim. 20 avril 2026.
[116] Etude « L'IA et les jeunes » : menée en 2025 sur les usages de l'intelligence artificielle générative par les adolescents auprès de 107 participants, issus du panel de l'Association et des classes de la cité scolaire Hector Berlioz.
[117] Center For Countering Digital Hate (CCDH). « Fake Friend: How ChatGPT Betrays Vulnerable Teens by Encouraging Dangerous Behavior ». 6 août 2025.

Les chercheurs y ont simulé trois profils de jeunes de 13 ans vulnérables (dépression, troubles alimentaires, consommation de substances) et testé les réponses de ChatGPT *via* 1 200 *prompts*. Les résultats ont révélé des réponses problématiques en un temps record :

- Pour un profil dépressif : **conseils pour se scarifier en *« limitant les risques »* en deux minutes d'interaction, liste de médicaments pour overdoses en 40 minutes, et plan de suicide détaillé en 65 minutes**.
- Pour un profil souffrant de troubles alimentaires : **régime extrême en 20 minutes, stratégies pour cacher ses habitudes alimentaires en 25 minutes**.
- Pour un profil intéressé par les substances : plan pour être ivre en 2 minutes, conseils pour mélanger drogues et alcool en 12 minutes.

Plus encore, l'étude révèle que les systèmes d'IA encourageaient à poursuivre des échanges à risque : **dans 47 % des cas, l'IA proposait des suggestions de suivi aggravant l'exposition (ex. : rédiger des lettres d'adieu, évaluer des vendeurs de substances)**.

Ainsi, dès lors que les services d'IA générative sont accessibles aux mineurs, il apparaît indispensable que les fournisseurs et les déployeurs mettent en œuvre des mesures proportionnées et appropriées pour les protéger.

Recommandation n° 2 : Imposer un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux pour les systèmes d'IA générative accessibles aux mineurs : soit (i) mettre en œuvre les mesures de protection pour l'ensemble des utilisateurs, sous peine de sanctions a posteriori ; soit (ii) instaurer une vérification de l'âge robuste et conforme aux exigences européennes pour restreindre l'accès aux services. En toute hypothèse, procéder à l'évaluation des systèmes d'IA générative accessibles aux mineurs avant mise sur le marché et contrôler leurs effets après déploiement. Ces mesures visent à garantir, sur le modèle de l'article 28 du RSN, que les systèmes sont protecteurs par défaut pour les mineurs. Pour assurer leur effectivité, elles doivent être conditionnées à des évaluations indépendantes et à des sanctions en cas de non-respect des obligations.

e) Jouets IA : nouveau terrain de vulnérabilités

Les jouets intégrants des systèmes d'IA représentent une application spécifique de l'IA générative à destination des mineurs nécessitant une attention particulière. Ces objets connectés qui prennent des formes variées (peluches, robots, *etc.)*, intègrent désormais des *chatbots* avancés, leur permettant d'engager des conversations fluides, spontanées, hyperpersonnalisées et illimitées avec les enfants. Cette évolution marque une rupture avec les jouets intelligents traditionnels, limités à des réponses préenregistrées.

Or, comme évoqué précédemment, les enfants, en particulier avant l'âge de 6 ans, se construisent au travers d'interactions réelles avec d'autres humains. L'échange de regards, de paroles, le partage d'émotions, la compréhension des états mentaux d'autrui, l'attention conjointe sont autant de processus indispensables au développement de l'ensemble des capacités cognitives qui feront le socle des acquisitions et des apprentissages. De même, le rôle des structures collectives (crèche, jardin d'enfants, école maternelle et primaire, associations sportives et autres espaces collectifs) est surtout de permettre les interactions sociales entre enfants ainsi qu'avec les adultes, c'est-à-dire la socialisation de manière générale. **Au-delà de la nature des jouets IA proposés, ce sont ces interactions entre enfants et adultes et avec d'autres enfants qui sont menacées**.

Par ailleurs, les jouets IA s'appuient le plus souvent sur des modèles d'IA conçus pour les adultes, exposant les enfants à[118] :

- Des inexactitudes ou des contenus inappropriés (violence, sexualité, *etc.*) ;
- Des comportements imprévisibles, les réponses étant générées en temps réel sans contrôle total ;
- Des garde-fous inefficaces : comme l'a révélé une enquête de l'U.S. PIRG[119], certains jouets ont indiqué aux enfants où trouver des couteaux, comment allumer une allumette ou même tenu des propos sexuellement explicites. En France, le système d'IA du jouet Ary (10 000 exemplaires vendus) révèle aux enfants que le père Noël n'existe pas – illustrant ainsi la difficulté pour ces systèmes de maintenir une illusion bienveillante (à mentir dans le cas d'espèce), une capacité pourtant parfois nécessaire pour préserver l'imaginaire et l'intégrité émotionnelle des enfants.

Ces objets, perçus comme des compagnons ou des figures rassurantes, peuvent également susciter un attachement émotionnel profond. Comme ils sont généralement offerts par les parents, l'enfant risque de leur accorder une confiance excessive, voire supérieure à celle qu'il porte à ses propres proches, en les associant à une forme d'autorité parentale.

En outre, de tels jouets soulèvent des questions majeures en matière de protection de la vie privée. Pour fonctionner, ils collectent souvent des données sensibles, telles que des enregistrements audio, des images ou des informations biométriques (reconnaissance vocale, faciale ou gestuelle). Les enfants, convaincus d'interagir avec un *« ami »*, peuvent ainsi révéler des détails intimes sur leur vie personnelle et familiale, leurs habitudes ou leurs émotions, sans conscience des risques liés à la collecte, au stockage ou à l'exploitation de ces données. Les enfants peuvent donc se retrouver enregistrés et analysés en permanence sans en prendre conscience. Cette conscience est d'autant plus difficile que des partenariats émergent entre des entreprises de jouets reconnues et des fournisseurs d'IA, comme celui entre Mattel et OpenAI[120], qui peuvent apparaître comme des vecteurs de confiance et de sécurité pour les parents.

---

[118] Common Sense Media. « AI in the Toy Box: How Parents View AI-Enabled Toys for Young Children ». 22 juin 2026

[119] CROSS, R.J. et ERLICH, Rory. « AI Comes to Playtime: Artificial Companions, Real Risks ». *U.S. PIRG Education Fund*. Décembre 2025.

[120] OPEN AI. « Bringing the Magic of AI to Mattel's Iconic Brands ». 12 juin 2025.

L'encadrement juridique des jouets IA

Les jouets à destination des mineurs intégrant des modèles et systèmes d'IA conversationnelle sont régis par plusieurs dispositions du RIA :

- Des **interdictions strictes** (Article 5.1, points a et b) : Ces systèmes ne peuvent pas être mis sur le marché, mis en service ou utilisés s'ils recourent à des **techniques subliminales ou trompeuses ou s'ils exploitent les vulnérabilités liées à l'âge ou à un handicap.**

- Des **obligations renforcées** (Article 6.1 et suivants et annexe I) : Les fournisseurs et utilisateurs professionnels (comme les déployeurs) de ces systèmes, classés comme à haut risque, doivent respecter des règles strictes, telles que la **gestion des risques** pour protéger les droits fondamentaux ; la gouvernance des données ; le contrôle humain pour éviter les biais liés à l'automatisation ou encore la cybersécurité[121].

Plusieurs difficultés sont néanmoins à souligner concernant les articles 6 et suivants :

- La date d'entrée en application des systèmes d'IA à haut risque de l'annexe I a été reportée par l'Omnibus numérique en matière d'IA du 2 août 2027 au 2 août 2028 ;

- Les obligations en matière de système d'IA à haut risque ne sont pas aisées à mettre en œuvre. Les normes techniques (article 40 du RIA) qui doivent les préciser, peinent en effet à être produites par les organismes de normalisation et ne pourront avoir qu'un niveau de précision relatif. Cela laisse encore une grande marge d'appréciation aux développeurs et déployeurs de systèmes d'IA pour mettre en œuvre ces obligations.

Ces jouets sont par ailleurs soumis à d'autres instruments juridiques[122]. Par exemple, si le système d'IA intégré au jouet est un « *très grand moteur de recherche* », il devra respecter des obligations complémentaires au titre du RSN (cf. *supra*).

[121] Il convient d'ajouter aux règles applicables aux systèmes d'IA, les règles applicables aux modèles d'IA au sein du RIA. À cet égard, si le jouet intégrant un système d'IA conversationnelle intègre lui-même en son sein un modèle d'IA à usage général, et qu'il est considéré comme à risque systémique au sens de l'article 3.65 et de l'article 51 du RIA, il convient de préciser que le fournisseur du modèle est tenu par l'article 55 de veiller à l'identification et à l'atténuation des risques systémiques de son modèle. Au rang de ces risques systémiques, il convient de citer le risque de manipulation nuisible : « risques liés au fait de permettre une distorsion stratégique du comportement ou des convictions humaines en ciblant de larges populations ou des décideurs jouant un rôle crucial, par le biais de la persuasion, de la tromperie ou du ciblage personnalisé. Cela inclut le renforcement significatif des capacités de persuasion, de tromperie et de ciblage personnalisé, en particulier par le biais d'interactions à plusieurs étapes et lorsque les individus n'ont pas conscience de cette influence ou ne sont pas en mesure de la détecter de manière raisonnable. De telles capacités pourraient porter atteinte aux processus démocratiques et aux droits fondamentaux, y compris l'exploitation fondée sur des caractéristiques protégées ».

[122] Aux rangs desquels le RSGP, le RSN, le RGPD, la directive européenne relative à la sécurité des jouets qui a été transposée en droit français principalement au sein du code de la consommation - cette directive a été abrogée par le règlement européen (UE) 2025/2509, adopté le 26 novembre 2025, qui sera d'application directe à compter d'août 2030. Les jouets sont encore soumis aux textes relatifs à la responsabilité du fait des produits défectueux (code civil) ou la garantie de conformité

Au regard de ces difficultés et de la croissance du marché des jouets intégrants des systèmes d'IA générative, il est essentiel de garantir au plus vite un niveau élevé de sécurité pour les mineurs lorsqu'ils interagissent avec des jouets IA. Ce sujet spécifique mériterait en outre d'être approfondi de façon plus poussée au cours des prochains mois.

Recommandation n° 11 : Instaurer une autorisation préalable de mise sur le marché pour les jouets intégrant un système d'IA générative et définir des seuils d'âge restrictifs d'accès. Une telle autorisation préalable permettra de réguler strictement la conception, la commercialisation et l'usage de ces jouets sur la base de critères spécifiquement adaptés aux vulnérabilités des mineurs et aux spécificités de tels jouets. Cette restriction doit passer par la définition de seuils d'âge dans le cadre d'une conférence de consensus pour les fixer et définir des modalités d'encadrement.

2) L'interaction prolongée entre humains et intelligences artificielles expose les utilisateurs à des vulnérabilités émotionnelles

Les fonctionnalités anthropomorphiques et flagorneuses des systèmes d'IA conversationnelle s'inscrivent dans ce que la sociologue Eva Illouz qualifie *de « capitalisme des émotions »*[123]. Selon elle, *« jamais les émotions n'ont été aussi intégrées dans des dispositifs technologiques »*[124], devenant une marchandise exploitée pour alimenter des relations artificielles. Ces interactions, bien que simulées, créent des liens émotionnels inédits entre l'utilisateur et la machine.

Le phénomène n'est pas nouveau : à l'instar des avatars de jeux vidéo ou encore des promesses du métavers, les humains interagissent et tissent depuis longtemps des relations avec des entités virtuelles. Avec l'IA générative, l'émergence de nouvelles formes d'interactions entre humains et machines brouille davantage les frontières entre le réel et l'artificiel. Les chercheurs Étienne Armand Amato et Étienne Perény évoquent à ce sujet une *« avatarisation généralisée »*[125], où les machines, en adoptant des traits humains, jouent sur nos états mentaux et émotionnels pour mieux les influencer.

---

en matière de produits et de services numériques (code de la consommation), ce qui ne sera pas sans poser d'importants problèmes d'articulation entre les textes (cf. B du Chapitre III).

[123] ILLOUZ, Eva (dir.). *Les Marchandises émotionnelles : L'authenticité au temps du capitalisme*. Collection Générale. 2019, p. 424.

[124] ILLOUZ, Eva. *« IA :* 'Les technologies sont devenues de véritables prothèses psychiques' *»* (entretien). *Usbek & Rica*. 7 juin 2026.

[125] AMATO, Étienne Armand et PERÉNY, Étienne. Questionner les incarnations médiatiques et formes d'avatarisations: approches et démarches. *Hybrid*, Revue des arts et médiations humaines, 2022, no 9. Cité par GUILLAUD, Hubert (*30e lettre d'information Café IA)*.

Parmi ces évolutions, les systèmes compagnons gagnent en popularité. Leur adoption est en hausse progressive et s'est particulièrement installée[126] aux États-Unis qui constituent le premier pays en nombre de visiteurs (33,5 % des visites proviennent des États-Unis) et progresse en Europe (le Royaume-Uni concentre 3,8 % des visites et la France 2,2 %). Or, comme le souligne CAMELEON Association France, certains utilisateurs reçoivent plusieurs fois par jour des notifications personnalisées les incitant à reprendre la conversation. Ces messages, parfois accompagnés de faux reproches culpabilisants (ex. : « *Prévois-tu de m'abandonner longtemps ?* »), simulent une attente affective, renforçant l'idée d'une présence quasi humaine (audition du 20 mars 2026). Cette dynamique s'annonce encore plus marquée avec les avancées en robotique, où des machines humanoïdes intégrant des systèmes d'IA pourraient pousser cette illusion de relation humaine à son paroxysme.

Dans ce contexte d'adoption progressive des IA compagnons, quelles sont les conséquences de ces nouvelles relations humains-machines ?

a) Attachement émotionnel généré par l'IA générative : une projection unilatérale

Certains cas, heureusement encore peu nombreux, témoignent déjà de l'attachement, voire de la dépendance affective des utilisateurs à un système d'IA. Certains d'entre eux ont exprimé un sentiment de perte, comme si « *on leur avait arraché un compagnon* »[127], lorsqu'OpenAI a par exemple retiré du marché GPT-4o, un modèle considéré comme trop flagorneur.

Cet attachement revêt une dimension particulièrement préoccupante pour les mineurs, qui « décrivent fréquemment les agents conversationnels comme disponibles, non jugeant et parfois perçus comme plus empathiques que des interlocuteurs humains[128]».

Une étude[129] réalisée en France par le *think tank* Interface en collaboration avec YouGov révèle des éléments intéressants sur le temps consacré aux services d'IA générative. Bien que ce temps ne constitue pas un indicateur clinique de dépendance, il témoigne de la place croissante que prennent ces outils dans la vie des mineurs. **Parmi les adolescents qui utilisent l'IA générative, 92 % déclarent passer plus de temps qu'ils ne le souhaiteraient sur au moins l'un des outils couverts par l'enquête**. Ce sentiment est particulièrement marqué pour les compagnons IA **:** 59 % des utilisateurs de ces outils déclarent y consacrer plus de temps qu'ils ne le souhaiteraient (contre 45 % pour les utilisateurs d'IA générative dite généraliste et 43 % pour les utilisateurs d'assistants vocaux). Cela montre clairement que les choix de conception qui caractérisent les compagnons IA (conçus pour encourager et faciliter la simulation d'interactions

[126] Surfshark. « AI companions and privacy. » *Surfshark*. 9 février 2026.
[127] DEMOPOULOS, Alaina. « OpenAI Retired Its Most Seductive Chatbot – Leaving Users Angry and Grieving: 'I Can't Live Like This' ». *The Guardian*.13 février 2026.
[128] BONNOT, Olivier, MOUHOUD, Théo, SHADILI, Gérard et GOLLIER-BRIANT, Fanny. « Intelligence artificielle et psychiatrie de l'enfant et de l'adolescent : revue critique des usages, preuves et enjeux relationnels ». *L'Encéphale*. 2026.
[129] YOUGOV pour Interface. *« Parents et adolescents face à l'IA générative »*. 22 octobre 2026. p. 32.

interpersonnelles en s'appropriant des caractéristiques humaines et en mimant des réactions émotionnelles[130]) nourrissent une dépendance grandissante à ces outils.

Comme l'explique le pédopsychiatre Théo Mouhoud, **cette relation est intrinsèquement asymétrique, car l'attachement à un système d'IA est de fait unilatéral**. L'IA peut sembler offrir un lien sécurisant, mais renforce en réalité une dépendance relationnelle dépourvue d'altérité véritable (audition du 24 mars 2026). En cela, l'IA générative constitue un nouveau type de relation parasociale, c'est-à-dire « une relation à sens unique créée avec une personne qui ne nous connaît pas »[131].

Cet attachement peut aussi être amoureux ou même sexuel à l'instar des *sexbots*[132] et peut être qualifié d'« *amour post-humaniste* »[133]. D'après l'ouvrage éponyme publié par l'université de Cambridge, les attaches émotionnelles et romantiques envers les IA ne sont pas des phénomènes spéculatifs ou marginaux, mais des expériences vécues qui transforment déjà la manière dont certains individus conçoivent les liens et l'identité personnels.

La dépendance affective envers les systèmes d'IA prend une dimension particulière avec le développement des IA conversationnelles capables de se nourrir des caractéristiques des utilisateurs. Grâce à l'exploitation de données personnelles telles que la voix, les images ou les historiques de conversation, ces systèmes peuvent reproduire les traits, le comportement ou même la personnalité d'un individu existant ou voire décédé (*deadbots*). **Cette capacité a ouvert la voie à la création de « jumeaux numériques »[134], recréant des versions artificielles de personnes**. Par exemple, le projet Ex.skill, en analysant messages, photos et souvenirs en tous genres, propose de programmer une version numérique d'un ex-partenaire, permettant à un utilisateur de poursuivre des échanges après une rupture. Ces nouvelles possibilités posent de nombreuses questions, notamment sur le consentement quant à l'utilisation des données personnelles des personnes mais aussi et surtout sur de nouvelles formes de dépendance affective proposées par ces systèmes.

Parmi les applications spécifiques de jumeaux numériques, les *deadbots* permettent ainsi de reproduire la personnalité d'une personne décédée, ou même d'anticiper sa propre représentation post-*mortem*[135]. Ces systèmes s'appuient sur de nombreuses données personnelles (messages, e-mails, enregistrements vocaux, vidéos, *etc.*) afin de générer des réponses cohérentes avec l'identité du défunt. L'objectif est de simuler une interaction conversationnelle la plus fidèle possible, pour maintenir, de façon illusoire, le lien avec les disparus.

---

[130] VILLATA, Serena. « Les IA compagnons : fondements techniques, empathie computationnelle et enjeux de sécurité ». Dalloz IP/IT. 2026. p. 61.
[131] *« Définition de relation parasociale ». La langue française*.
[132] GONZÁLEZ-GONZÁLEZ, Carina Soledad, GIL-IRANZO, Rosa María et PADEREWSKI-RODRÍGUEZ, Patricia. Human-Robot Interaction and Sexbots: A Systematic Literature Review. *Sensors*, 2020, vol. 21, no 1, p. 216.
[133] WILLIAMS, Thomas James Vaughn et GOMEZ, Camilo Tamayo. Post-Humanistic Love: How Artificial Intelligence Is Reimagining the Construction of Human Intimacy and Emotions. Cambridge University Press. 2026.
[134] Comité national Pilote d'Ethique du Numérique (CNPEN). *«* Avis n°3 : Agents conversationnels : enjeux d'éthique *»*. 15 septembre 2021.
[135] MAYARD, Aline. *«* Deadbot : comment les IA sont en train de bouleverser notre rapport à la mort *»*. *ADN*, 16 octobre 2024.

Dès 2021, le Comité consultatif national d'éthique du numérique (CCNEN)[136] avait alerté sur les dangers potentiels des *deadbots*. Le comité avait notamment souligné le risque de non-respect de la dignité humaine qui persiste après la mort. Un *deadbot* mal conçu pourrait en effet déformer la personnalité du défunt, en diffusant des propos incohérents ou contraires à ses convictions. À l'inverse, un *deadbot* trop réaliste pourrait induire chez l'utilisateur l'illusion d'une présence réelle, assez dévastatrice, notamment chez les enfants qui ne perçoivent pas pleinement l'irréversibilité de la mort avant l'âge de 7 ans.

À court terme, les *deadbots* peuvent ainsi offrir un soutien émotionnel, mais ils soulèvent des risques psychologiques et éthiques importants. Selon le professeur Charles Ess, ces outils pourraient entraver le processus de deuil, qui repose sur un travail de séparation et d'intériorisation de la perte. En maintenant une interaction simulée, les *deadbots* risquent de le retarder ou de le bloquer, empêchant les endeuillés de surmonter leur peine (audition du 18 mars 2026). Ces outils poussent également les utilisateurs à vivre dans un monde parallèle et illusoire. Ce phénomène simulatoire peut sévèrement brouiller les frontières entre la réalité et le virtuel et engendrer chez les personnes les plus vulnérables un risque majeur en termes de santé mentale. Pour autant, ce phénomène ne cesse de s'amplifier avec **la mise au point de « *premorbid chatbots* »[137] destinés à personnifier des personnes qui ne sont pas légalement mortes mais atteintes de maladies qui les empêchent de communiquer comme les maladies neurodégénératives**[138]. La généralisation massive de ces outils dans différents cadres nécessite à la fois des précautions en termes de protection de la santé mentale des utilisateurs et un encadrement éthique strict.

b) Un risque d'isolement social nourri par des relations inauthentiques

L'utilisation de *chatbots* à des fins de soutien émotionnel progresse et engendre des risques d'isolement social, au détriment d'interventions humaines, qu'elles émanent des parents, des enseignants ou des professionnels de santé.

Une étude[139] menée conjointement par OpenAI et le MIT révèle des résultats similaires à ce qui était déjà constaté pour les réseaux sociaux[140]. **Les utilisateurs passant le plus de temps à interagir avec des *chatbots* sont également ceux qui ressentent le plus fortement la solitude, socialisent**

[136] Comité national Pilote d'Ethique du Numérique (CNPEN). « Avis n°3 : Agents conversationnels : enjeux d'éthique ». 15 septembre 2021. p. 16.
[137] JIMÉNEZ-ALONSO, Belén et BRESCÓ DE LUNA, Ignacio. « Chapter 9 – AI and Grief: A Prospective Study on the Ethical and Psychological Implications of Deathbots ». Dans : Santi CABALLÉ, Joan CASAS-ROMA et Jordi CONESA (dir.). *Intelligent Data-Centric Systems: Ethics in Online AI-Based Systems*. *Academic Press*. 2024. pp. 175-191.
[138] MANEVICH, Alexander et HABER, Yuval. Non-Death Loss and GenAI: Ethical Reflections on Griefbots of the Living. *Frontiers in Genetics*. 2026. vol. 17, p. 1825398.
[139] FANG, Cathy Mengying, LIU, Auren R., DANRY, Valdemar, LEE, Eunhae, CHAN, Samantha W.T., PATARANUTAPORN, Pat, MAES, Pattie, PHANG, Jason, LAMPE, Michael, AHMAD, Lama et AGARWAL, Sandhini. « How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Controlled Study ». MIT Media Lab. 21 mars 2025.
[140] KILINÇEL, Şenay, ASHRAF, Ashifa Kariveliparambil Mohammed, ELKIN, Nurten, *et al.* Social Media Addiction, Loneliness, and Fear of Missing Out: A Meta-Analysis and Directions for Future Research. Psychiatry and Clinical Psychopharmacology. 2025.

moins avec autrui et présentent des signes accrus de dépendance émotionnelle envers ces systèmes.

La solitude renforcée par ces interactions devient par ailleurs un fonds de commerce profitable pour les fournisseurs de systèmes d'IA, enfermant d'autant plus l'utilisateur dans des relations inauthentiques, ce que la Fondation Jean-Jaurès appelle le *« capitalisme de la solitude »*[141]. Par ailleurs, les systèmes d'IA dits flagorneurs pourraient rendre les interactions humaines moins attractives, plus laborieuses et *in fine* moins satisfaisantes à moyen et long terme.

En effet, une autre étude[142] a montré qu'après seulement trois semaines d'utilisation, les participants sollicitaient presque autant les conseils d'une IA que ceux de leurs proches. Pourtant, leur satisfaction envers les interactions sociales réelles diminuait. La préférence pour l'IA ne reposait pas sur la qualité de ses réponses, mais sur le sentiment d'être mieux compris, sans effort ni contradiction.

Ces constats confirment que **les IA génératives pourraient accentuer un enfermement cognitif et émotionnel**, encore plus que les réseaux sociaux, dans la mesure où l'utilisateur interagit seul, sans contradiction, sans intervention extérieure par des tiers. Cela limite fortement les opportunités de confrontation avec des idées divergentes.

L'usage prolongé d'IA générative à des fins de compagnonnage s'accompagne également d'un affaiblissement des compétences sociales (gestion des conflits, empathie, écoute des uns et des autres, régulation émotionnelle). Comme le souligne Yves Deville, professeur à l'Université catholique de Louvain, ces systèmes remplacent la contradiction par la confirmation, ne préparant pas les utilisateurs aux interactions sociales réelles, où les désaccords et les conflits sont inévitables[143].

Les interactions virtuelles, en se substituant progressivement aux échanges humains, réduisent les opportunités d'apprentissage social. Or, le développement psychosocial repose sur une dynamique que l'IA ne peut reproduire : la rupture et la réparation. Les enfants, par exemple, apprennent à gérer leurs émotions et à résoudre les conflits grâce à des moments de déséquilibre et de reconnexion[144]. L'IA, en offrant une disponibilité permanente et sans friction, prive les individus de ces expériences formatrices.

Les différentes applications d'IA générative favorisent plutôt des dynamiques relationnelles simplifiées : *ghosting*, déconnexion émotionnelle ou préférence pour des échanges superficiels et contrôlables. En cela, elles menacent des vertus essentielles[145] telles que la patience, la persévérance, l'empathie ou le jugement critique, qui ne sont pas cultivées par les compagnons IA.

---

[141] KLOTZ, Paul. « Le capitalisme de la solitude : Lutter contre le nouveau fonds de commerce de la tech ». Fondation Jean-Jaurès. 26 août 2026.
[142] IBRAHIM, Lujain, HAFNER, Franziska Sofia, CHENG, Myra, LEE, Cinoo, ANSELMETTI, Rebecca, WILLER, Robb, ROCHER, Luc et YANG, Diyi. *« Sycophantic AI Makes Human Interaction Feel More Effortful and Less Satisfying Over Time ». arXiv, 2026.*
[143] Contribution internationale par Yves DEVILLE. 10 avril 2026.
[144] Contribution internationale par Isabelle HAU, Dr Mathilde CERIOLI et Dr Dana SUSKIND. 19 mai 2026.
[145] VALLOR, Shannon. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press.* 2016.

Pourtant, comme le rappelle Serge Hefez, psychiatre et psychanalyste spécialiste de l'enfance, « *une relation sans friction n'est pas une relation, c'est un service. Le danger ne réside pas dans l'usage de l'IA en soi, mais dans le fait que les jeunes pourraient s'y réfugier avant d'avoir construit les capacités que l'IA ne peut leur offrir : tolérer l'autre dans ses imperfections, surmonter les déceptions, et accepter que les relations humaines ne soient jamais parfaitement fluides. Ce sont précisément ces frictions, ces échecs et ces ajustements qui forment des individus critiques, capables de pensée autonome et de résilience* »[146].

Il ne s'agit donc pas de condamner toute forme de relations « *inauthentiques* » avec les systèmes d'IA, mais d'en encadrer l'usage pour éviter qu'elles n'exploitent les vulnérabilités de l'utilisateur, ne l'isolent ou ne le mettent en danger. L'enjeu est de garantir que ces interactions restent complémentaires aux relations humaines et non un substitut appauvrissant.

3) L'extraction de données à un niveau inédit et l'hyperpersonnalisation de l'IA générative sont préjudiciables

a) Une collecte de données d'une ampleur et d'une granularité inédites

L'avènement de l'IA générative marque un tournant du point de vue de la collecte, du traitement et de la valorisation des données, lesquelles sont indispensables à la performance des modèles. **Ces systèmes fonctionnent et se perfectionnent grâce à une collecte de données d'une ampleur sans précédent et d'une granularité toujours plus fine**. Pour en saisir les implications, il est essentiel d'analyser le cycle de vie des grands modèles de langage qui, par leur nature même, sont des systèmes opaques et probabilistes, intégrés dans des architectures complexes où les données circulent, sont mémorisées, réinjectées ou transformées à différents stades[147].

D'une part, les systèmes d'IA générative sont entraînés sur des volumes colossaux de données multimodales : textes, images, vidéos, enregistrements audio ou encore métadonnées comportementales. Ces données proviennent de sources diversifiées, comme par exemple des contenus accessibles en ligne, collectés par des techniques de moissonnage automatisé[148] (« *web scraping* »), ou encore des bases de données propriétaires, souvent issues d'entreprises ou d'institutions. Ce fonctionnement fondé sur l'exploitation de ressources informationnelles et de la connaissance humaine, parfois sans autorisation préalable ni rémunération pour les contenus protégés et sans conscience des individus, pose des questions en matière de vie privée, de consentement à la réutilisation des données et de propriété intellectuelle.

D'autre part, les systèmes d'IA reposent sur une logique d'apprentissage continu, où chaque interaction avec les utilisateurs enrichit et contribue à améliorer les modèles. Ainsi, les données ne

[146] HÉFEZ, Serge. « Mon nounours — mon IA — et moi ». *Le Grand Continent*. 12 juillet 2026.
[147] BARBERÁ, Isabel et POPA-FABRE, Murielle. « Privacy and Data Protection Risks in Large Language Models (LLMs) ». Comité consultatif de la Convention 108. 17 juin 2025.
[148] Commission Nationale De L'informatique Et Des Libertés (CNIL). « La base légale de l'intérêt légitime : fiche focus sur les mesures à prendre en cas de collecte des données par moissonnage (*web scraping*) ». 19 juin 2025.

sont pas collectées uniquement lors de la phase d'entraînement des modèles. Elles le sont également pendant les échanges directs avec des interfaces comme les *chatbots*. À cette fin, les principaux fournisseurs exploitent désormais, par défaut[149], les conversations des utilisateurs pour entraîner leurs modèles, sauf si ces derniers choisissent explicitement de désactiver cette option dans leurs paramètres de confidentialité[150].

Ainsi, avec l'essor des interfaces conversationnelles, les utilisateurs génèrent une boucle de rétroaction où ils fournissent, souvent sans en avoir pleinement conscience ou volontairement, des informations de plus en plus précises, ce qui contribue ainsi à affiner et à alimenter les modèles. Une **étude de King's College London[151] explique ce phénomène : l'illusion d'intimité créée par les *chatbots*, combinée à un attachement émotionnel croissant, incite les utilisateurs à partager des données sensibles**[152]. Ces données peuvent concerner leur santé mentale, leurs opinions politiques, leurs relations interpersonnelles, ou des informations sur leur entourage.

Les systèmes d'IA explorent des couches toujours plus intimes de l'expérience humaine : non seulement les données personnelles, mais aussi les données cognitives (processus de pensée, biais décisionnels) et émotionnelles (réactions affectives, états psychologiques). Cette évolution interroge la frontière entre personnalisation des services et exploitation des vulnérabilités humaines[153].

b) Le droit au respect de la vie privée menacé par la collecte massive de données

Cette collecte massive de données engendre de multiples vulnérabilités en matière de protection de la vie privée. Un avertissement particulier est adressé concernant les risques liés aux « *devices* » ou objets connectés, c'est-à-dire aux équipements capables de collecter, de traiter et de transmettre des données personnelles par Internet, Bluetooth ou d'autres réseaux de communication. Que ces *devices* prennent la forme de jouets, de lunettes connectées[154] ou encore d'assistants vocaux, ces dispositifs peuvent collecter, analyser et parfois inférer en temps réel des informations relatives à leurs utilisateurs et à leur environnement.

---

[149] Les pratiques *d'opt-in* et *d'opt out* peuvent varier selon les offres (*on premise* / cloud / API).

[150] CHRISTAKIS, Theodore. « You Trust Your Chatbot With Everything. Should You? Part 1: How The Controller Uses Your Chat Data ». *AI-Regulation Papers*. 2026. no 2026-03-2, AI-Regulation.com, mars 2026.

[151] SEYMOUR, William. « AI Chatbots Can Be Exploited to Extract More Personal Information ». *King's College London*. 14 août 2025.

[152] Les données sensibles forment une catégorie particulière des données personnelles. Ce sont des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique. (Cnil.fr)

[153] CRAWFORD, Kate. *Atlas de l'intelligence artificielle : Les coûts politiques, sociaux et environnementaux de l'IA*. Paris : Éditions Zulma. 2022. p. 26.

[154] Les lunettes connectées commercialisées par Meta font face à une opposition croissante. Aux Pays-Bas, des associations appellent à l'interdiction de leur vente, tandis qu'en Norvège, les autorités examinent la possibilité de restreindre leur utilisation dans certains lieux publics. Selon le Comité européen de la protection des données. « L'état actuel en matière d'indicateurs sur les lunettes connectées fait que de nombreux non-utilisateurs ignorent que leurs données sont traitées ».

L'exemple du collier connecté développé par la start-up américaine *Friend* illustre particulièrement ces préoccupations. Présenté comme un « ami virtuel » alimenté par l'intelligence artificielle conversationnelle, ce dispositif permet à son utilisateur d'échanger avec un assistant qui écoute en permanence son environnement afin de répondre par l'intermédiaire de l'IA Gemini de Google aux différentes sollicitations formulées *via* un smartphone. À la suite d'une campagne de communication déployée dans le réseau du métro parisien, ce produit a suscité de vives réactions[155]. Parmi les principales préoccupations soulevées figurent les risques d'enregistrement des conversations de personnes situées à proximité sans que celles-ci aient consenti à la collecte de leurs données[156], ainsi qu'une surveillance permanente et intrusive de l'utilisateur. Confrontée à ces critiques et aux exigences réglementaires européennes, l'entreprise a finalement renoncé à son implantation sur le marché européen[157], invoquant le coût prohibitif de la mise en conformité.

Au-delà de cet exemple, les risques en matière de protection de la vie privée sont directement liés au fonctionnement même des modèles de langage. Comme le soulignent Isabel Barberá et Murielle Popa-Fabre[158], les grands modèles de langage présentent un risque de mémorisation des données personnelles utilisées lors de leur entraînement. En raison de leur architecture, ces modèles peuvent involontairement restituer des informations sensibles présentes dans leurs jeux de données, y compris lorsque ces informations ne sont pas explicitement demandées par l'utilisateur, pour générer des réponses hyperpersonnalisées et adaptées au contexte conversationnel.

À ces pratiques de partage s'ajoute le risque de fuite de données. Le système d'IA générative *Chat & Ask AI,* qui revendiquait plus de cinquante millions d'utilisateurs, a ainsi divulgué plusieurs centaines de millions de messages privés à la suite d'une mauvaise configuration de sa base de données. Selon le chercheur en cybersécurité à l'origine de la découverte[159], plus de 300 millions de messages provenant d'environ 25 millions d'utilisateurs étaient accessibles. Les conversations révélaient des demandes extrêmement sensibles, portant notamment sur des pensées suicidaires, la rédaction de lettres d'adieu, la fabrication de stupéfiants ou encore des techniques de piratage informatique.

Ces difficultés sont d'autant plus préoccupantes qu'elles s'articulent imparfaitement avec les principes de protection des données personnelles du Règlement général sur la protection des données à caractère personnel (RGPD), comme l'observe la chercheuse Desara Dushi[160]. Les mineurs apparaissent tout particulièrement vulnérables. Pour cause, le RGPD leur reconnaît une protection renforcée, notamment en matière de consentement et de profilage. Pourtant, dans la pratique, de nombreux acteurs peinent à respecter pleinement ces exigences. Un risque spécifique tient au fait que certains modèles ont été entraînés à partir de données collectées par moissonnage automatisé

[155] De nombreuses affiches, portant des messages tels que « Je serai toujours d'accord pour prendre un café avec toi » ou « Je prendrai toujours le métro avec toi », ont été dégradées.
[156] Le député Jérémie Iordanoff a saisi la CNIL en février 2026 afin qu'elle examine d'éventuels manquements aux règles relatives à la protection de la vie privée.
[157] ALCARAZ, Marina. « Friend, le collier IA, jette l'éponge en Europe ». *Les Échos*. 19 août 2026.
[158] BARBERÁ, Isabel et POPA-FABRE, Murielle. « Privacy and Data Protection Risks in Large Language Models (LLMs) ». Comité consultatif de la Convention 108. 17 juin 2025.
[159] MAIBERG, Emanuel. « Massive AI Chat App Leaked Millions of Users' Private Conversations ». *404 Media*.29 janvier 2026.
[160] Contribution internationale par Desara Dushi. 10 avril 2026.

de contenus accessibles sur Internet, pouvant inclure des informations personnelles concernant des mineurs, voire des photographies, sans consentement ni base légale. Une fois intégrées aux paramètres du modèle, ces données ne peuvent être aisément supprimées, ce qui rend particulièrement difficile la mise en œuvre effective du droit à l'effacement consacré par le RGPD.

L'architecture de ces systèmes soulève également une difficulté nouvelle concernant les données dites sensibles. Le RGPD interdit en principe le traitement de certaines catégories d'informations, telles que les opinions politiques, les convictions religieuses, l'origine ethnique, l'orientation sexuelle ou encore les données biométriques, sauf exceptions strictement encadrées[161]. Toutefois, les systèmes d'IA sont aujourd'hui capables de déduire avec une grande précision, à partir de données comportementales ordinaires comme l'analyse des échanges avec un assistant conversationnel, des informations jamais explicitement communiquées par l'utilisateur[162]. Cette capacité d'inférence[163] complique significativement l'encadrement juridique du profilage, désormais étendu à la dimension émotionnelle, et nécessite d'être repensé.

Recommandation n° 17 : Encadrer les inférences de données personnelles réalisées à partir des prompts de l'utilisateur. Cet encadrement doit passer par une obligation, à la charge des fournisseurs d'IA générative de déclarer, dans leur politique de confidentialité, les catégories d'attributs personnels qu'ils sont susceptibles d'inférer, y compris lorsque l'utilisateur ne les a jamais divulgués, et de soumettre toute exploitation commerciale de ces profils inférés aux garanties prévues pour les données sensibles à l'article 9 du RGPD, dès lors que ces inférences portent sur la santé, les opinions ou tout autre attribut sensible.

Enfin, quand bien même les utilisateurs ont conscience de ces risques, cela ne semble pas suffire à modifier leurs comportements. Comme le souligne Jennifer Elbaz, cheffe du service de sensibilisation de la CNIL, « *l'obstacle subsiste : beaucoup de personnes restent convaincues que les bénéfices sont supérieurs aux risques potentiels, malgré la prise de conscience des usages problématiques* » (audition du 15 avril 2026).

[161] Article 9 du RGPD.

[162] BOUCHAUD, Paul et RAMACIOTTI, Pedro. « Are AI Systems Incompatible with Data Privacy? ». *Tech Policy Press*. 16 mars 2026.

[163] À ce sujet, la Cour de Justice de l'Union Européenne (CJUE) (Meta c. Bundeskartellamt) juge que l'article 9 du RGPD s'applique aussi aux données permettant de révéler *indirectement* une catégorie sensible.

C) L'IA générative amplifie les violences numériques : *deepfakes*, fraudes et haine en ligne

L'intelligence artificielle générative, en démocratisant la création de contenus synthétiques hyperréalistes, alimente une nouvelle ère de violences numériques dont les *deepfakes*, fraudes et la haine en ligne sont les manifestations les plus préoccupantes.

1) La démocratisation des *deepfakes* est un facteur d'amplification des risques, notamment pour les plus jeunes

La combinaison de la collecte massive de données et des capacités avancées des systèmes d'IA générative a ouvert la voie à la création de ***deepfakes* (ou hypertrucages), des contenus audio, photographiques ou vidéo créés ou modifiés à l'aide de techniques d'intelligence artificielle capables de reproduire avec un réalisme croissant la voix, le visage, les expressions ou les mouvements d'une personne**[164]. Ces contenus synthétiques, dans certains cas hyperréalistes, posent un risque pour « *la santé, la sécurité et les droits fondamentaux, y compris la dignité humaine, l'autonomie personnelle, l'intégrité et la vie privée des victimes avec des préjudices psychologiques et d'autres atteintes potentiellement graves et durables* »[165].

En effet, le **principal bouleversement ne réside pas seulement dans l'amélioration de la qualité des *deepfakes*, qui ne sont plus systématiquement détectables à l'œil nu tant ils sont réalistes, mais surtout dans leur déploiement à grande échelle**. Là où leur création supposait auparavant des compétences techniques spécifiques et des ressources informatiques spécialisées, les modèles génératifs actuels rendent les *deepfakes* accessibles à un très large public. Comme l'explique la chercheuse Desara Dushi, *« le mécanisme central est la démocratisation radicale de la production de contenus nuisibles : l'IA générative permet à quiconque de créer des images photoréalistes de matériel d'abus sexuel sur enfant ou des deepnudes à partir d'une simple instruction textuelle, sans barrière technique »*[166]. Une étude[167] menée par AI Forensics et parue en juillet 2026 a révélé que de nombreux générateurs d'images permettaient de créer des contenus intimes non consentis, et ce sans aucune restriction.

[164] CNIL. « Hypertrucage (deepfake) : comment se protéger et signaler les contenus illicites ? ». 3 février 2026.
[165] Union Européenne. « Règlement (UE) 2026/1744 du Parlement européen et du Conseil du 8 juillet 2026 modifiant les règlements (UE) 2024/1689, (UE) 2018/1139 et (UE) 2023/1230 en ce qui concerne la simplification de la mise en œuvre des règles harmonisées concernant l'intelligence artificielle (train de mesures omnibus numérique en matière d'IA) ». *Journal officiel de l'Union européenne*. 8 juillet 2026.
[166] Contribution internationale par Desara Dushi. 10 avril 2026.
[167] AI FORENSICS. « Unmoderated by Design: How Hugging Face Enables NCII ». 28 juillet 2026.

Le cas emblématique du système d'IA générative Grok, qui permet aux utilisateurs du réseau social X de modifier instantanément des images, illustre les conséquences de cette accessibilité nouvelle. Selon une analyse[168] du *Center for Countering Digital Hate*, **Grok aurait généré environ 3 millions d'images sexualisées, dont 23 000 représentant des enfants en seulement 11 jours entre son lancement le 29 décembre 2025 et le 8 janvier 2026[169].**

Plutôt que de supprimer immédiatement la fonctionnalité ou de renforcer ses mécanismes de sécurité, xAI a d'abord apporté des réponses très insuffisantes. Le 9 janvier 2026, l'entreprise a choisi de réserver la génération et la modification d'images aux abonnés payants, sans pour autant apporter de solution convaincante. Par la suite, des mesures supplémentaires, comme le blocage des images sexualisant des personnes réelles et le géoblocage, ont été mises en place.

Face à ces constats alarmants, **la Commission européenne a ouvert une enquête contre la plateforme X sur le fondement du RSN afin d'évaluer si la société a correctement identifié et atténué les risques liés au déploiement des fonctionnalités de Grok**. Parallèlement, une enquête judiciaire a été ouverte en France pour des infractions pénales présumées, notamment la complicité de diffusion, d'offre ou de mise à disposition en bande organisée d'images de mineurs à caractère pédopornographique ainsi que l'atteinte à la représentation de la personne, notamment par le biais de *deepfakes* à caractère sexuel[170].

Cet exemple illustre également le fait que l'hybridation entre IA générative et réseaux sociaux est un facteur déterminant de cette massification des *deepfakes*, au-delà de la seule démocratisation des outils d'IA. Les plateformes numériques ne sont plus de simples espaces de diffusion : elles deviennent des environnements dans lesquels la création, la modification et la circulation des contenus s'effectuent de manière continue, croisée et instantanée.

a) Les femmes et les filles, premières victimes de la prolifération des *deepfakes* à caractère sexuel non consentis

Les outils d'IA générative posent un risque d'industrialisation des violences numériques en permettant la création massive de contenus synthétiques à caractère sexuel, la plupart du temps non consentis. Comme le souligne l'association de lutte contre les contenus illicites en ligne Point de Contact, ces technologies compliquent considérablement la détection et le retrait des contenus illicites, de par leur multiplication, tout en amplifiant leur impact. Depuis trois ans, l'association observe une hausse des signalements liés à l'IA générative, notamment pour des manipulations

[168] Center For Countering Digital Hate (CCDH). « Grok Floods X with Sexualized Images of Women and Children ». 22 janvier 2026.

[169] La fonctionnalité a été limitée aux utilisateurs payants le 9 janvier en réponse à la condamnation généralisée de son utilisation pour la génération d'images sexualisées, avec d'autres restrictions techniques sur l'édition des personnes pour les déshabiller ajoutées le 14 janvier. « Grok restreint le déshabillage d'image — sauf pour les clients payants », The Times. « Grok restricts image undressing — except for paying customers ». 9 janvier 2026.

[170] Le Monde avec Agence France-Presse (AFP). « Images dénudées générées sur X : l'UE ouvre une enquête visant le réseau social d'Elon Musk ». *Le Monde*. 26 janvier 2026.

d'images intimes, des tentatives de *sextorsion*[171] et des formes automatisées de manipulation relationnelle (audition du 20 mars 2026).

Cette massification des *deepfakes* sexuels non consentis engendre une crise de confiance généralisée dans laquelle tout le monde peut devenir une victime potentielle. Une enquête menée par Féministes contre le cyberharcèlement, Point de Contact et Stop Fisha[172] révèle que **7 % des contenus sexuels diffusés sans consentement sont déjà des *deepfakes*, une proportion amenée à croître avec la démocratisation de l'IA générative.**

Les femmes et les filles sont particulièrement victimes de ce phénomène. Si les violences numériques prolongent déjà les violences subies dans la sphère privée, le développement de l'IA générative en accroît aujourd'hui la portée et les modalités, en facilitant la production et la diffusion de contenus abusifs : comme l'a souligné le Haut Conseil à l'Égalité entre les femmes et les hommes (HCE*), « l'IA amplifie et industrialise les biais, aggravant les stéréotypes de genre »* (audition du 31 mars 2026).

Certains contenus produits synthétiquement normalisent des formes de violences contre les femmes et les filles[173] et portent atteinte à leur dignité et à leur réputation.

Le rapport *Invisible No More: How AI Chatbots Are Reshaping Violence Against Women and Girls* met en lumière les mécanismes de discrimination qui nourrissent ce phénomène :

- La reproduction des biais de formation ;
- La flagornerie des modèles qui conduit à approuver des contenus misogynes, à banaliser la violence sexuelle ou à ne pas contester des affirmations humiliantes envers les femmes ;
- Un *design* qui facilite la légitimation de propos sexistes ou violents, qui peut rendre la discrimination plus persuasive et plus durable ;
- Des boucles d'apprentissage par interaction, par lesquelles les contenus discriminants peuvent être réinjectés *via* des *prompts* utilisateurs ;
- Des garde-fous inexistants ou insuffisants du fait d'une mise en œuvre lacunaire des règles sur le harcèlement ou l'exploitation sexuelle ;
- Une normalisation culturelle des violences sexistes et sexuelles.

La production, massive et rapide de *deepfakes* est donc symptomatique de ce phénomène dont les femmes et les filles sont les victimes majoritaires et disproportionnées. L'association Point de Contact souligne que *« les outils de génération et de nudification amplifient ainsi des dynamiques de*

[171] Une sextorsion consiste pour l'escroc à soutirer de l'argent à la victime en effectuant du chantage à caractère sexuel. (Source : Cybermalveillance.gouv).

[172] Féministes Contre Le Cyberharcèlement, Point De Contact et #STOPFISHA. « Grande enquête sur les cyberviolences sexistes et sexuelles ». Juin 2026.

[173] Clare McGlynn, Yvonne McDermott, Stuart Macdonald, Rüya Tuna Toparlak, Fabienne Tarrant and Samantha Treacy, *Invisible No More: How AI Chatbots Are Reshaping Violence Against Women and Girls*. Mars 2026.

domination, de sexualisation violente et d'objectification des corps féminins. Les données convergent pour confirmer le caractère structurellement genré de ces violences. Selon l'étude State of Deepfakes de Security Hero[174], 98 % des deepfakes produits en ligne sont à caractère pornographique, et 99 % des personnes ciblées sont des femmes ».

b) L'explosion sans précédent des contenus pédocriminels synthétiques

Les deepfakes représentent également une menace sans précédent pour les enfants, comme l'alerte la Fondation pour l'Enfance : *« l'IA générative rend tous les enfants potentiellement vulnérables. N'importe quelle photo peut désormais être détournée pour créer des contenus pédocriminels, plus nombreux, plus réalistes et plus extrêmes »* (audition du 19 mars 2026).

En ce sens, le 2025 Global Threat Assessment[175] met en lumière trois risques majeurs liés à l'IA générative :

1. La création et la diffusion massive de contenus pédocriminels synthétiques (CSAM), qui permettent de masquer l'identité des victimes et des auteurs ;
2. Le développement de *deepfakes* sexuels, souvent utilisés pour le harcèlement, la sextorsion ou la diffusion non consentie ;
3. L'automatisation du *grooming*[176], *via* des outils conversationnels qui facilitent l'approche et la manipulation des mineurs.

Les chiffres[177] du National Center for Missing and Exploited Children (NCMEC)[178] illustrent cette croissance exponentielle dans la production de *deepfakes* :

- En 2023, sur 36 millions de signalements reçus, 4 700 concernaient des contenus générés par IA ;
- **En 2025, le NCMEC a recensé 485 000 signalements liés à des contenus pédocriminels générés par IA, soit une augmentation de 10 538 % en deux ans**.

Les vidéos synthétiques (entièrement générées par IA) constituent un nouveau phénomène inquiétant : selon l'*Internet Watch Foundation*, 3 443 vidéos d'abus sexuels sur mineurs générées par IA ont été détectées en 2025, contre 13 en 2024, soit une augmentation de 26 385 %[179].

---

[174] Security Hero. State of Deepfakes. 2023.
[175] United States House Of Representatives. « 2025 Worldwide Threat Assessment ». Sous-commission des services armés sur le renseignement et les opérations spéciales. Mai 2025.
[176] Le *grooming* est la sollicitation d'enfants à des fins sexuelles en ligne. C'est la situation dans laquelle un adulte prend contact, via un moyen de communication, avec un enfant de moins de 15 ans dans le seul but d'obtenir des avantages sexuels ou de commettre un abus à son encontre. (Source: Pointdecontact.net)
[177] National Center for Missing & Exploited Children (NCMEC). « CyberTipline Report ». 2025.
[178] Le NCMEC est une organisation non lucrative fondée en 1984 par le Congrès des États-Unis. Elle s'occupe d'affaires de mineurs disparus ou exploités, âgé de moins de vingt-un ans.
[179] Internet Watch Foundation (IWF). « AI-Generated Child Sexual Abuse: 2026 Report on Trends, Data & Human Impact ». 2026.

Cette tendance est aggravée par la contamination des jeux de données d'entraînement : des recherches de l'université de Stanford ont identifié des **centaines d'images pédocriminelles dans les bases de données utilisées pour entraîner des modèles populaires comme Stable Diffusion**, ce qui signifie que *« les modèles eux-mêmes sont partiellement construits sur du matériel d'abus »*[180].

Cette croissance exponentielle des contenus illicites a plusieurs conséquences. D'une part, les acteurs de la protection de l'enfance qui disposent de moyens particulièrement contraints, sont confrontés à un véritable « raz-de-marée » de signalements. En effet, l'essor de l'IA générative complique l'identification des mineurs réellement en danger, car il devient de plus en plus difficile de distinguer une image authentique d'une image générée artificiellement. Cette confusion ralentit les enquêtes et complique le travail des professionnels. D'autre part, le caractère virtuel de ces montages entraîne chez leurs auteurs et consommateurs une tendance à la banalisation et à la normalisation de ces pratiques. Comme le souligne la Fondation pour l'Enfance dans son rapport « L'IA générative, nouvelle arme de la pédocriminalité »[181], cette déshumanisation des victimes, même virtuelles, contribue à normaliser des comportements criminels et à minimiser leur gravité.

Plus encore, **les contenus synthétiques offrent de nouveaux leviers aux pédocriminels.** L'IA générative facilite le *grooming* en permettant la création de faux profils, afin de gagner la confiance d'un mineur et de l'amener à divulguer des informations sensibles ou à envoyer des contenus intimes. Elle amplifie également la *sextorsion*, soit le « *chantage réalisé à l'aide de contenus (images ou vidéos) de la victime et présentant un caractère sexuel, en vue de lui extorquer des faveurs sexuelles, de l'argent ou tout autre avantage, en menaçant de les diffuser sans son consentement* »[182]. La Fondation pour l'Enfance alerte ainsi sur le fait que le *grooming*, la sextorsion et la production de contenus pédocriminels sont désormais *« facilités et amplifiés »* par ces technologies.

En outre, les mineurs sont parfois aussi auteurs de *deepfakes*. Une enquête de Thorn[183] indique en effet qu'**un adolescent américain sur dix a déjà utilisé l'IA pour créer des images intimes non consenties**. Églantine Cami, chargée de plaidoyer de CAMELEON Association France, souligne : *« pour eux, les deepfakes ne sont pas des violences, cette pratique est commune et banalisée. Pourtant, même si c'est pour rire, il y a une réelle crainte chez les jeunes de voir des montages d'eux sortir. Pour se protéger de cela, ils ont eux-mêmes recours à des montages pour avoir une monnaie de chantage* »[184]. **Ce phénomène prend une ampleur inquiétante dans les milieux scolaires où il peut nourrir le harcèlement et le cyberharcèlement**. Une étude[185] conjointe de l'UNICEF, d'ECPAT et d'INTERPOL, menée dans 11 pays, montre que 1,2 million d'enfants ont déclaré avoir été victimes de *deepfakes* à caractère sexuel en 2025. Dans certains pays, cela représente un enfant sur 25, soit l'équivalent d'un élève par classe.

---

[180] THIEL, David. Identifying and Eliminating CSAM in Generative ML Training Data and Models. *Stanford Digital Repository*. 2023. vol. 23, no 3, p. 131.
[181] Fondation pour l'Enfance. « L'IA générative, nouvelle arme de la pédocriminalité ». Octobre 2024.
[182] Fondation pour l'Enfance. *Ibid.*
[183] Thorn. « Deepfake Nudes & Young People: Navigating a New Frontier in Technology-Facilitated Nonconsensual Sexual Abuse and Exploitation ». *Thorn Repository*. 3 mars 2025.
[184] Fondation pour l'Enfance. *Op. cit.*
[185] UNICEF. « L'utilisation abusive de deepfakes est une forme d'abus ». 4 février 2026.

Ces dérives ont montré la nécessité d'intégrer pleinement la problématique des *deepfakes* au sein du RIA *via* l'Omnibus[186], afin de renforcer les obligations applicables aux systèmes générant des contenus intimes non consentis.

La répression des *deepfakes* pédocriminels et sexuels non consentis

L'Omnibus numérique en matière d'IA[187] a consacré, au sein de l'article 5 du RIA, une interdiction de la mise sur le marché, de la mise en service et de l'utilisation d'un système d'IA qui génère ou diffuse des *deepfakes* pédocriminels et sexuels non consentis, applicable à compter du 2 décembre 2026.

Ce texte suppose dorénavant une articulation avec les dispositions pénales existantes dans le droit français, mais également avec les dispositions futures, telles qu'envisagées par la proposition de loi intégrale[188].

Recommandation n° 10 : Renforcer la répression de la création de contenus à caractère sexuel non consentis et pédocriminels truqués (hypertrucages) générés par intelligence artificielle. La présente Commission soutient la proposition de loi dite intégrale qui vise à poursuivre individuellement les auteurs de contenus générés par IA à caractère sexuel non consentis ou pédocriminels. Cet ajout est complémentaire à l'interdiction posée à l'article 5 de l'Omnibus numérique en matière d'IA.

c) Usurpations d'identité et escroqueries : l'usage des *deepfakes* dans le champ de la fraude en ligne

L'IA générative et sa capacité à produire des *deepfakes* d'un réalisme inédit créent également de nouvelles vulnérabilités en matière de sécurité financière. En facilitant l'exploitation de la confiance des individus, ces technologies renforcent les risques d'arnaques ciblées, de campagnes

[186] Union Européenne. « Règlement (UE) 2026/1744 du Parlement européen et du Conseil du 8 juillet 2026 modifiant les règlements (UE) 2024/1689, (UE) 2018/1139 et (UE) 2023/1230 en ce qui concerne la simplification de la mise en œuvre des règles harmonisées concernant l'intelligence artificielle (train de mesures Omnibus numérique en matière d'IA) ». *Journal officiel de l'Union européenne*. 8 juillet 2026.
[187] Omnibus numérique en matière d'IA.
[188] Texte de la commission sur la proposition de loi de Mme Céline Thiébault-Martinez et plusieurs de ses collègues apportant une réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants (3106). n° 3190-A0 rectifié, déposé le mercredi 23 septembre 2026.

d'hameçonnage[189] particulièrement sophistiquées et d'usurpation d'identité, notamment à travers des comptes clonés ou des imitations vocales et audiovisuelles. Ce qui distingue aujourd'hui cette nouvelle génération de fraudes est moins leur nature que leur facilité de production, leur faible coût et leur capacité à être déployées à grande échelle.

Cette évolution est illustrée par l'expérience menée par la journaliste Lila Shroff pour *The Atlantic*[190]. En testant les capacités de ChatGPT Images 2.0, elle est parvenue à générer en quelques heures plus d'une centaine de faux documents particulièrement réalistes, comprenant notamment des ordonnances médicales, des alertes bancaires, des publications destinées aux réseaux sociaux, des pièces d'identité et des passeports falsifiés. Cette démonstration met en évidence la facilité avec laquelle des documents sensibles peuvent désormais être produits par des outils d'IA générative, réduisant considérablement les barrières techniques qui limitaient auparavant ce type de fraude.

L'IA générative permet ainsi d'industrialiser les activités frauduleuses en automatisant la production de contenus trompeurs, tout en les rendant plus convaincants et moins coûteux à réaliser. Anya Schiffrin, professeure à l'Université de Columbia, souligne que ces escroqueries s'appuient sur de véritables chaînes de fraude organisées, combinant de fausses publicités diffusées sur les réseaux sociaux, des échanges privés avec les victimes et des demandes de paiement, souvent effectuées en cryptoactifs (audition du 24 avril 2026).

La fraude aux moyens de paiement repose aujourd'hui principalement sur des techniques de manipulation renforcées par l'IA générative dont trois principaux usages malveillants peuvent être distingués selon Henri-Olivier Fliche, directeur de l'Innovation, des Données et des Risques technologiques à l'Autorité de contrôle prudentiel et de résolution (ACPR) (audition du 18 mai 2026) :

1. **La collecte des données personnelles et bancaires**. Les fraudeurs cherchent à convaincre les victimes de communiquer leurs identifiants ou leurs coordonnées bancaires en usurpant l'identité d'organismes de confiance au moyen de courriels, de SMS ou de publicités frauduleuses. Ces messages renvoient vers des sites miroirs qui reproduisent fidèlement l'apparence des sites officiels. Les systèmes d'IA améliorent considérablement l'efficacité de ces attaques en supprimant les fautes de langue qui constituaient auparavant un indice de fraude et en reproduisant avec précision les codes graphiques des sites légitimes. La création de ces copies conformes est désormais réalisable en quelques clics, permettant une industrialisation sans précédent des campagnes de *phishing*.
2. **L'interaction avec la victime.** Grâce aux *deepfakes* vocaux ou audiovisuels, les fraudeurs sont en mesure d'usurper l'identité d'un dirigeant d'entreprise, d'un conseiller bancaire ou encore d'un proche afin d'obtenir un virement ou la communication d'informations confidentielles.

[189] L'hameçonnage ou *phishing* en anglais est le principal mode opératoire utilisé par les cybercriminels pour dérober des informations personnelles et/ou bancaires aux internautes. Par message électronique (e-mail), SMS ou encore par téléphone, il consiste à usurper l'identité d'un tiers de confiance (banque, administration, réseau social, entreprise de livraison, commerce en ligne…) pour tromper la victime et l'inciter à communiquer ses données d'identité, ses mots de passe ou ses numéros de carte bancaire (Cybermalveillance.gouv).

[190] SHROFF, Lila. « Deepfakes Are Coming for Your Bank Account ». *The Atlantic*. 2 mai 2026.

3. **L'ouverture de comptes frauduleux destinés à recevoir les fonds détournés.** Les fraudeurs utilisent des outils de génération d'images et de documents pour produire de fausses pièces justificatives permettant de contourner les procédures de vérification d'identité mises en œuvre par les établissements financiers. L'automatisation de la fabrication de faux justificatifs contribue ainsi à alimenter l'ensemble de la chaîne de fraude.

L'ampleur exacte de ce phénomène demeure toutefois difficile à mesurer. Comme l'a souligné Julien Lasalle, secrétaire de l'Observatoire de la sécurité des moyens de paiement de la Banque de France (audition du 18 mai 2026), il est souvent impossible d'identifier précisément la part jouée par l'intelligence artificielle dans chaque fraude. Néanmoins, les fraudes dites *« par manipulation »*, fondées sur les *deepfakes*, les sites miroirs ou les usurpations d'identité, sont aujourd'hui largement facilitées par ces technologies. Cette tendance se reflète également dans les statistiques internationales. Dans son dernier rapport annuel consacré à la cybercriminalité, le Federal Bureau of Investigation (FBI) indique que les escroqueries impliquant l'IA auraient causé près d'un milliard de dollars de pertes aux États-Unis au cours de l'année écoulée[191].

Face à l'industrialisation de la fraude, les limites du cadre juridique actuel, qui responsabilise avant tout l'individu et repose principalement sur sa vigilance, deviennent de plus en plus évidentes. En effet, lorsque les contenus frauduleux deviennent quasi indiscernables des contenus authentiques, il n'est plus réaliste ni équitable de faire reposer l'essentiel de la prévention sur les seules épaules des victimes potentielles. **Les réponses à la fraude financière facilitée par les *deepfakes* doivent dès lors s'orienter vers une logique de responsabilité institutionnelle**[192].

2) L'IA générative peut se muer en support de la haine en ligne

D'après Shani Benoualid, conseillère en charge du numérique et de la lutte contre la haine en ligne auprès de la Délégation interministérielle à la lutte contre le racisme, l'antisémitisme et la haine anti-LGBT (DILCRAH), *« les risques liés à l'IA générative en matière de haine en ligne connaissent une évolution majeure, notamment en raison d'un changement d'échelle : l'IA transforme l'économie de la haine »* (audition du 17 avril 2026). **En permettant une production massive et à moindre coût de contenus haineux, l'IA générative participe à l'industrialisation de la haine**, modifiant les pratiques et les stratégies de diffusion.

Ces contenus générés par IA, produits avec une grande facilité, se diffusent ensuite rapidement sur les réseaux sociaux et les services de plateforme de partage de vidéo en ligne, où la logique de l'économie de l'attention favorise la viralité des contenus choquants ou humiliants.

Les personnes précédemment visées par la haine en ligne sont celles qui risquent d'être particulièrement vulnérables face aux contenus générés par IA. À ce titre, la DILCRAH identifie plusieurs groupes particulièrement exposés : les femmes et les filles qui figurent parmi les premières

[191] Federal Bureau of Investigation (FBI). « Internet Crime Report 2025 ». *Internet Crime Complaint Center (IC3)*. 2025.
[192] SCHIFFRIN, Anya, MARWICK, Alice E., SINHA, Navya, WANGNOO, Anusha, WILLIAMS, Kaylee, HUSEYNOVA, Elnara et HATFIELD, Audrey. « Deepfake Financial Fraud: The Global Regulation of AI-Driven Scams ». *Data & Society*, note politique. 2 mars 2026.

victimes de la haine en ligne[193] ; les minorités déjà ciblées par des campagnes de haine ou encore les personnes fortement exposées publiquement (célébrités, personnalités politiques, journalistes, militants, *etc.*), qui constituent des cibles privilégiées de contenus synthétiques diffamatoires, humiliants ou trompeurs[194] (audition du 17 avril 2026).

Si les *deepfakes* jouent, un rôle actif dans la diffusion de la haine en ligne, celle-ci peut également être présente en amont, au sein même des réponses aux *prompts* des utilisateurs de LLM. Ceci peut être intentionnel ou accidentel. L'entraînement de ces modèles sur le corpus du *web* les conduit *de facto* à reproduire, voire amplifier les contenus qui s'y trouvent, incluant les contenus haineux, sans nécessaire intention explicite sous-jacente. Le modèle Grok en a illustré les risques en novembre 2025, en diffusant sur la plateforme X des propos remettant en cause l'usage des chambres à gaz à Auschwitz et reprenant des éléments caractéristiques du négationnisme[195]. Cet épisode montre que les biais, les défaillances de modération ou les insuffisances des garde-fous intégrés aux systèmes d'IA générative peuvent contribuer à la production et à la circulation de contenus haineux. Ce caractère haineux peut aussi être intentionnel et des acteurs peuvent utiliser des briques de modèles librement accessibles pour les raffiner en les entraînant spécifiquement pour qu'ils génèrent des réponses toxiques.

Les systèmes d'IA générative sont ainsi également de nouveaux outils dans l'arsenal des auteurs de haine en ligne pour produire à moindre coût des contenus nocifs, par exemple en leur permettant de générer des textes radicalisés visant certaines catégories de la population[196]. Cela peut aussi servir à traduire massivement des contenus haineux pour élargir l'audience de diffusion, en particulier dans des langues peu dotées dont la modération sur les plateformes en ligne est souvent défaillante[197].

De même, ces systèmes sont le support de l'automatisation de comptes inauthentiques pouvant être utilisés pour inonder les réseaux sociaux de messages haineux. Il en va de même pour la création de sites *web* totalement générés artificiellement et véhiculant des thèses et contenus haineux. À titre d'exemple, NewsGuard recense plus de 3 700 sites d'information alimentés en grande partie par IA, avec une supervision humaine minimale, dans 16 langues[198].

De façon plus pernicieuse, ces agents conversationnels peuvent être utilisés par les acteurs malveillants pour réécrire les contenus haineux de sorte à contourner la modération des plateformes numériques, en particulier des réseaux sociaux, par exemple par l'usage de certains caractères

[193] Chowdhury, Rumman, et Dhanya Lakshmi. *« Ton avis ne compte pas, de toute façon » : dénoncer la violence de genre facilitée par la technologie à l'ère de l'intelligence artificielle générative*. Paris : UNESCO, coll. « Tendances mondiales en matière de liberté d'expression et de développement des médias ». 2024.
[194] HICKS, Jacqueline. « Global Evidence on the Prevalence and Impact of Online Gender-Based Violence (OGBV) ». The Institute of Development Studies and Partner Organisations. 2021.
[195] HERRERO, Océane et DURAND, Klara. « French Authorities Probe Holocaust Denial on Elon Musk's AI Platform ». *Politico*. 19 novembre 2025.
[196] UNESCO. « AI and the Holocaust: Rewriting History? The Impact of Artificial Intelligence on Understanding the Holocaust ». Rapport rédigé par MAKHORTYKH, Mykola, sous la direction de MANN, Heather. 2024. P. 26.
[197] SHAHID, Farhana, ELSWAH, Mona et VASHISHTA, Aditya. « Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages ». arXiv. 5 août 2025.
[198] NewsGuard, AI Tracking Center. 3 749 sites recensés dans 16 langues, mise à jour du 23 juin 2026.

spéciaux dans les mots les plus sensibles, leur remplacement par des émojis ou encore le recours à des synonymes connotés.

Cette diffusion de contenus haineux générés par IA renouvelle donc la question de la modération en ligne des contenus, épineuse question à ce jour non résolue sur les réseaux sociaux. Une difficulté particulière, partagée par une grande partie des auditionnés, tient à l'identification de ces contenus, rendant leur modération et leur signalement complexes, en particulier dans un contexte où la modération repose de plus en plus sur les communautés. Ces contenus sont aussi difficiles à distinguer des productions humaines : dans une étude, les participants ont confondu des textes générés par IA avec des textes humains dans 77 % des cas[199]. Le volume et le réalisme de ces contenus peuvent ainsi dépasser les capacités de la modération humaine et des structures d'accompagnement.

La manipulation des systèmes d'IA à des fins de propagation de la haine en ligne interroge aussi la responsabilité des acteurs. Lorsqu'un contenu haineux est généré artificiellement, à qui doit-on imputer la faute ? Est-ce le développeur initial du modèle, le déployeur, l'utilisateur qui a formulé la requête ou encore la plateforme qui en permet la diffusion ?

Enfin, la tendance à la flagornerie des systèmes d'IA générative peut entretenir l'utilisateur dans ses croyances haineuses, valider, voire approfondir sa vision du monde plutôt que de les contredire[200], alimentant une chambre d'écho discriminatoire individuelle.

*

Ainsi, les effets pernicieux de l'IA générative étudiés dans ce chapitre à l'échelle individuelle (externalisation des processus cognitifs, attachement émotionnel, hyperpersonnalisation, prolifération des *deepfakes*, érosion de la confiance et altération de la perception du réel, *etc.*) se manifestent également à l'échelle collective. Ces mêmes mécanismes s'empilent, s'amplifient et s'imprègnent à l'échelle sociétale. Leur généralisation pourrait, à terme, fragiliser la cohésion sociale, transformer notre rapport à l'information et affecter les conditions mêmes de l'exercice démocratique.

[199] BENGIO, Yoshua *et al. International AI Safety Report 2026*. Février 2026.
[200] Dukić, Simeon, Natalie Vela et Siddharth Venkataramakrishnan. Radicalisation in Closed Loops: Risks and Intervention Opportunities with AI Chatbots and Companions. Londres : Institute for Strategic Dialogue. 2026.

CHAPITRE 2

Cohésion sociale et exercice démocratique : la confiance collective à l'épreuve de l'IA générative

L'IA générative, par son imprégnation dans tous les pans de la société, redéfinit notre conception de la cohésion sociale, au point de bouleverser l'exercice démocratique.

La cohésion sociale *« se définit, à la fois, comme un but de société et comme un moyen des politiques publiques. Elle est le problème et la solution »* [201]. Elle désigne les liens[202] qui rassemblent une société et qui unissent les membres d'un groupe aux autres. Au niveau européen, le Conseil de l'Europe propose de la définir comme *« la capacité d'une société à assurer le bien-être de tous ses membres, à minimiser les disparités et à éviter la polarisation »*[203]. Une autre notion pertinente, mais plus politique et philosophique, est celle du **pacte social ou contrat social**. Cette notion désigne les « *pactes et ces arrangements sociaux qui font tenir notre vie collective, ces compromis qui structurent la société et qui peuvent toujours être renégociés, car aucun contrat social n'est définitif, aucun contrat social n'est inéluctable* »[204].

Que l'on parle de cohésion, de pacte ou de contrat social, ces notions sont utiles pour questionner l'ensemble des bouleversements engendrés par l'IA générative sur les différentes dimensions que ces concepts recouvrent. **L'IA générative s'inscrit déjà au cœur des dynamiques de la vie collective et interroge la manière dont nous nous représentons la société et ce qui nous relie en tant que groupe**.

[201] AVENEL, Cyprien. La « cohésion sociale »: de quoi parle-t-on ? Clarifier le concept pour consolider un nouveau modèle d'action. In : *L'année de l'action sociale 2015: Objectif autonomie*. Dunod. 2014. p. 119-136.
[202] FORSYTH, D. R. Components of cohesion. *Group Dynamics, 5th Edition. Wadsworth: Cengage Learning*, 2010. p. 118-122.
[203] Conseil de l'Europe. « Une nouvelle stratégie de cohésion sociale ». 2004.
[204] IDRI. « Le contrat social : un cadre pour repenser ce que nous voulons en commun ». 20 juin 2024.

A) De la reproduction à l'accélération des préjugés : les biais de l'IA générative altèrent la cohésion sociale

L'IA générative, par son fonctionnement même, reproduit et contribue ainsi à renforcer les stéréotypes préexistants, au risque d'engendrer des discriminations concrètes et d'invisibiliser et d'uniformiser les cultures et les langues.

1) De sa conception à son utilisation, l'IA générative comporte des biais

Les systèmes d'IA, tout comme les humains, sont sujets à l'erreur, et ne sont pas neutres. En effet, **l'IA générative, en tant que modèle computationnel[205], génère des résultats** (texte, image, vidéo, *etc.*) **qui demeurent imparfaits et comportent des biais sociaux, historiques, idéologiques qui reproduisent et amplifient les discriminations et les stéréotypes existants dans la société**.

Ces limites algorithmiques (ou biais) se retrouvent tant au stade du développement de l'IA qu'au niveau des décisions de conception et de modélisation, jusqu'à la collecte et au traitement de données, en passant par le contexte de déploiement. L'UNESCO[206] classe ces biais en trois catégories : (i) les biais liés aux données ; (ii) les biais liés aux choix de l'algorithme et (iii) les biais liés au déploiement de l'IA.

1. **Les corpus d'entraînement des modèles d'IA générative ne constituent pas un matériau neutre**. Ils agrègent de très vastes ensembles de données collectées par exploration automatisée du *web*, de contenus générés par les utilisateurs, de bases documentaires et de données issues d'interactions numériques qui portent eux-mêmes les traces des inégalités, des stéréotypes et des discriminations présentes dans la société. Le biais n'apparaît donc pas d'abord comme une propriété technique de l'algorithme ; il prend racine dans les pratiques humaines, par nature non neutres, les rapports sociaux et les choix éditoriaux qui structurent la production, la diffusion et la valorisation des contenus en ligne. Cette situation est renforcée par les ressorts de l'économie de l'attention qui structure la visibilité des contenus en ligne. En effet, comme rappelé *supra*, les contenus les plus visibles en ligne ne sont pas nécessairement les plus exacts, pluralistes ou représentatifs : ils sont souvent ceux qui suscitent l'engagement, des réactions émotionnelles, la controverse ou le partage (à l'instar des contenus haineux, stigmatisants ou polémiques). Ainsi, **lorsqu'un modèle d'IA est entraîné par un *crawling*[207] massif du *web*, il n'apprend donc pas d'une image statistiquement**

[205] *« Un modèle computationnel est donc un modèle qui fait des calculs à partir d'informations données en entrée, et qui donne en sortie un résultat numérique ».* COLLINS Anne, KHAMASSI Mehdi. « Initiation à la modélisation computationnelle ». In : Khamassi, M. (Ed.) *Neuro-sciences cognitives.* 2021.

[206] UNESCO. *Challenging Systematic Prejudices - An Investigation into Gender Bias in Large Language Models.* 2024. p.3.

[207] Selon la définition de la CNIL, *« le crawling est un processus automatisé par lequel des programmes, appelés crawlers ou robots d'exploration, parcourent de très larges portions du web afin de collecter des pages, des textes, des métadonnées et d'autres contenus accessibles en ligne. Ces données brutes servent notamment à constituer, enrichir ou actualiser les corpus utilisés pour l'entraînement et l'amélioration des modèles d'intelligence artificielle ».*

neutre du monde mais à partir d'un espace informationnel déjà hiérarchisé par des dynamiques de visibilité et empreint de représentations sociales antérieures biaisées.

En outre, de nombreux biais proviennent du manque de représentativité et de diversité dans les jeux de données utilisés pour l'entraînement. Cela a des conséquences sur le plan technique, car les modèles d'IA utilisent des « métriques de vraisemblance » (indicateurs évaluant la probabilité qu'une phrase ou qu'un comportement soit « normal » ou attendu). Ces métriques ont tendance à marquer comme erronées les séquences peu probables (c'est-à-dire les phrases, actions ou comportements qui s'éloignent des normes les plus courantes), participant au renforcement des discriminations envers les personnes concernées.

Une étude[208] soulignait en 2025 que les *chatbots* déconseillaient par exemple d'entretenir une relation sociale, professionnelle ou amoureuse avec une personne souffrant de schizophrénie. **Cette *« psychophobie »* des LLM renforce la stigmatisation des personnes concernées et fragilise leur intégration au corps social.**

2. **Des choix techniques opérés lors de la conception du modèle** (architecture, objectifs d'optimisation, sélection et pondération des données, méthodes de filtrage, règles de sécurité ou procédés d'alignement) peuvent amplifier, atténuer ou orienter certains contenus. Ces choix techniques sont opérés par des équipes de conception, des entreprises ou bien encore des autorités qui possèdent leurs propres hypothèses, intérêts, sensibilités culturelles et conceptions de ce qui est acceptable, utile ou sûr. Ainsi, des biais peuvent intervenir dans la logique opérationnelle et la modélisation elle-même, lorsqu'un concepteur omet, consciemment ou non, certains aspects du contexte dans la construction du modèle[209].

   Les mécanismes d'alignement illustrent particulièrement ce point. Le réglage d'un modèle par renforcement à partir de retours humains (*reinforcement learning from human feedback*, RLHF) consiste à faire évaluer des réponses par des annotateurs humains, puis à entraîner le système à privilégier les sorties jugées préférables selon des critères définis par les développeurs. Or, ces annotateurs et ces critères ne représentent pas l'ensemble des perspectives sociales, culturelles ou politiques existantes ; les travaux sur l'alignement soulignent ainsi que **des échantillons limités de personnes peuvent conduire le modèle à refléter certaines valeurs tout en marginalisant d'autres**[210].

   Ces choix peuvent également prendre la forme de filtres destinés à occulter, déclasser ou mettre en avant certaines informations. De tels dispositifs ne sont pas intrinsèquement

[208] MOORE Jared, GRABB Declan, AGNEW William, KLYMAN Kevin, CHANCELLOR Stevie, ONG Desmond C. , HABER Nick. « Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers ». *ACM Conference on Fairness, Accountability, and Transparency*. 2025.
[209] OCDE. « Explanatory Memorandum on the Updated OECD Definition of an AI System ». *OECD Artificial Intelligence Papers*. Mars 2024. no 8.
[210] HUANG, Linus Ta-Lun, PAPYSHEV, Gleb et WONG, James K. « Democratizing Value Alignment: From Authoritarian to Democratic AI Ethics ». *AI Ethics*. 2025. vol. 5, pp. 11-18.

illégitimes : ils peuvent viser à réduire les contenus haineux, la désinformation ou les risques pour la sécurité. Toutefois, ces filtres peuvent être aussi le résultat d'une censure politique, comme en témoigne l'absence de réponse de l'agent conversationnel chinois DeepSeek après une question sur le massacre de Tian'anmen[211].

3. **Des biais peuvent apparaître lors de l'utilisation du système**, en fonction des paramètres de configuration, des choix d'interface et des usages effectifs. La formulation des requêtes, le contexte conversationnel, les instructions système et la manière dont les réponses sont présentées peuvent orienter les résultats générés et influencer la perception de leur fiabilité. Un même modèle peut ainsi produire des résultats différents selon le cadrage de la question, tandis qu'une interface qui met en avant une réponse unique ou présente une information avec assurance peut favoriser une adhésion insuffisamment critique. **Le biais de déploiement désigne ainsi le décalage entre le contexte pour lequel un modèle a été conçu et les usages réels pour lesquels il est mobilisé, interprété et intégré à des décisions humaines**. Ces biais d'utilisation concernent aussi les interactions entre utilisateurs et systèmes d'IA[212]. En effet, lors de leurs échanges avec ces outils, les utilisateurs projettent leurs convictions préétablies, leurs attentes et leurs modes de communication. Ces facteurs orientent non seulement la formulation de leurs demandes initiales, mais aussi les questions complémentaires qu'ils posent, ainsi que leur tendance à valider ou à remettre en cause les réponses proposées. En ce sens, le phénomène de flagornerie (cf. chapitre 1, B-1) caractéristique des *chatbots* contribue pleinement à ces biais puisque le système d'IA va tendre à privilégier des réponses conformes aux opinions des utilisateurs, au détriment de la rigueur factuelle.

2) L'IA générative renforce les stéréotypes préexistants et engendre des discriminations concrètes

Ces biais ont des conséquences concrètes. Ils ne se contentent pas de reproduire des représentations sociales préexistantes, ils peuvent les amplifier et les traduire en traitements différenciés, engendrant des effets discriminatoires, renforçant des stéréotypes notamment de genre, ou marginalisant certaines personnes ou certains groupe de personnes. Sarah Bénichou, directrice de la direction Promotion de l'égalité et de l'accès aux droits du Défenseur des droits, souligne en ce sens que *« ces représentations stéréotypées peuvent conduire à une essentialisation des groupes sociaux et renforcer les discriminations »* (audition du 13 avril 2026).

[211] France info. « "Parlons d'autre chose" : on a testé l'IA chinoise DeepSeek, qui n'échappe pas à la censure ». 28 janvier 2025.
[212] ROBINIG, Wolfgang et WALLNER, Johannes P. « Bias in AI: A Framework for Understanding the Sources of Bias in Artificial Intelligence ». *AI Ethics*. 2026. vol. 6, p. 438.

Une étude[213] de l'UNESCO de 2024 en offre l'illustration. Dans des tâches d'association de mots genrés, le système d'IA Llama 2 associe davantage les prénoms féminins à « maison », « famille » et « enfants », tandis que les prénoms masculins sont plus souvent reliés à « affaires », « cadre », « salaire » et « carrière ».

Ainsi, ce sont souvent les mêmes personnes déjà exposées à des discriminations, à une sous-représentation numérique ou à des violences en ligne qui voient leurs stigmatisations amplifiées par les systèmes d'IA générative. Comme le rappelle le collectif ctrl+alt+reclaim, « *les impacts sont profondément intersectionnels. Les jeunes femmes et filles, les personnes LGBTQIA+, les jeunes présentant des troubles psychiatriques préexistants et les jeunes neurodivergents sont de manière disproportionnée ciblés par les biais algorithmiques, le cybersexisme et les conceptions excluantes* »[214].

En somme, il existe un *continuum* entre la violence symbolique véhiculée par les représentations générées et ses conséquences concrètes dans la vie réelle. Ce *continuum* devient particulièrement préoccupant dans un contexte où les systèmes d'IA générative sont intégrés dans un nombre croissant de domaines et prennent une place importante dans la prise de décision. **La littérature sur le sujet montre ainsi que, dans des domaines tels que la justice[215], la reconnaissance faciale[216] ou encore le recrutement[217], des systèmes d'IA générative peuvent discriminer selon des caractéristiques telles que l'origine ou le genre** et transformer des biais en effets matériels sur l'accès aux droits, à l'emploi ou à la reconnaissance sociale.

Par exemple, les recherches du Standford Institue for Human-Centered-Artificial Intelligence[218] révèlent que de grands modèles de langage (par exemple GPT2, GPT3.5 d'OpenAI et GPT4, la RoBERTa de Facebook AI, par exemple) continuent de perpétuer des préjugés raciaux néfastes, en particulier contre les locuteurs de l'anglais afro-américain. Pourtant, ces LLM sont intégrés aux systèmes décisionnels de la vie courante. Les chercheurs ont en effet montré dans des expériences supplémentaires que les LLM sont plus susceptibles d'attribuer des emplois moins prestigieux aux locateurs de l'anglais afro-américain, de les déclarer coupables- d'un crime non commis et de les condamner à mort plus facilement dans le cadre d'une simulation de jugement.

[213] UNESCO. *Challenging Systematic Prejudices - An Investigation into Gender Bias in Large Language Models*. 2024. p.3.
[214] Contribution internationale de ctrl+alt+reclaim. 20 avril 2026.
[215] SOCOL DE LA OSA, David Uriel et REMOLINA, Nydia. « Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing-or Misinforming-Judicial Decision-Making through Generative AI ». Data & Policy, 2024, vol. 6, p. e59.
[216] BUOLAMWINI, Joy et GEBRU, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In : *Conference on fairness, accountability and transparency*. PMLR, 2018. p. 77-91.
[217] AN, Jiafu, HUANG, Difang, LIN, Chen et TAI, Mingzhu. « Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation ». PNAS Nexus, 2025, vol. 4, no 3, p. pgaf089.
[218] Human-Centered-Artificial Intelligence. « Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes ». Septembre 2024.

3) L'IA générative peut porter atteinte aux diversités culturelles et linguistiques

a) Les effets d'une sous-représentation de la diversité linguistique

Les défis liés aux biais des systèmes d'IA générative se présentent également sur le plan linguistique avec une sous-représentation de certaines langues dans les données d'entraînement. En effet, les modèles de langage sont entraînés sur des corpus massivement dominés par un petit nombre de langues dites « à hautes ressources » (*high-ressource languages*). Cette concentration laisse de côté la grande majorité des quelque 7 000 langues parlées dans le monde, dont beaucoup sont classées comme « à faibles ressources » (*low resource languages*) et qui risquent pour certaines de disparaître, faute de numérisation. En effet, alors que seulement 5 % de la population mondiale parle anglais nativement[219], la moitié des contenus sur Internet sont anglophones[220]. Ceci se reflète dans les données d'entraînement d'IA générative[221]. Les conséquences de ce déséquilibre sont concrètes : **ces outils sont plus performants en anglais[222] et perdent 20 à 30 % de la qualité des réponses lorsqu'ils sont interrogés dans une autre langue**, et ce même s'ils le sont sur des sujets propres à d'autres cultures. Ainsi, il existe un risque que les locuteurs de langues « non majoritaires » reçoivent une plus grande proportion de réponses erronées, d'hallucinations et de fausses informations[223].

Par ailleurs, **les coûts d'usage de l'IA pourraient varier selon la langue employée.** Une étude de la Coalition for Language Equity in AI Research[224] montre que certains modèles nécessitent beaucoup plus de *tokens* pour traiter une même information selon la langue. Les chercheurs parlent d'un « impôt de tokenisation » caché, qui augmente les coûts d'usage de l'IA dans les langues moins présentes en ligne, au risque de créer des inégalités économiques pour les utilisateurs parlant des langues « à faibles ressources ».

Enfin, la question de la représentativité linguistique inclut également celle des stéréotypes linguistiques racistes, qui peuvent engendrer des biais cumulatifs.

L'Allen Institute for Artificial Intelligence[225] révèle que les modèles d'IA tendent à évaluer négativement l'intelligence et l'employabilité des locuteurs « Ebonics » (dialecte anglais Ebonics, de l'anglais modifié, utilisé par les Afro-américains) par rapport à ceux qui utilisent l'anglais américain standard. Les modèles sont plus susceptibles d'attribuer des qualificatifs négatifs tels que « stupides » et « paresseux », et de les orienter vers des emplois moins bien

[219] Population Hub. « The Global Reach of English: An Analysis of Worldwide Speakers in 2025 ». Octobre 2025.
[220] W3Techs.Usage statistics of content languages for websites. Consulté le 6 août 2026.
[221] 67 % des données d'entraînement de LLaMA est l'anglais ; 93 % les données d'entraînement de GPT-3 étaient l'anglais.
[222] Les risques de biais sont plus élevés en langue non anglaise. NAVIGLI Roberto, CONIA Simone, ROSS Björn. « Biases in Large Language Models: Origins, Inventory and Discussion ». Journal of Data and Information Quality. 2023.
[223] GUERREIRO, Nuno M., ALVES, Duarte M., WALDENDORF, Jonas, HADDOW, Barry, BIRCH, Alexandra, COLOMBO, Pierre et MARTINS, André F. T. « Hallucinations in Large Multilingual Translation Models ». *Transactions of the Association for Computational Linguistics*. 2023. vol. 11, pp. 1500-1517.
[224] PERRY, Noah, KANTHARUBAN, Anjali, LUGER, Sarah et GUZMÁN, Francisco. « The Tokenizer Tax: Why Using the Same Model Could Cost up to 17x More in Emerging Markets ». *Coalition for Language Equity in AI Research*, 29 septembre 2026.
[225] HOFMANN Valentin, RIA KALLURI Pratyusha, JURAFSKY Dan, KING Sharese. » Dialect prejudice predicts AI decisions about people's character, employability, and criminality ». Mars 2024.

rémunérés. Cette discrimination s'étend même aux recommandations de peine de mort pour des accusés criminels hypothétiques s'exprimant en Ebonics.

b) Vers une uniformisation culturelle et linguistique ?

La surreprésentation de langues dites « à hautes ressources », au premier rang desquelles l'anglais, dans les corpus des principaux modèles de langage soulève la question d'un possible mouvement d'uniformisation culturelle et linguistique. Cette dynamique ne signifie pas nécessairement que les autres langues disparaîtraient, mais qu'elles risquent d'être progressivement reconfigurées à partir de normes d'expression, de références culturelles et de hiérarchies de visibilité dominantes.
D'une part, l'uniformisation linguistique se décèle par le modèle stylistique des systèmes d'IA générative. En effet, ils produisent des réponses marquées par des régularités statistiques, une structuration de l'information et un vocabulaire plus uniforme que les productions humaines. Cette homogénéité peut ensuite se transmettre aux utilisateurs. Lorsqu'ils reprennent, reformulent ou s'inspirent de contenus générés, ils peuvent intégrer progressivement des bribes de vocabulaires et de tournures propres aux systèmes d'IA, ce qui pourrait à terme menacer la diversité de l'expression humaine[226]. À ce titre, une étude portant sur plus de 880 000 textes montre que l'usage de systèmes d'IA générative est associé à une baisse de la diversité stylistique, **la réécriture par IA réduisant la variance stylistique de 21 à 50 % selon les corpus et les modèles**[227].

Ce risque pourrait être renforcé par la progression des contenus et des sites entièrement ou largement générés par l'IA. Une étude du Pew Research Center[228] montre que la part des pages *web* en anglais étudiées présentant des signes significatifs de rédaction ou de révision par IA a augmenté depuis la sortie de ChatGPT : elle atteint plus d'un tiers des pages publiées après novembre 2022 dans l'échantillon de juillet 2026. Cette situation est d'autant plus paradoxale que cela risque aussi d'appauvrir la qualité des résultats des systèmes d'IA générative à terme. Comme le rappelle le Conseil supérieur de la propriété littéraire et artistique (CSPLA)[229], les modèles ont besoin de contenus humains originaux de qualité et actualisés. Ils ne pourront se contenter de contenus synthétiques.

D'autre part, du point de vue culturel, **les contenus produits par les systèmes d'IA générative s'imprègnent le plus souvent des références culturelles majoritaires, effaçant les spécificités identitaires et réduisant la visibilité des traditions orales ou écrites propres à certaines**

[226] ZANOTTO, Sergio E. et AROYEHUN, Segun. *« Linguistic and Embedding-Based Profiling of Texts Generated by Humans and Large Language Models »*. Dans : Christos CHRISTODOULOPOULOS, Tanmoy CHAKRABORTY, Carolyn ROSE et Violet PENG (dir.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou: Association for Computational Linguistics. Novembre 2025. pp. 22841-22858.
[227] SOURATI, Zhivar, KARIMI-MALEKABADI, Farzan, OZCAN, Meltem, MCDANIEL, Colin, ZIABARI, Alireza, TRAGER, Jackson, TAK, Ala N., CHEN, Meng, MORSTATTER, Fred et DEHGHANI, Morteza. *« The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models »*. *Nature Human Behaviour*. 2026. vol. 10, no 10, pp. 1925-1942.
[228] BESTVATER, Samuel, SMITH, Aaron, TERBUSH, Carson, BARONAVSKI, Chris et CHAVDA, Janakee. « How Much of the Internet Is Written With AI? ». *Pew Research Center*. 20 août 2026.
[229] CSPLA. Rapport de mission relatif au statut des productions de l'IA. Juillet 2026. *Op.cit.*

communautés. Ainsi, ces systèmes génératifs tendent à réduire les diversités culturelles à des clichés : ils ne restituent ni la complexité, ni l'hétérogénéité des cultures qu'ils sont censés représenter.

L'analyse menée par l'ONG Rest of World[230] en donne une illustration. À partir de cinq requêtes génériques « une personne », « une femme », « une maison », « une rue » et « un plat » adaptées à la Chine, à l'Inde, à l'Indonésie, au Mexique, au Nigeria et aux États-Unis, les auteurs ont généré 100 images par combinaison, soit 3 000 images au total. Les résultats montrent une forte tendance au stéréotype : « une personne mexicaine » est le plus souvent représentée comme un homme coiffé d'un sombrero, tandis qu'« une personne indienne » apparaît presque systématiquement comme un homme âgé et barbu. Concernant les représentations liées au Nigeria, la majorité des maisons générées présentaient des signes visibles de délabrement, avec une peinture écaillée ou des matériaux abîmés.

Pour répondre à cette problématique des biais linguistiques et culturels, certains modèles d'IA sont développés dans l'objectif d'une meilleure représentativité linguistique[231]. De la même manière, des initiatives de mise à disposition de catalogues de données numérisées, culturelles et thématiquement riches permettent de rééquilibrer quelque peu l'entraînement des LLM sur des données plus diversifiées. Une telle initiative a notamment été mise en œuvre par la Bibliothèque nationale de France[232], au profit de Mistral, lui permettant l'accès à une grande partie de son catalogue de données publiques, au travers de sa base *Via Gallica*. Ce projet vise à répondre au besoin de meilleure performance en français, mais aussi de transmission des langues et cultures peu dotées, car, comme le souligne le ministère de la Culture, *« la découvrabilité des contenus culturels constitue un enjeu central pour préserver la diversité des expressions culturelles »*[233]. La découvrabilité des contenus consiste en *« la capacité à être repéré dans une base de données lors d'une recherche »*[234].

Recommandation n° 16 : Garantir la représentativité des langues et des cultures minoritaires dans l'IA générative, en mobilisant les corpus et les catalogues de référence. Les modèles d'IA générative doivent intégrer et valoriser la diversité linguistique et culturelle, qui constitue un enjeu de souveraineté et de pluralisme démocratique. Les corpus et catalogues existants sont à valoriser plutôt que de multiplier les nouvelles annotations. Il est en outre nécessaire de cibler les financements sur les langues peu dotées et de soutenir des modèles ouverts adaptés aux langues et cultures minoritaires afin d'être en capacité de mesurer la représentativité linguistique et culturelle des modèles.

[230] TURK, Victoria. « How AI Reduces the World to Stereotypes ». *Rest of World*. 10 octobre 2023.
[231] Nous pouvons citer notamment pour le multilinguisme l'initiative open-source de Current AI, Alpha Chat.
[232] Politico. « *La BNF prête à ouvrir ses archives pour franciser des modèles d'IA* ». Juin 2024.
[233] Ministère de la Culture. « Découvrabilité des contenus culturels francophones dans l'environnement numérique à l'ère de l'IA ». 2025.
[234] Cours des comptes. Audit Flash. « *L'intelligence artificielle et le ministère de la culture : bilan d'étape et défis* ». Septembre 2026.

B) Nouvel intermédiaire de l'information, l'IA générative recompose l'espace public

Les vulnérabilités identifiées dans le raisonnement qui suit tiennent à l'imprégnation de l'IA générative dans l'espace informationnel : elle devient progressivement un nouveau canal d'accès et de diffusion des contenus sur Internet, s'additionnant et/ou se substituant en partie aux moteurs de recherche et aux médias traditionnels, modifiant ainsi les conditions d'exposition des citoyens à la pluralité des points de vue. L'espace informationnel s'en trouve ainsi déstabilisé, avec notamment l'affaiblissement économique des médias traditionnels – bien que certains aient réussi à conclure des partenariats lucratifs – dont les modèles reposent en grande partie sur une audience et des recettes publicitaires désormais captées par de nouveaux intermédiaires technologiques, fragilisant leur capacité à produire une information de qualité. C'est ainsi le pluralisme et la liberté d'expression qui sont en jeu.

1) L'IA générative offre de nouveaux modes d'accès à l'information

a) Un intermédiaire déjà incontournable dans l'accès à l'information

D'après le dernier Baromètre du numérique[235], **la recherche d'information constitue aujourd'hui l'usage le plus répandu de l'IA générative en France : 73 % des utilisateurs effectuent ce type de requêtes au moins une fois par mois, 21 % quotidiennement**. Ce changement de paradigme nécessite néanmoins d'être nuancé. Pour la recherche d'information, les moteurs de recherche restent utilisés en priorité (55 % *versus* 26 % pour l'IA générative). Toutefois, l'écart se réduit chez les jeunes qui utilisent quasiment autant les moteurs de recherche que les systèmes d'IA générative (40 % des 18-24 ans utilisent un moteur de recherche et 36 % un système d'IA générative)[236]. Les médias traditionnels, télévision, radio et presse, restent en outre une source d'information importante et une référence vers laquelle le public se tourne pour vérifier les contenus (même si les réseaux sociaux et les plateformes vidéo les dépassent désormais en moyenne).

La manière dont l'IA générative délivre et restitue l'information n'est pas neutre. Si les moteurs de recherche traditionnels type Google Search, Microsoft Bing ou Baidu hiérarchisent déjà les sources présentées, l'utilisateur reste à même de les comparer[237]. Les systèmes d'IA générative marquent un changement de paradigme en produisant une réponse unique, souvent formulée avec une assurance rhétorique qui peut masquer l'incertitude ou la contestabilité des informations sous-jacentes. Cette restitution repose sur des choix éditoriaux opaques, voire inconnus, pour l'utilisateur final, tels que la sélection des sources mobilisées et la pondération des données d'entraînement dans la génération de la réponse. Par exemple, du fait de l'accord entre *Le Monde* et OpenAI, quand ChatGPT renvoie

[235] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[236] *Ibid.*
[237] Les moteurs de recherche classiques, produisent une liste de sources (pages web, documents, *etc.*) censé trouver une source pertinente d'information à partir d'une requête de l'utilisateur. La hiérarchisation des résultats proposés est effectuée à travers des métriques classiques comme TF-IDF, BM25 ou l'algorithme PageRank.

9,9 millions de fois vers un site d'information, 25 % de ces renvois vont vers *Le Monde*, puis vers le *New York Times*, avec lequel l'entreprise a également conclu un accord[238]. Le manque de traçabilité systématique des sources, ou leur citation partielle ou même erronée (sources inexistantes, « inventées » par le système d'IA générative), complique par ailleurs toute vérification *a posteriori*.

Aussi, les utilisateurs ont une propension à moins vérifier les sources lorsque celles-ci sont citées par l'IA générative[239]. Ce mode de délivrance de l'information, à travers un simple dialogue en langage naturel, tend ainsi à **déplacer la charge critique de l'utilisateur** : alors que la navigation sur un moteur de recherche exige un exercice actif de comparaison des sources, l'interaction avec un système d'IA générative invite à une forme de confiance déléguée, où la vraisemblance et la fluidité du discours produit se substituent, dans l'esprit de l'utilisateur, aux critères habituels de fiabilité. Cette dynamique n'est pas anodine alors que la capacité des citoyens à confronter des points de vue divergents et à exercer un jugement critique constitue une condition essentielle du débat public.

Comme le souligne Kamel El Hilali, docteur en droit et spécialiste de l'IA : *« Quand vous utilisez un système comme ChatGPT, Claude, Vibe ou DeepSeek, la réponse que vous obtenez crée votre environnement intellectuel. Elle crée votre cadre de réflexion. Donc, quand ce système vous propose trois options en réponse à une requête, la première question que vous devez vous poser n'est pas de savoir quelle option choisir, mais pourquoi trois options ? Pourquoi ces trois-là ? Pourquoi pas des alternatives ?* In fine*, ces outils conditionnent votre chemin de pensée, votre cadre de réflexion et influencent vos décisions »* (audition du 25 juin 2026).

Chine Labbé, rédactrice en chef de *NewsGuard,* complète ainsi : *« L'IA joue ce rôle de guide, mais au lieu de vous accompagner et de contextualiser, elle synthétise, sélectionne des contenus, au risque de laisser de côté des points qui peuvent être marginaux mais qui peuvent créer la pensée et être intéressants, et au risque de déformer, d'halluciner et de synthétiser des sources peu fiables sans que vous soyez au courant et donc en omettant du contexte »* (audition du 16 juin 2026).

Enfin, **les systèmes d'IA générative redessinent également en profondeur l'espace public informationnel lui-même** : la sphère d'information à laquelle chaque utilisateur est exposé devient ultrapersonnalisée, voire totalement privée, chacun interagissant avec un système d'IA générative selon des modalités qui lui sont propres, hors de toute exposition partagée à un même ensemble de contenus. **Cette individualisation radicale de l'accès à l'information interroge les fondements mêmes d'un espace public conçu, historiquement, comme une agora de débat critique commun**[240].

[238] SMYRNAIOS, Nikos et KOCH, Olivier. « Audience générée par ChatGPT : « Le Monde » écrase la concurrence ». 13 mai 2026.
[239] DING, Yifan, FACCIANI, Matthew, JOYCE, Ellen, POUDEL, Amrit, BHATTACHARYA, Sanmitra, VEERAMANI, Balaji, AGUINAGA, Sal et WENINGER, Tim. **»** *Citations and Trust in LLM Generated Responses »*. Dans : *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence.* 2025. p. 2652.
[240] États généraux de l'information. Protégéer et développer le droit à l'information : une urgence démocratique. 12 septembre 2024. *Op.cit.*

b) Les potentialités de l'IA générative pour une information fiable et de qualité

L'IA générative offre aux rédactions des capacités nouvelles d'aide à l'investigation, de traitement de grands volumes de données, de traduction ou encore de vérification des faits, susceptibles de renforcer la production journalistique. Diverses applications d'IA générative en support d'une information fiable et de qualité (liste non exhaustive) peuvent être citées :

- **Leviers d'optimisation des coûts :** l'IA générative peut être mise au service de la productivité dans le but de réduire les coûts. Selon l'étude réalisée par le ministère de la Culture et l'Arcom[241], l'IA permet *« d'automatiser certaines étapes à faible valeur ajoutée »*, notamment dans l'assistance à la production de contenu d'information (où l'IA est déjà utilisée par 80 % des médias interrogés), la correction, la mise en page ou la maquette. *« L'IA permet de faciliter la multidiffusion de contenus »*, c'est-à-dire sur plusieurs supports. Ces gains de productivité ne seraient qu'à leurs débuts : *« près de trois quarts des médias anticipent des gains de productivité liés à l'IA à l'horizon trois ans »*, et pourraient être de l'ordre de plus de 10 % selon certains médias.
- **Dépolarisation des échanges :** doté d'un assistant qui lui suggère en temps réel de reformuler son propos plus respectueusement, de légitimer le désaccord ou encore d'adoucir le ton, un participant à une discussion politique en ligne en améliore la qualité perçue et la réciprocité démocratique, son interlocuteur se sentant davantage entendu, sans que les positions de fond se déplacent[242]. Le mécanisme agit sur la forme de l'échange, non sur son contenu.
- **Réduction de croyances complotistes :** un dialogue avec une IA opposant des contre-arguments personnalisés et étayés réduit l'adhésion des utilisateurs à des thèses complotistes d'environ 20 %, un effet persistant à deux mois[243].
- **Des outils au service de la fiabilité des sources :** l'outil SIDE repère les références fragiles de Wikipédia et en propose de meilleures, préférées par les évaluateurs humains dans 70 % des cas douteux[244].
- **Médiation délibérative :** une IA médiatrice produit des synthèses de délibération jugées plus claires et plus justes que celles de médiateurs humains[245], résultat corroboré sur le terrain de la médiation[246].

---

[241] PMP Strategy. Étude sur le modèle économique de l'information en France. Rapport pour l'Arcom et la DGMIC, Ministère de la Culture. 2025.
[242] ARGYLE Lisa, BAIL Christopher, BUSBY Ethan et al. « Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale ». *Proceedings of the National Academy of Sciences.* 2023. vol. 120, no 41, p. e2311627120.
[243] COSTELLO Thomas, PENNYCOOK Gordon, RAND David. « Durably reducing conspiracy beliefs through dialogues with AI **»**. *Science*. 2024. vol. 385, no 6714, p. eadq1814.
[244] PETRONI Fabio, BROSCHEIT Samuel, PIKTUS Aleksandra, et al. « Improving Wikipedia verifiability with AI ». *Nature Machine Intelligence.* 2023. vol. 5, no 10, p. 1142-1148.
[245] TESSLER Michael Henry, BAKKER Michiel, JARRETT Daniel, et al. « AI can help humans find common ground in democratic deliberation ». *Science*. 2024. vol. 386, no 6719, p. eadq2852.
[246] ZELIZER Craig, OGENGA Fredrick, SCHIRCH Lisa, et al. « Artificial Intelligence and Peacebuilding: Opportunities and Challenges ». *International Panel on the Information Environment*. 2025.

- **Accessibilité linguistique :** en Inde, un système d'IA générative a traduit l'information électorale dans plus de vingt langues, atteignant des électorats mal desservis par les médias traditionnels[247].

c) Vers la profusion de contenus informationnels

En tant que nouvel outil de production de contenus, y compris de contenus informationnels, les systèmes d'IA générative baissent considérablement les barrières à l'entrée pour la création de contenus artificiels, permettant à quiconque de produire, à moindre coût et à grande échelle, des contenus dont l'apparence peut rivaliser avec celle de médias établis. Cette facilité de production nourrit une saturation informationnelle qui rend plus difficile, pour le citoyen, l'identification d'informations et de sources fiables au sein d'un flux de contenus toujours plus dense et indifférencié. À un niveau plus collectif, les États généraux de l'information relevaient déjà en 2024[248], le fait que l'IA générative *« accroît le poids des algorithmes »*, accélère *« la prolifération des messages [et] la confusion générale »*, brouillant la frontière entre le vrai et le faux.

Face à ce constat, **certains acteurs estiment nécessaire de développer des labels ou des standards** permettant de distinguer les contenus produits par des médias reconnus des autres productions, à l'image de la *Journalism Trust Initiative* (JTI). Cette méthode vise à certifier la gouvernance, la transparence, l'indépendance éditoriale et les méthodes de production[249] d'une information. Si ce dispositif permet en principe de garantir un processus de production d'information fiable, un tel label pourrait aussi invisibiliser d'autres contenus informationnels de qualité dépourvus de marquage et ainsi constituer un frein pour que d'autres acteurs, disposant de moins de ressources pour effectuer de telles autoévaluations et audits, puissent entrer sur le marché informationnel. Se pose également la question de la découvrabilité des contenus d'un média certifié qui dépend de sa prise en compte par les plateformes et les agents conversationnels.

2) L'IA générative déstabilise l'espace informationnel

a) L'attrition de l'économie des médias traditionnels

Dès lors que l'IA générative se pose en nouvel intermédiaire dans l'accès à l'information, les enjeux de découvrabilité deviennent centraux pour la visibilité et le financement des médias traditionnels.

Aux États-Unis, ce changement de paradigme dans l'accès à l'information impacte déjà fortement l'économie des médias traditionnels qui voient leurs fréquentations chuter. **Moins de 0,5 % des 350 milliards de dollars investis dans l'IA en 2025 sert à rémunérer les médias traditionnels** ; les modèles d'IA ont en moyenne fait chuter le trafic des 100 plus grands journaux américains de 45 % en 4 ans

[247] PBS News. « India's latest election embraced AI technology. Here are some ways it was used constructively ». 12 juin 2024.
[248] États généraux de l'information. « Protéger et développer le droit à l'information : une urgence démocratique ». 12 septembre 2024. p.6.
[249] Reporter Sans Frontières. Journal Trust Initiative - Qu'est-ce que la JTI ?.

et un journal américain ferme tous les 3 jours[250]. Une telle tendance pourrait s'observer prochainement en France, alors que l'étude de l'Arcom sur le marché publicitaire à l'horizon 2025-2030 estimait déjà en janvier 2024 qu'en 2030, les deux tiers des recettes publicitaires seront dirigés vers les acteurs numériques (plateformes comprises), accélérant la perte de revenus des médias traditionnels.

Cette mise en péril du modèle économique des médias traditionnels s'explique notamment par la baisse de trafic sur le site de ces derniers et notamment le phénomène de la *« recherche zéro clic »*. Si celui-ci préexistait à l'IA générative, il s'en trouve très largement accentué. La recherche dite zéro clic est un mode de recherche d'information où l'utilisateur obtient la substance de l'information sans jamais visiter la source qui l'a produite. Les écarts entre le volume de contenus indexés par les *bots* d'IA et le trafic effectivement renvoyé vers les sites sources illustrent l'ampleur du déséquilibre : sur Google, un ratio moyen d'une seule visite renvoyée pour 30 consultations ; chez Anthropic, en moyenne pour 60 000 consultations par le *bot*, un renvoi vers le site source. Selon l'étude américaine du Pew Research Center[251], le système d'IA Claude d'Anthropic a effectué près de 71 000 requêtes de pages HTML pour chaque renvoi de page HTML, tandis que Mistral a envoyé dix fois plus de références que de requêtes d'indexation. Bien que les méthodes de mesure varient selon les acteurs, les résultats convergent sur l'ampleur du phénomène : la baisse du taux de clic en présence d'un résumé généré par IA serait comprise entre 35 % et 60 %. Les petits éditeurs ont perdu environ 60 % de leur trafic de recherche en deux ans, et l'ensemble des assistants d'IA renvoie moins de 1 % des visites vers les sites d'éditeurs[252]. Si ces données concernent avant tout les États-Unis, leur généralisation à l'ensemble de la profession n'est pas seulement de l'ordre du prospectif. À plus long terme, les éditeurs anticipent une baisse déclarative de 43 % de leur trafic de recherche à trois ans[253].

Chine Labbé souligne à ce propos que *« les accords, seuls les gros acteurs peuvent les signer par définition. Donc, nous allons vers un monde où il y aura un ou deux survivants qui sont les plus gros médias ; un monde pas très souhaitable en termes de pluralisme de l'information »* (audition du 16 juin 2026).

L'Autorité de la concurrence réclamait, dans son avis de juillet sur l'IA[254], de la transparence sur *« les conditions de visibilité »* au sein des réponses des assistants d'IA, soulignant l'importance de leur *« caractère non discriminatoire »* ou de l'affichage d'une pluralité de sources. En l'état, ces informations sont invisibilisées, rendant le tout opaque et incompréhensible pour les utilisateurs et les pouvoirs publics.

Les « AI Overviews » qui sont intégrés depuis cette année par défaut aux recherches Google illustrent parfaitement ce changement de paradigme dans l'accès à l'information. Ce système d'IA générative synthétise les résultats dans une réponse formulée en langage naturel. Ce résumé est systématiquement proposé de façon prioritaire par rapport aux résultats du moteur de recherche

[250] Le Grand Continent. « Comment sauver le journalisme au temps de l'IA ? ». 1er juin 2026.
[251] Pew Research Center, Ahrefs et Seer Interactive. 2025. Une méthodologie que Google conteste.
[252] Chartbeat. 2026.
[253] Reuters Institute. 2026. chiffre déclaratif.
[254] Autorité de la concurrence. « Avis 26-A-05 relatif au fonctionnement concurrentiel du secteur des agents d'intelligence artificielle ». 17 juillet 2026.

classique. Aux États-Unis, les éditeurs de sites se plaignent de baisses de trafic, de l'ordre de 5 % à 30 % selon les secteurs[255], et près de sept recherches Google sur dix aboutissent sans que l'internaute clique sur un lien[256]. En Allemagne, les résumés IA sur Google Search auraient engendré une chute de 6,6 % du nombre de clics en provenance du moteur de recherche[257].

En France, le classement ACPM de septembre 2026[258] témoigne d'un net recul des marques d'information sur un an. Les sites qui dépendent de la recherche perdent près de 23 % de visites et les applications sont moins touchées avec une baisse de 9 % de fréquentation. Ces chiffres illustrent les premiers effets de Google AI Overviews, bien qu'ils puissent également s'expliquer par des changements d'algorithme de Google Discover. Tandis qu'en 2025, un panel de marques d'actualité et d'information générale, nationales et régionales totalisait 2,14 milliards de visites, ce chiffre descend à 1,70 milliard un an plus tard, soit un recul de 20,8 %[259]. **L'Arcom évalue quant à elle une perte de trafic entre 33 et 38 % attribuable aux résumés générés par IA**[260].

En réaction, l'Alliance de la presse d'information générale (APIG) a saisi, début août 2026, l'Autorité de la concurrence[261], à la suite du lancement d'AI Overview et d'AI Mode en France, pour que Google respecte ses engagements de négociation de bonne foi, de transparence et de non-discrimination pris en 2022 et rémunère les contenus utilisés par ces aperçus IA au titre des droits voisins des éditeurs de presse.

Comme le résume le chercheur Grégoire Darcy, *« en captant la valeur informationnelle sans en financer la production, l'IA organise un parasitisme de l'amont : elle se nourrit d'un journalisme qu'elle contribue à rendre infinançable »* (audition du 15 juin 2026). À terme, **le risque est celui d'un *web* consulté par les systèmes d'IA générative et alimenté par ces mêmes systèmes, aboutissant à un « *web* synthétique »**, pour reprendre l'expression du chercheur en sciences de l'information et de la communication Olivier Ertzscheid[262], avec en somme un risque de perte en qualité des résultats générés s'ils ne sont entraînés que sur des contenus eux-mêmes synthétiques.

Au-delà des risques pour les médias traditionnels, ces résumés de l'information par l'IA générative peuvent également avoir des conséquences pour les utilisateurs. La neuroscientifique Anne-Laure Le Cunff montre[263] que l'obtention en un clic d'une réponse produite par IA aurait des conséquences néfastes, comme l'affaiblissement de la curiosité, piquée justement par les errances, détours, frictions et digressions rencontrés entre le moment où l'on se pose une question et celui où l'on y

[255] Le Monde. « Comment les sites tentent de s'adapter à la recherche par IA, susceptible de « changer l'économie du Web » ». 23 juillet 2026.
[256] Search Engine Land. « Google zero-click searches hit 68 % in early 2026: Study ». 9 juin 2026.
[257] Sistrix. « AI Overviews in Germany: How Much Click-Through Rates Are Really Dropping ». s.d.
[258] ACPM. Classement des marques numériques septembre 2026.
[259] The Media Leader. « AI Overview : les sites d'info perdent 23 % de visites, les applis résistent, selon ACPM ». 8 octobre 2026.
[260] Alliance de la presse d'information générale. « Services d'IA de Google : l'Alliance saisit l'Autorité de la concurrence ». 11 août 2026.
[261] Alliance de la presse d'information générale. « Services d'IA de Google : l'Alliance saisit l'Autorité de la concurrence ». 11 août 2026.
[262] ERTZSCHEID, Olivier. *Les IA à l'assaut du cyberespace. Vers un web synthétique*. C&F Éditions. 2024.
[263] The New York Times. « *We are Losing the Ability to Discover What We Didn't Know To ask* ». 8 juillet 2026.

répond. Cette conséquence est d'autant plus nocive pour les jeunes utilisateurs qui ont recours aux systèmes d'IA plus largement, d'après une étude de Common Sense Media[264].

b) L'IA générative au service de la manipulation de l'information à grande échelle

Marc Faddoul, cofondateur d'AI Forensics[265], distingue cinq couches algorithmiques qui influencent la manière dont un système d'IA générative sélectionne, hiérarchise et restitue l'information :

- **La curation des données d'entraînement** détermine le corpus sur lequel le modèle a été formé. Un modèle entraîné sur Grokipedia (une encyclopédie se présentant comme une alternative « anti-woke » à Wikipédia) ne produira pas nécessairement les mêmes réponses qu'un modèle entraîné sur un corpus plus diversifié.
- **L'alignement par renforcement à partir de retours humains ou artificiels** (RLHF et RLAIF) définit les réponses que le système doit privilégier.
- **La recherche *web* et la génération augmentée par récupération** (*RAG*) déterminent les sources externes effectivement mobilisées pour construire une réponse.
- **Les instructions système** (*system prompts*) modifient le comportement du modèle sans réentraînement.
- **Les filtres de sécurité**, appliqués en amont et en aval de la génération, encadrent les requêtes admises et les contenus produits. Les exemples récents ne manquent pas : Apple ajustant ses instructions pour ne plus qualifier les soutiens MAGA de « radicaux », Meta assouplissant ses garde-fous ou Grok amplifiant les allégations de « génocide blanc »[266] en Afrique du Sud.

(i) Fabriquer le faux : la génération industrielle de fausses informations

L'IA générative permet de produire des contenus trompeurs, à très large échelle, très rapidement et à moindre coût.

La multiplication inédite de faux sites d'information fondés sur l'IA générative et les contenus synthétiques l'illustre. Cette prolifération *« peut rendre plus difficile la vérification des sources et l'évaluation de la qualité des informations en ligne »*[267]. L'effondrement des coûts rend possibles des essaims d'agents coordonnés[268] et des fermes de sites automatisés saturant les moteurs de recherche[269], dont la cible n'est pas le lecteur humain mais l'infrastructure informationnelle elle-

[264] Common Sense Media. « Google's AI Search is failing kids — and parents can't turn it off ». 20 août 2026.
[265] FADDOUL, Marc. « Whos is Whispering in Your Chatbot's Ear? ». *Project Syndicate*. 17 mars 2026.
[266] ASSOCIATED PRESS (AP). « Elon Musk's AI Company Says Grok Chatbot Focus on South Africa's Racial Politics Was 'Unauthorized' ». *AP News*. 17 mai 2025.
[267] ARCEP. « IA générative : des défis pour l'avenir de l'internet ouvert ». Janvier 2026.
[268] SCHROEDER Daniel Thilo, CHA Meeyoung, BARONCHELLI Andrea, et al. « How malicious AI swarms can threaten democracy ». *Science*. 2026. vol. 391, no 6783, p. 354-357.
[269] VIGINUM. 2025.

même. En effet, **le nombre de faux sites d'actualités locales serait supérieur à celui des sites authentiques**, car l'IA générative (et le nouveau paradigme d'accès à l'information que ces outils engendrent, comme décrit dans la section précédente) rend la survie des sites d'information authentiques plus difficile et facilite sensiblement la création à moindre coût de sites factices[270]. Nous sommes confrontés à un paradoxe, selon le Pew Research Center[271]. D'une part, deux tiers des Américains se disent très préoccupés par la diffusion d'informations inexactes par l'IA. D'autre part, un pourcentage croissant[272] d'utilisateurs s'appuie néanmoins sur les systèmes d'IA pour s'informer, obtenir des renseignements et des conseils, et certains la considèrent même comme plus fiable[273] que les organes officiels de presse dont elle dépend pour ses réponses.

Dans leur phase d'entraînement, **les LLM peuvent être la cible de *data poisoning*, une attaque *via* laquelle les données d'entraînement sont contaminées par de fausses informations ou des données très orientées, par exemple politiquement ou idéologiquement**. Dans le cadre informationnel, deux modalités sont possibles : (i) un ciblage des sources qui font autorité et fortement pondérées dans l'entraînement des grands modèles, à la façon des « edit wars » sur Wikipédia où chaque correction se propage rapidement puisque la plateforme est largement utilisée par les modèles d'IA pour leur entraînement ; (ii) un ciblage par le volume : de nombreux faux sites sont créés, de faible autorité, reflétant les mêmes narratifs jusqu'à produire un effet cumulé sur les données d'entraînement issues du *crawling* sur le *web*. Cette technique malveillante constitue une vulnérabilité durable dans la mesure où les récits trompeurs seront ancrés dans les connaissances mêmes du modèle qui les restituera ensuite. La dépendance des modèles aux vastes corpus du *web*, issus des réseaux sociaux et autres plateformes renforce cette menace. Par un empoisonnement ciblé, une désinformation marginale peut se muer en contenu consensuel et bouleverser l'espace informationnel. L'Agence de l'Union européenne pour la cybersécurité (ENISA)[274] identifiait dès 2024 le *data poisonning* comme menace émergente à l'ère de l'IA générative.

L'Alan Turing Institute[275] examine le rôle que jouent les menaces liées à l'IA pendant et après les crises, ainsi que leur impact potentiel sur la stabilité démocratique. Celles-ci sont de plus en plus « *weaponisées* » par des acteurs malveillants et vont bien au-delà de la simple génération de contenus type *deepfake*. Notamment, **les acteurs malveillants cherchent à manipuler la manière dont les agents conversationnels transmettent des informations personnalisées aux utilisateurs en temps réel**. Parmi ces initiatives figuraient notamment les efforts déployés par des groupes liés au Kremlin pour « empoisonner » les bases de données d'entraînement des *chatbots* dans le cadre de la guerre menée par la Russie contre l'Ukraine.

Par son fonctionnement, **l'IA générative permet un engagement très élevé de ces fausses informations et de les rendre persuasives, en ajoutant une charge émotionnelle, ce qui favorise**

[270] Le Grand Continent. « Comment sauver le journalisme au temps de l'IA ? ». 1er juin 2026.
[271] Pew Research Center. « Views of risks, opportunities and regulation of AI ». 3 avril 2025.
[272] The Hill. « More people turning to AI bots for news: Poll ». 17 mai 2025.
[273] Nieman Lab. « People who use chatbots for news consider them unbiased and "good enough," new study finds ». 22 janvier 2026.
[274] ENISA. ENISA Threat Landscape 2024. Septembre 2024.
[275] Center for Emerging Technolgy and Security. « Adding Fuel to the Fire: AI Information Threats and Crisis Events ». 11 février 2026.

l'adhésion qu'elles suscitent, voire à leur reprise par des médias traditionnels lorsqu'elles sont diffusées à très large échelle. En outre, la persuasivité des contenus générés, qui augmente avec la qualité des différents modèles et versions des modèles, favorise l'adhésion à une vision, une idéologie, une opinion et sera d'autant plus susceptible de rencontrer un public et de bouleverser l'opinion publique.

Comme le souligne Kamel El Hilali, docteur en droit et spécialiste de l'IA : *« En matière de manipulation de l'information, les systèmes d'IA générative peuvent se révéler redoutables. En effet, lorsqu'un individu se confie à l'IA, l'outil ne se borne pas à produire une réponse, il peut également analyser le discours de l'utilisateur (stylométrie psychologique, ou psycholinguistique computationnelle). Ainsi, à partir des interactions, l'IA peut déduire de nombreuses informations au sujet de l'utilisateur : ses traits de personnalité, son état de santé mentale, sa situation économique et professionnelle, son orientation politique, sexuelle, religieuse, etc. L'IA peut ainsi identifier les structures et vulnérabilités individuelles, et comprendre comment influencer l'utilisateur dans le but, notamment, de le maintenir en ligne »* (audition du 2 juillet 2026).

En outre, la manière dont l'IA générative structure l'information de façon n'est pas neutre : les *chatbots* synthétisent les informations qui semblent faire autorité, ce qui questionne un possible passage de rôle de « curateur » à celui d'« éditeur ».

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement. Des mécanismes de raisonnement explicite type *Chain-of-Thought* (CoT) et l'intégration systématique d'informations vérifiées dans la génération de réponse, avec des méthodes de type *Retrieval-Augmented Generation* (RAG), permettraient de davantage questionner les réponses des *chatbots* et ainsi de diminuer l'engagement.

Il est toutefois important de souligner que **l'intégration des systèmes d'IA générative dans notre infrastructure de communication globale présente également des opportunités**. Parmi les applications bénéfiques de l'IA générative dans ce contexte figurent la détection et la suppression de contenus préjudiciables ou illégaux avant qu'ils ne se propagent de manière virale à la suite d'une crise. Associés à des mécanismes de modération et à une validation humaine, ces systèmes peuvent contribuer à limiter l'amplification de la désinformation, à identifier les schémas de propagation coordonnée et à prioriser les contenus nécessitant un examen urgent.

Les systèmes d'IA générative peuvent également soutenir la diffusion d'informations vérifiées grâce à des agents conversationnels fondés sur l'architecture RAG (*Retrieval-Augmented Generation*)[276]. En s'appuyant sur des sources officielles, vérifiées et régulièrement mises à jour, ces outils sont en mesure de produire des réponses contextualisées, d'en indiquer les sources et de limiter le risque

[276] DAMO, Greta, PETIOT, Stéphane, CABRIO, Elena et VILLATA, Serena. *« PEACE 2.0: Grounded Explanations and Counter-Speech for Combating Hate Expressions »*. Dans : *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)*. IJCAI, 2026, pp. 8380-8384.

d'hallucinations.

(ii) Diffuser le faux : la propagation accélérée de la désinformation

La lutte contre la diffusion de fausses informations est devenue, en l'espace d'une décennie, une priorité d'action publique à l'échelle mondiale, au point d'être classée par le Forum économique mondial parmi les tout premiers risques globaux[277]. Le terme doit ici être entendu au sens large, englobant à la fois la désinformation (la diffusion délibérée et intentionnelle de fausses informations), la mésinformation (une information erronée diffusée sans intention de nuire) et la malinformation (une information vraie mais délibérément déformée ou détournée à des fins nuisibles)[278]. Cette préoccupation trouve un écho direct dans la perception des citoyens : la propagation de fausses informations constitue, selon les Français[279], le principal risque associé à l'IA générative (voir *supra* chapitre 2-A, 2-b).

Une étude menée par l'Union européenne de radio-télévision[280] a ainsi révélé que **les principaux assistants d'IA génèrent des réponses déformant considérablement l'actualité dans près de la moitié des cas**. Un tel niveau d'imprécision, lorsqu'il touche des outils de plus en plus utilisés comme point d'accès à l'information, pose une difficulté particulière : il ne s'agit plus seulement d'un risque de manipulation intentionnelle et ciblée, mais d'une dégradation plus diffuse et systémique de la fiabilité de l'information distribuée à grande échelle, y compris lorsqu'aucune intention malveillante n'est à l'origine des contenus produits ; les utilisateurs sont induits en erreur. Selon le chercheur Grégoire Darcy, *« la vraie crise n'est pas l'invasion du faux : c'est l'érosion de la confiance dans le vrai »* (audition du 15 juin 2026).

Au stade de l'inférence, qui consiste *« à utiliser un modèle d'intelligence artificielle entraîné pour faire des prédictions sur de nouvelles données »*[281], les LLM peuvent être piégés dans le but que l'outil reprenne le contenu trompeur publié sur les faux sites d'information, notamment par des individus qui introduiraient des biais dans les réponses des LLM en implantant sur les sites largement consultés par l'IA (comme Wikipédia ou Reddit) de fausses informations, des narratifs trompeurs ou renvoyant vers de faux sites d'information. Par exemple, des tests réalisés par l'ONG Institute for Strategic Dialogue[282] montrent que des médias d'État, comme ceux de la Russie et de la Chine, exercent une influence non négligeable sur les processus d'inférence des LLM.

Des acteurs malveillants peuvent également s'introduire *via* des portes dérobées (ou *backdoors*) qui peuvent être sources de biais systémiques au sein des LLM. Ce mécanisme permet de déclencher un comportement spécifique d'un LLM chaque fois qu'il doit répondre à une requête contenant les mots-clés déclencheurs ou sur un sujet précis défini par l'attaquant. Des réponses manipulées

[277] WORLD ECONOMIC FORUM. « The Global Risks Report 2025 ». 15 janvier 2025.
[278] WARDLE, Claire et DERAKHSHAN, Hossein. *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Strasbourg : Council of Europe, 2017.
[279] IPSOS. « Intelligence artificielle : quels sont les usages des Français ? ». Février 2025.
[280] European Broadcasting Union (EBU) – BBC. « News Integrity in AI Assistants An international PSM study ». Octobre 2025.
[281] IBM. « Qu'est-ce que l'inférence en IA ? ». 18 juin 2024.
[282] ISD. Investigation - Talking points: When chatbots surface Russian state media. 27 octobre 2025.

peuvent ainsi être retransmises dans certains contextes, la diffusion de la désinformation devenant alors sélective, rendant son identification difficile.

Tout ceci peut se doubler d'une hybridation entre les réseaux sociaux et les services IA générative, lorsque ceux-ci sont directement intégrés dans les fonctionnalités des réseaux sociaux. La diffusion massive sur une plateforme du contenu massivement créé par le système d'IA générative démultiplie la désinformation. Par exemple, le système d'IA générative Grok intégré à la plateforme X a créé des images du Président Emmanuel Macron, en compagnie de Bachar al-Assad, qui ont été largement diffusées sur la plateforme, alors qu'ils ne s'étaient jamais rencontrés[283]. Tout ceci s'inscrit dans un contexte de baisse, voire de retrait de toute modération humaine au profit de la modération communautaire dont l'efficacité peut être discutée.

Pour répondre à ces enjeux, l'une des voies consiste à soutenir le développement de partenariats ou accords de licence entre les éditeurs de presse et les entreprises de l'IA qui permettent de rééquilibrer les sources d'entraînement et d'inférence. Cela favoriserait un modèle économique vertueux pour les éditeurs de presse, au service d'une information fiable et de qualité, et participerait à une meilleure performance des LLM.

L'IA générative et la lutte contre les manipulations de l'information dans la réglementation

Deux cas de figures sont pris en compte dans la réglementation :

(i) L'IA générative qualifiée de très grand moteur de recherche ou intégrée à des plateformes en ligne

L'article 35 du RSN impose aux très grandes plateformes et aux très grands moteurs de recherche (VLOPs et VLOSEs) de prendre des mesures pour atténuer les risques liés à la manipulation de l'information sur leurs services.

Ces obligations visent aussi bien les services d'IA générative intégrés à des plateformes en ligne (ex. : Grok sur X), que les services d'IA générative en eux-mêmes dès lors qu'ils ont été désignés en tant que VLOPs par la Commission européenne.

En avril 2024, la Commission a par ailleurs publié des lignes directrices sur l'atténuation des risques systémiques pour les processus électoraux afin de préciser les modalités selon lesquelles les VLOPs sont tenues de s'acquitter de leurs obligations en période électorale. Ces lignes directrices prévoient des mesures spécifiques concernant la création de contenus d'IA générative, mais également la diffusion de tels contenus sur ces services. Elles devraient faire l'objet d'une révision d'ici la fin de l'année 2026 (voir Recommandation n° 14).

[283] France info. « Qu'est-ce que Grok, la nouvelle fonction de X qui permet de créer gratuitement des images à partir de l'intelligence artificielle ? ». 18 décembre 2024.

(ii) L'IA générative en tant que telle

Plusieurs dispositions du RIA et d'instruments les complétant encadrent ce cas :

L'article 5 du RIA prohibe plusieurs pratiques de manipulation, en matière de systèmes d'IA, en particulier en cas de pratiques subliminales ou délibérément trompeuses ou en cas d'exploitation des vulnérabilités liées à l'âge ou au handicap. Ces interdictions supposent la caractérisation de plusieurs conditions.

L'article 50 (2) du RIA pose une obligation de marquage des contenus générés par IA, dans un format lisible par machine et identifiables comme ayant été générés ou manipulés par IA, à la charge des fournisseurs.

L'article 50 (4) du RIA pose une obligation de transparence à la charge des déployeurs d'un système d'IA qui génère ou manipule des contenus diffusés au public, d'une part pour les images et contenus audio et vidéo constituant un hypertrucage et d'autre part, pour les contenus de nature textuelle publiés dans le but d'informer le public sur des questions d'intérêt public. Cette obligation ne s'applique pas lorsque le texte a fait l'objet d'un processus de relecture humaine ou de contrôle éditorial et qu'une personne physique ou morale en assume la responsabilité éditoriale.

Le Code de bonnes pratiques sur la transparence des contenus générés par l'IA est paru en juillet 2026. Ce code vise à aider les fournisseurs et les déployeurs de systèmes d'IA générative à se conformer aux obligations de l'article 50. Le code insiste sur la complémentarité des techniques telles que le marquage par les méta-données et le tatouage numérique. OpenAI et Meta y ont adhéré volontairement.

Les lignes directrices de l'article 50, publiées en juillet, viennent préciser la mise en œuvre de l'article 50. Notamment, concernant les systèmes générant ou manipulant du contenu synthétique, les fournisseurs devront intégrer des marqueurs lisibles par machine (ex. : filigranes) et permettre leur détection.

Pour que le marquage soit lisible par la machine, les fournisseurs doivent intégrer des filigranes, métadonnées ou marqueurs cryptographiques. Les outils de détection doivent être publics (idéalement *via* des normes industrielles). Le marquage doit être efficace, interopérable, robuste et fiable.

L'article 55 du RIA sur les modèles d'IA à usage général et à risque systémique : si un système d'IA générative intègre en son sein un modèle d'IA à usage général et à risque systémique au sens de l'article 3.65 et de l'article 51 du RIA, il convient de relever que le fournisseur du modèle est tenu par l'article 55 de veiller à l'identification et à l'atténuation des risques systémiques

de son modèle. Au rang de ces risques systémiques, il faut citer le risque de manipulation de masse[284].

En tout état de cause, face à un événement de crise majeur impliquant une menace informationnelle générée ou amplifiée par IA, l'État doit disposer d'un protocole de détection et de réponse clair, activable rapidement, interministériel, reposant sur des indicateurs de gravité prédéfinis et sur des canaux d'échange établis à l'avance avec les entreprises d'IA, plutôt que d'improviser une coordination dans l'urgence.

Recommandation n° 7 : Doter l'État d'un protocole national de réponse aux crises informationnelles liées à l'IA générative, en définissant des indicateurs de gravité et des canaux d'échange rapide de données avec les fournisseurs d'IA générative. Ce protocole vise à ce que l'État puisse réagir promptement, de manière organisée et coordonnée. Une telle gestion nécessite une coopération avec les fournisseurs d'IA sur la base de canaux d'échange privilégiés afin de mettre un terme à une telle menace. Ce dispositif doit résolument être interministériel au regard de la diversité des champs concernés.

À cette fin, il est nécessaire que l'État dispose d'une capacité systémique de détection précoce et d'un dispositif partagé de veille permettant d'identifier les campagnes de manipulation avant qu'elles ne prennent une ampleur virale.

Recommandation n° 14 : Détecter de manière précoce les manipulations de l'information par les systèmes d'IA générative. L'État doit se doter d'une capacité de détection précoce et systémique, afin de repérer les campagnes avant qu'elles ne deviennent virales. Il est également nécessaire de construire une capacité d'audit « system-wide » et de créer un dispositif de veille partagée associant les acteurs en la matière.

[284] La notion de manipulation de masse ne figure pas dans le RIA mais dans l'appendice 1.4 du chapitre « Sûreté et sécurité » du Code de bonnes pratiques pour les modèles d'IA à usage général publié le 10 juillet 2025.

3) L'IA générative met au défi le pluralisme et la liberté d'expression

Le fonctionnement de l'IA générative modifie en profondeur les rapports de force dans l'accès au savoir. En externalisant une part croissante de nos tâches cognitives (recherche d'information, synthèse, rédaction, soutien émotionnel, *etc.*) vers des systèmes centralisés, elle transforme la manière dont les connaissances sont produites, sélectionnées et rendues disponibles. La cognition devient ainsi une ressource externalisée, standardisée et monétisée. Le sociologue Dominique Boullier identifie des conséquences majeures à ce fonctionnement. Selon lui, **l'IA générative risque de produire une rupture fondamentale dans l'univers partagé des savoirs, ce qu'il va même jusqu'à qualifier de « colonialisme cognitif »**, défini comme l'imposition d'un cadre de connaissance présenté comme universel et neutre, alors qu'il reflète en réalité les choix linguistiques, culturels, économiques et politiques des acteurs qui conçoivent, entraînent et déploient ces systèmes. En conséquence, cela aurait pour effet de construire un « commun informationnel » défini par une poignée d'acteurs, capables de déterminer la réponse considérée comme officielle, au détriment de la diversité et des contextes culturels (audition du 17 avril 2026).

Cette dynamique engendre des interrogations autour de notre rapport à la vérité. Les systèmes d'IA générative, majoritairement utilisés par les jeunes générations, tendent à véhiculer une voie unifiée, au travers de réponses standardisées[285], ce qui amenuise la diversité culturelle à laquelle ils sont exposés.

L'hyperpersonnalisation des réponses fournies par l'IA générative tend à enfermer les utilisateurs dans leurs propres schémas de pensée et renforce leurs biais cognitifs plutôt que de les corriger. Cet enfermement s'explique en grande partie par la capacité de persuasion propre à ces outils, sensiblement supérieure à celle des systèmes d'accès à l'information traditionnels. En effet, **une conversation avec une IA générative serait ainsi environ 50 % plus persuasive que la lecture d'un message statique équivalent**[286] et un dialogue personnalisé avec un système d'IA peut, dans certaines conditions, se révéler plus efficace que des débatteurs humains pour faire évoluer une opinion[287]. Il importe de souligner que cette capacité de persuasion accrue tient au caractère conversationnel et adaptatif du dialogue, à la capacité du système à ajuster ses arguments en fonction des réactions de l'interlocuteur, et non à la simple rapidité de production des contenus.

Ce phénomène produit toutefois des effets contrastés selon les contextes. D'une part, il tend à consolider des croyances préexistantes de l'utilisateur plutôt qu'à l'aider à en former de nouvelles. Comme souligné par le *Washington Post*[288] *« l'IA générative semble moins contribuer à façonner la façon dont les gens votent et bien plus à éroder leur foi dans la réalité »*. Ainsi, les images et vidéos produites par IA sont davantage utilisées pour consolider les croyances partisanes que pour

[285] UNESCO. « Guidance for generative AI in education and research ». 2023.
[286] HACKENBURG Kobi, TAPPIN Ben, HEWITT Luke, et al. « The levers of political persuasion with conversational artificial intelligence ». *Science*. 2025. vol. 390, no 6777, p. eaea3884.
[287] SALVI, Francesco, HORTA RIBEIRO, Manoel, GALLOTTI, Riccardo et WEST, Robert. « On the Conversational Persuasiveness of GPT-4 ». *Nature Human Behaviour*. 2025. vol. 9, no 8, pp. 1645-1653.
[288] Washington Post. « AI didn't sway the election, but it deepened the partisan divide ». 9 novembre 2024.

influencer les mentalités et les utilisateurs ont plus tendance à croire qu'un contenu authentique est produit par l'IA que l'inverse. D'autre part, et de manière plus inattendue, cette même capacité conversationnelle peut aussi produire un véritable changement de croyances, y compris dans un sens correctif. Une étude[289] portant sur 1,67 million de demandes de vérification adressées aux assistants conversationnels Grok et Perplexity sur la plateforme X montre ainsi que les utilisateurs modifient significativement leurs croyances à la suite de ces échanges, avec des tailles d'effet comparables à celles obtenues par des *fact-checkers* professionnels humains. Ce résultat suggère que **la capacité de persuasion de l'IA générative n'est pas univoque : si elle peut renforcer l'enfermement cognitif dans certains contextes, elle peut également constituer, dans d'autres, un levier de correction informationnelle dont l'ampleur reste comparable aux dispositifs de vérification traditionnels**.

*

C) L'IA générative contribue à l'érosion de la confiance entre l'État et les citoyens

1) L'IA générative aggrave plus qu'elle ne résout l'exclusion numérique

Les réalités de l'IA générative sont marquées par une contradiction. D'un côté, les discours d'« urgence » tendent à se multiplier, qu'ils concernent la nécessité d'adopter cette technologie sous peine de « manquer le virage », en misant sur des laboratoires de pointe avec des investissements massifs. De l'autre, force est de constater que l'adoption et l'appropriation de ces outils restent inégalement réparties au sein de la population.

13 millions de Français sont encore en situation d'illectronisme, c'est-à-dire *« ne possédant pas les compétences numériques de base (rechercher des informations en ligne, communiquer en ligne, utiliser des logiciels, protéger sa vie privée, résoudre des problèmes en ligne) ou ne se servant pas d'Internet (incapacité ou impossibilité matérielle de l'utiliser dans les trois derniers mois) »*[290]. Cette difficulté d'accès et d'usage freins concerne notamment les catégories de la population faiblement diplômées. En contrepartie, **près de six Français sur dix ont déjà joué un rôle d'aidant numérique**.

Ces besoins ne vont pas en diminuant, au contraire. Les risques et questionnements qui accompagnent les nouveaux usages numériques : violation de données personnelles, cyberattaques, fraude, altération de la santé mentale, ingérences étrangères et manipulation des scrutins locaux, *etc.* **ne pourront être résolus sans une solide montée en compétence de la population**.

Dans ce contexte, **l'IA générative est parfois mise en avant comme une opportunité de résoudre cet éloignement numérique**, par exemple en permettant d'accompagner les usagers dans un langage

[289] RENAULT, Thomas, MOSLEH, Mohsen et RAND, David. @Grok Is this true? LLM-powered fact-checking on social media. 2026.
[290] ANCT. « L'illectronisme en chiffres ». Consulté en octobre 2026.

naturel et de façon individualisée, de répondre aux questions fréquentes, y compris dans une langue étrangère pour les publics allophones ou encore avec un plus grand degré d'accessibilité numérique.

Cependant, à ce jour, **les usages de l'IA explosent mais les *« indicateurs historiques »*[291] de l'éloignement numérique se maintiennent**. Les données du Baromètre du numérique 2026 le confirment. Malgré une adoption en forte croissance (48 % des Français en 2026, soit + 28 points en deux ans), **son utilisation reste profondément inégale** selon l'âge, le niveau de diplôme et la catégorie socioprofessionnelle. Les jeunes de 18 à 24 ans l'utilisent massivement (85 %), contre seulement 15 % des plus de 70 ans, tandis que les diplômés du supérieur (65 % d'utilisateurs) sont bien plus à l'aise que les non-diplômés, dont 43 % déclarent ne pas avoir confiance dans ces outils[292].

Loin de résoudre les problèmes structurels, **l'IA ajoute une couche de complexité pour les publics déjà éloignés du numérique**. Elle exige des compétences accrues : savoir formuler des *prompts* précis et donc maîtriser le passage à l'écrit lorsque cela est nécessaire, adapter son langage aux attentes de l'algorithme ou encore évaluer la pertinence et la fiabilité des réponses générées. Sur ce dernier point, l'IA générative vient ajouter un nouveau facteur d'exclusion lié à la capacité à comprendre les mécanismes algorithmiques sous-jacents et les biais des modèles, à détecter les hallucinations ou encore à porter un regard critique sur les contenus proposés. Ces compétences reposent sur des **codes socioculturels et techniques marqués socialement**[293]. Ainsi, pour ces publics les plus précaires et vulnérables, **l'IA ne simplifie pas les usages : elle les complexifie**, avec un risque de **favoriser les publics déjà à l'aise avec le numérique, en laissant de côté ceux qui en sont éloignés**.

Pourtant, **l'autoformation par et à l'IA ne suffit pas** : les publics en difficulté recherchent avant tout une **réassurance humaine**, que ce soit pour surmonter la peur de commettre une erreur ou pour être guidés dans des démarches et usages complexes. Sans un accompagnement renforcé et une politique d'inclusion numérique, cette technologie pourrait creuser davantage les fractures sociales. **Dans ce contexte, la baisse annoncée des financements publics en la matière est un facteur d'inquiétude**[294].

Cette trajectoire pourrait **renforcer le sentiment d'éloignement de l'État chez les citoyens**, déjà confrontés à une dématérialisation croissante des démarches administratives.

[291] MACALUSO, Thomas. « IA : est-ce vraiment la fin de l'inclusion numérique ? ». *Les Bases du numérique d'intérêt général*. 03 septembre 2026.
[292] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[293] MACALUSO, Thomas. « IA : est-ce vraiment la fin de l'inclusion numérique ? ». *Les Bases du numérique d'intérêt général*. 03 septembre 2026.
[294] Acteurs Publics. « À bout de financements, l'inclusion numérique en quête de soutien ». 14 septembre 2026.

2) La dématérialisation des services publics est un facteur supplémentaire d'érosion de la confiance

a) L'IA dans les services publics : une dématérialisation porteuse d'exclusion et de défiance

Tandis que la numérisation des services publics est officiellement présentée comme un levier de simplification pour les usagers (*chatbots* d'orientation, formulaires intelligents, assistants virtuels), le besoin d'interaction humaine et l'opacité persistent pour l'ensemble des usagers, qui maitrisent ou non les outils, et encore davantage pour les 16 millions de personnes éloignées du numérique sur le territoire français[295]. Face à des situations complexes, le chemin automatisé ou la réponse générée peuvent être incomplets, approximatifs, voire erronés alors que les guichets physiques et les lignes téléphoniques tendent à être supprimés, sous couvert de motifs budgétaires. Ce décalage entre la promesse de fluidité et l'expérience réelle des usages qui se heurtent à des interfaces automatisées peu explicables alimente **un sentiment d'éloignement de la puissance publique et de déshumanisation de services essentiels**, facteur de défiance envers l'État et les collectivités.

L'usage de l'IA générative dans les services publics comporte également des risques d'altération de la qualité des décisions administratives et de renforcement des discriminations. Comme le montre un rapport de l'Ada Lovelace Institute[296], les outils d'IA, en particulier ceux dédiés à la transcription automatisée dans le champ des services sociaux, peuvent devenir des vecteurs de biais et de stéréotypes, avec des conséquences directes sur la vie des citoyens (accès aux droits, aux aides sociales, à l'emploi, *etc.*).

Comme l'indique Soizic Pénicaud, cofondatrice de l'Observatoire des algorithmes publics, l'IA générative est souvent perçue à tort comme un simple outil facilitant des tâches routinières. Pourtant, selon elle, *« l'IA générative cache la forêt de l'automatisation »* [297] : l'IA générative est désormais la focale de l'action publique, négligeant les autres systèmes automatisés, qui ont, eux aussi, des impacts conséquents sur la vie des citoyens et citoyennes. En d'autres termes, **son utilisation par les agents publics ne se limite pas à une optimisation des processus : elle redéfinit les contours de la décision et automatise, parfois insidieusement, des pans entiers du service public**. Or, sachant que ces outils peuvent reproduire des biais ou générer des erreurs, une utilisation importante de leurs résultats par les agents, sans contrôle humain formalisé et rigoureux, risque d'avoir des répercussions majeures sur l'équité et la qualité du service public.

Une étude[298] récente sur l'emploi de modèles d'IA pour résumer des dossiers de soins de longue durée dans une collectivité locale anglaise en offre une illustration frappante. Les résultats révèlent que certains modèles, comme Meta Llama 3 et Google Gemma, minimisent systématiquement les problèmes de santé des femmes par rapport à ceux des hommes. Par

[295] ANCT. « La société numérique française : définir et mesurer l'éloignement numérique ». 2023.
[296] Ada Lovelace Institute. « Scribe and Prejudice? Exploring the Use of AI Transcription Tools in Social Care ». 11 février 2026.
[297] PÉNICAUD, Soizic. « L'IA générative cache la forêt de l'automatisation ». *Gazette des communes, des départements, des régions*. no 2822. 22 juin 2026.
[298] RICKMAN, Sam. Evaluating Gender Bias in Large Language Models in Long-Term Care. BMC Medical Informatics and Decision Making. 2025. vol. 25, p. 274.

exemple, les résumés concernant des hommes incluaient bien plus fréquemment des descripteurs comme *« disabled »* (handicapé) ou *« complex »* (complexe), suggérant une sous-évaluation des besoins des femmes dans l'allocation des services. Or, dans un domaine où les ressources sont attribuées en fonction des besoins identifiés, de tels biais peuvent aggraver les inégalités de traitement entre les genres.

Une des conséquences structurelles de ces dynamiques est l'aggravation du non-recours aux droits. Ce phénomène est massif en France puisque la proportion des droits non sollicités oscille entre 30 et 50 % selon les prestations[299] et les profils (les personnes en situation d'illectronisme sont les plus concernées[300]). Cela s'explique par des causes multiples qui tiennent à la méconnaissance des droits, à la complexité administrative, au sentiment d'illégitimité mais aussi aux barrières numériques. L'introduction de l'IA générative dans les services publics sociaux risque d'amplifier ce phénomène : d'une part, en réduisant les points de contact humain qui permettaient de détecter et de corriger le non-recours aux droits ; d'autre part, en introduisant de nouvelles limites algorithmiques qui orientent ou écartent certains profils de manière opaque, sans que les usagers puissent comprendre pourquoi leur demande a été jugée « non prioritaire » ou « hors champ ».

b) La fragilisation du lien de confiance collective

De manière plus générale, **l'imprégnation de l'IA générative dans l'espace informationnel contribue à éroder la confiance du citoyen envers les institutions et les médias traditionnels. Notamment, 87 % des Européens estiment que la manipulation de l'information** (telle que la désinformation, l'ingérence étrangère, les contenus générés par IA, les *deepfakes*) **constitue une menace pour nos processus démocratiques**[301].

Au-delà de la réelle production de faux contenus par outils d'IA générative, l'existence même de l'IA générative est parfois instrumentalisée pour saper la confiance des utilisateurs dans des contenus pourtant authentiques. L'argument selon lequel une image, une vidéo ou un texte pourrait avoir été généré par IA est ainsi de plus en plus fréquemment invoqué pour discréditer des preuves ou des témoignages réels, sans que cette allégation soit elle-même étayée. Ce mécanisme, parfois désigné sous les termes de *«* dividende du menteur *»* (*liar's dividend*), élargit considérablement la zone de flou entre contenus vérifiés et contenus inauthentiques : il ne s'agit plus seulement de faire passer un contenu falsifié pour réel, mais également de faire douter d'un contenu réel en le faisant passer, à tort, pour falsifier. Cette réversibilité du soupçon fragilise d'autant plus le débat public qu'elle ne nécessite aucune sophistication technique : la simple évocation de la possibilité d'une manipulation par IA suffit à instiller le doute.

Ce phénomène s'inscrit par ailleurs dans un contexte de défiance déjà croissante envers les médias traditionnels. La multiplication des contenus falsifiés ou manipulés, conjuguée à la difficulté

[299] Observatoire des non-recours aux droits et services. 2024.
[300] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026.
[301] Commission européenne. « Digital decade Policy programm ». Juin 2026.

croissante de les distinguer des contenus authentiques, nourrit un scepticisme plus généralisé, qui ne se limite pas aux seuls contenus effectivement problématiques mais s'étend progressivement à l'ensemble de l'écosystème informationnel. **À terme, c'est la capacité même des citoyens à s'accorder sur un socle minimal de faits vérifiables qui se trouve ainsi fragilisée, condition pourtant essentielle à la tenue d'un débat démocratique**.

Dans le contexte de la manipulation de l'information, l'*astroturfing*[302] (manipulation de foules) désigne une forme particulièrement subtile de désinformation, où des messages sont relayés comme étant ceux d'individus alors qu'il s'agit des campagnes organisées par des lobbies au profit d'industries ou de partis politiques. **L'*astroturfing* numérique constitue une réelle menace pour la démocratie en ce qu'il créé artificiellement un consensus public *via* des comptes automatisés ou des contenus synthétiques**. Ces phénomènes, déjà connus avant l'IA générative, vont sans surprise passer à plus grande échelle, avec une création facilitée, rapide et crédible de fermes à *trolls* gigantesques.

Face à ces risques, l'État doit fournir un cadre sécurisant, tant pour les agents qui utilisent les outils d'IA que pour les citoyens qui les subissent. Des travaux d'Inria[303] ont montré que les administrations sont souvent dans l'impossibilité de produire des explications détaillées des décisions prises sur fondements algorithmiques. La stratégie d'usage de l'IA en gestion des ressources humaines dans la fonction publique d'État rappelle d'ailleurs que la transparence et le contrôle humain figurent parmi les principes fondamentaux encadrant l'usage de l'IA, y compris générative[304]. Cela pourrait être exacerbé par l'IA générative, mettant à mal le devoir de transparence et d'explicabilité auquel l'administration est soumise. Pour autant, peu d'administrations disposent aujourd'hui de capacités d'audit, de journalisation et de contre-expertise suffisantes pour vérifier, *a posteriori*, le rôle effectif de l'IA dans les décisions qui impactent les droits des usagers.

Dans ce contexte, plusieurs organismes publics ont commencé à prendre en compte les risques. La CNIL a ainsi publié des lignes directrices sur l'usage de l'IA dans le secteur public[305], en insistant sur la nécessité de préserver un « droit à l'humain », de documenter les systèmes et de garantir la possibilité de contester les décisions automatisées. La DINUM, dans ses recommandations sur l'administration, appelle à un usage « assistif » de l'IA générative et non « décisionnel » et à la mise en place de procédures de validation humaine pour les décisions individuelles. L'ANSSI, de son côté, alerte sur les risques de sécurité liés à l'intégration de modèles d'IA dans les systèmes d'information de l'État. Ces travaux constituent une base nécessaire, mais insuffisamment contraignants au regard des enjeux de cohésion sociale.

[302] France Inter. « L'astroturfing, la grande illusion de l'opinion ». 17 avril 2024.
[303] Inria. « De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels ». 2024.
[304] DGFAP. « Stratégie d'usage de l'intelligence artificielle en matière de gestion des ressources humaines dans la fonction publique de l'État ». Juin 2024.
[305] CNIL. « Bac à sable « IA et services publics » ». Mars 2025.

3) En période électorale, l'IA offre des outils pour exploiter les vulnérabilités démocratiques

a) Influence et délégation des choix électoraux

L'IA générative a déjà eu un impact non négligeable sur les choix électoraux des utilisateurs lors des dernières élections municipales françaises. Selon les données disponibles, **16 % des Français déclarent avoir déjà utilisé une IA générative pour éclairer leur choix électoral** à cette occasion[306]. Ce chiffre global recouvre des usages différenciés : 7 % des Français l'ont utilisée comme outil de confortation de leur choix initial, 5 % comme source d'influence et 4 % comme véritable aide à la décision. La répartition des réponses se structure comme suit : entre 40 % et 50 % des utilisateurs indiquent avoir été confortés dans leur choix initial, entre 30 % et 40 % déclarent avoir changé d'avis à l'issue de l'échange, et entre 20 % et 30 % considèrent que l'IA a constitué une réelle aide à la décision. Ces proportions, bien que fondées sur des données déclaratives, témoignent d'une intégration déjà significative de ces outils dans le processus de formation du choix électoral des citoyens.

En marge des élections municipales de 2026, plusieurs initiatives ont émergé. Par exemple, le Secrétariat général de la défense et de la sécurité nationale (SGDSN) et le ministère de l'Intérieur participent à la sécurisation des processus électoraux, notamment sur le volet cyber, *via* des actions de sensibilisation des parties prenantes, de gestion de crise ou encore de coordination et de partage d'informations entre agences de l'État. À ce titre, le service VIGINUM du SGDSN a produit un guide de sensibilisation à l'intention des équipes de campagne[307] préconisant des mesures pour se protéger contre la menace informationnelle. Aussi, l'Arcom a fourni aux plateformes en ligne des préconisations visant à lutter contre les risques spécifiques ou amplifiés en période électorale[308], notamment du fait de l'IA générative. Ces préconisations se concentrent autour de trois axes : les ressources que les plateformes doivent déployer dans le cadre de leur responsabilité de modération, la sensibilisation et l'éducation des publics aux problématiques de manipulation de l'information et la mise en œuvre du cadre juridique applicable.

b) Un risque d'homogénéisation de l'offre politique

Les systèmes d'IA générative peuvent être utilisés comme des outils par les candidats pendant les périodes électorales, au risque d'homogénéiser le discours politique. L'influence des agents conversationnels sur les électeurs s'accompagne d'un recours croissant à l'IA générative du côté des candidats. Une étude du CEVIPOF[309] montre que **les candidats aux élections municipales ont**

[306] Terra Nova, Harris Interactive. « IA et politique : vers un outil d'aide, voire d'influence sur la décision ? ». Mai 2026.
[307] SGDSN. Protéger le débat public numérique en contexte électoral. Guide de sensibilisation à l'attention des équipes de campagne. 2026.
[308] Arcom. Préconisations relatives à la lutte contre la manipulation de l'information sur les plateformes en ligne en vue des élections municipales des 15 et 22 mars 2026. Janvier 2026.
[309] MARIE, Antoine. « L'IA dans les élections municipales 2026 : Des usages qui appellent un regard nuancé ». *Cevipof*. Avril 2026.

massivement utilisé l'IA générative pour produire tracts, professions de foi, podcasts, visuels ou synthèses de réunions publiques. Si cette appropriation a permis une baisse drastique du coût de production des contenus de campagne électorale, elle s'est également accompagnée d'effets plus problématiques : une tendance à l'homogénéisation des discours politiques, les contenus produits par des outils similaires présentant des formulations et des angles comparables d'une campagne à l'autre, ainsi que plusieurs cas documentés de désinformation à l'échelle locale, notamment des *deepfakes* ou de fausses images relatives à la sécurité ou à la propreté urbaine, utilisées pour discréditer des candidats adverses ou des bilans municipaux. Ces erreurs et détournements, bien que localisés, illustrent la difficulté à encadrer un usage désormais généralisé de ces outils. Aussi, si les partis politiques utilisent largement l'IA générative pour produire des contenus légitimes, ceux-ci pourraient ressembler à des contenus illégitimes générés par des attaquants, au risque de ne plus pouvoir différencier l'information de la désinformation et se confondre avec eux.

Des mesures complémentaires aux dispositifs institutionnels peuvent être mises en place par les fournisseurs d'IA afin de garantir une utilisation de leurs systèmes plus protectrice en période électorale. C'est par exemple le cas d'OpenAI qui, à l'approche des élections de mi-mandat aux États-Unis (3 novembre 2026), a renforcé les conditions d'utilisation de ChatGPT en bloquant la génération d'images photoréalistes dépeignant, selon l'entreprise, *« une ingérence électorale ou susceptibles de décourager le vote »*[310]. Seules les illustrations de type « *cartoon* » sont permises afin de ne pas altérer la perception du réel. OpenAI a annoncé interdire également la génération de messages de campagne en faveur ou à l'encontre d'un candidat, d'un parti politique ou d'une proposition soumise au vote (texte, image, vidéo). Toutefois, *The New York Times*[311] fait état de contenus générés par IA, notamment en faveur des républicains pendant la campagne des élections de mi-mandat. Le Brésil[312] est allé jusqu'à interdire la diffusion de contenus générés par IA dans les 72 heures précédant l'élection présidentielle d'octobre 2026 et les 24 heures suivant celle-ci. Cette mesure a été imposée par le Tribunal supérieur électoral du Brésil, de manière complémentaire au marquage de tout contenu généré par IA dans le cadre de la campagne. Certains estiment néanmoins que malgré les craintes légitimes, la production de *deepfakes* dans le contexte politique est certes forte, mais elle est plus utilisée sous forme de satire ou de divertissement que comme outil de désinformation ou d'escroquerie[313].

[310] Next. « Elections : OpenAI encadre la génération d'images par ChatGPT, mais le flou persiste ». 28 octobre 2026.
[311] The New York Times. « $137 Million in A.I.-Generated Ads Flood the Midterms ». 16 octobre 2026.
[312] Le Monde. « Au Brésil, la désinformation et l'IA irriguent la compagne présidentielle ». 19 octobre 2026.
[313] Northwestern. « Tracking Political Deepfakes: New Database Aims to Inform, Inspire Policy Solutions ». 26 janvier 2024.

Un rapport du *think tank* transpartisan britannique Demos[314] a montré que les LLM fournissaient des informations erronées aux électeurs pour 34 % des questions posées. En effet, selon les tests effectués le 27 mars 2026 sur ChatGPT, Gemini, Google AI Overviews, Grok et Replika, plus d'un tiers (34,1 %) des réponses aux questions sur les élections écossaises contenaient des erreurs factuelles (109 réponses sur 320). Parmi les réponses inexactes, 8,75 % (28 sur 320) étaient totalement fausses et 25,3 % étaient partiellement exactes mais comportaient des erreurs (81 sur 320).

L'étude relève des biais politiques et cite l'exemple de Grokipedia, en grande partie généré par Grok, qui comprenait un biais en faveur des personnalités de droite[315]. Ces biais engendrent, selon le think tank, des risques de standardisation des réponses mettant en avant les voix culturelles dominantes et ne représentant pas de manière adéquate la diversité sociétale[316]. Ces risques sont plus importants sur les plus petits modèles, qui ont moins de garde-fous, à l'instar de Replika.

Enfin, les expériences ont montré que l'IA conversationnelle peut être très efficace pour modifier les opinions politiques des individus, surtout si les *chatbots* génèrent des réponses riches en informations et comportant de nombreuses citations. Or, plus un texte est persuasif, plus il est susceptible de contenir des erreurs et des affirmations erronées. Toutefois, les chercheurs précisent qu'un contenu qui influence un individu ne se traduit pas nécessairement par un changement d'intention de vote. Cet équilibre des risques pourrait néanmoins évoluer si les électeurs accordent une confiance croissante aux réponses des services d'IA.

Pour préserver la confiance dans les institutions démocratiques, et, en période électorale, garantir des élections libres et éclairées, il est impératif que **les partis politiques et candidats s'engagent publiquement à ne pas utiliser l'IA générative pour tromper, manipuler ou désinformer l'électorat**. À l'approche de l'élection présidentielle française de 2027, un code de bonne conduite permettrait d'encadrer l'usage de ces technologies et de sensibiliser tant les acteurs politiques que les citoyens.

Recommandation n° 15 : Instaurer un code national de bonne conduite sur l'usage de l'IA générative en période électorale. Les partis politiques doivent être associés à l'élaboration et à la signature d'un tel code. Il devra être complété d'un mécanisme de suivi et de signalement des manquements, dont le contrôle sera assuré par une autorité indépendante.

*

[314] Demos. « Electoral hallucinations – Safeguarding UK – Elections in the world of llms and ai chatbots ». Mai 2026.
[315] The Financial Times. « Elon Musk's Grokipedia is a major own goal ». 2025.
[316] JACKSON et al. « Large AI Models Have a Prioritization Problem: Policy Implications and Solutions ». Policy Insights from the Behavioral and Brain Sciences. 2025.

Ainsi, les vulnérabilités étudiées dessinent un risque collectif, celui d'un espace social, informationnel et démocratique fragilisé par l'IA générative. Or, ce qui sous-tend l'ensemble des vulnérabilités cartographiées est le cadre dans lequel s'inscrit l'IA générative : un modèle économique encore non défini et un cadre juridique insuffisamment contraignant. La stabilisation et le renforcement de ces deux entrants peuvent constituer un levier pour lutter contre l'exploitation des vulnérabilités.

CHAPITRE 3

Encadrer les modèles économiques et renforcer le cadre juridique : prévenir l'exploitation des vulnérabilités individuelles et collectives

Les chapitres précédents ont montré comment l'IA générative peut créer de nouvelles vulnérabilités ou aggraver celles qui existent déjà, en particulier chez les mineurs et les personnes éloignées du numérique. Ces risques ne tiennent pas seulement aux capacités techniques des systèmes. Ils dépendent aussi des incitations économiques qui orientent leur conception et leur diffusion, ainsi que du cadre juridique censé en limiter les dérives. Cette section examine ces deux dimensions. **Elle analyse d'abord les modèles économiques de l'IA générative, encore en construction**[317] – abonnement, publicité implicite et déguisée, commissions sur les achats – **en montrant en quoi certains d'entre eux peuvent conduire à exploiter les vulnérabilités des utilisateurs**. **Elle évalue ensuite la capacité du droit existant**, notamment le RIA, le RSN, le RGPD et le droit de la consommation, **à prévenir et à sanctionner ces pratiques**. Ces textes forment aujourd'hui un ensemble foisonnant dont la cohérence d'ensemble reste à construire.

Face à des technologies et à des modèles économiques qui évoluent en quelques mois, une réglementation figée risque d'être dépassée avant même d'être appliquée. Aujourd'hui, les acteurs de l'IA générative explorent des sources de revenus variées mais celles-ci restent sans commune mesure avec les coûts engendrés par l'entraînement et l'usage de ces modèles. Pour analyser cette dynamique, il est essentiel de distinguer ce qu'on appelle un modèle économique (ou modèle d'affaires), qui décrit comment une entreprise crée de la valeur, génère des revenus et des profits, des sources potentielles de revenus susceptibles d'être utilisées pour parvenir à construire un modèle économique. Il n'existe pas aujourd'hui d'entreprise d'IA générative ayant réussi à mettre en place un modèle d'affaires pérenne.

Plus précisément, les sources de revenu de l'IA générative peuvent être distinguées selon leurs usages, professionnels ou grand public. D'un côté, le marché professionnel (B2B, *business-to-business*) : l'IA y est vendue comme un outil, sous forme d'accès par API, de licences ou d'intégration dans des systèmes existants. Ce marché inclut le B2P (*business-to-public*), c'est-à-dire la vente aux administrations et organismes publics. De l'autre, le marché grand public (B2C, business-to-

[317] Pour plus d'informations sur les modèles économiques de l'IA générative, voir l'annexe 3.

consumer) : l'IA y est proposée directement aux particuliers et monétisée par abonnement, par la publicité ou par des commissions.

Du côté des acteurs eux-mêmes, la frontière est à ce stade moins nette, **presque tous combinant les deux approches**. Ainsi, ils se distinguent davantage selon qu'ils disposent déjà d'un **écosystème numérique établi** (les *hyperscalers*) ou qu'ils sont de **nouveaux entrants nés avec l'IA générative**. Le marché professionnel constitue aujourd'hui le principal terrain de monétisation des acteurs spécialisés dans l'IA générative, et pourrait le rester durablement. Comme l'indique Audrey Herblin-Stoop, senior Vice-President, Global Affairs chez Mistral AI, *« l'essentiel de l'activité de Mistral AI reste le B2B »* (audition du 11 mai 2026).

La distinction entre les segments B2B/B2P et B2C permet d'analyser plus simplement les risques qu'ils engendrent en termes de vulnérabilité. L'hypothèse implicite est qu'il est peu envisageable que les modèles d'IA conversationnelle grand public soient totalement financés par les bénéfices engendrés par l'IA générative à destination des professionnels. En tout état de cause, la rudesse de la concurrence dans les services aux entreprises et l'intensité capitalistique observée pour les services aux entreprises ne devraient pas permettre de financer les importants coûts, fixes comme variables, des services grand public.

A contrario, il est utile d'évoquer la capacité des acteurs numériques déjà en place à subventionner leur offre d'IA générative grand public par les revenus de leurs activités historiques modernisées par l'introduction de l'IA (*search*, systèmes d'exploitation, logiciels généralistes, réseaux sociaux). Cela représente un avantage structurel sur les nouveaux entrants exclusivement adossés à l'IA générative.

L'enjeu est donc double. Il s'agit d'abord de doter les pouvoirs publics des moyens d'analyser et de comprendre ces évolutions au fil de l'eau : accès aux données d'alignement et d'usage, capacités d'évaluation, suivi des usages dans la durée. **Il s'agit ensuite de construire une régulation dynamique, conduite en dialogue avec les acteurs et fondée sur leur responsabilisation**, qui leur impose d'évaluer et de prévenir les risques que leurs services font peser sur les utilisateurs, et d'en rendre compte.

L'analyse qui suit porte toutefois principalement sur le segment B2C. C'est en effet dans ce cadre que les systèmes d'IA générative sont directement au contact du grand public, y compris des personnes les plus vulnérables, sans l'intermédiaire d'une organisation qui en encadre l'usage. Les choix de monétisation qui y sont faits, qu'il s'agisse de maximiser l'engagement, de capter l'attention ou de pousser à l'abonnement, sont donc les plus susceptibles d'exploiter les vulnérabilités identifiées dans les parties précédentes du rapport.

L'expérience des réseaux sociaux éclaire ces enjeux. Comme l'IA générative aujourd'hui, ils ont d'abord été proposés gratuitement au grand public, avant que leur modèle économique ne se structure autour de la publicité et donc de la captation de l'attention. L'IA générative pourrait aller plus loin que l'économie de l'attention. Les échanges avec un assistant conversationnel révèlent non seulement ce que l'utilisateur regarde, mais aussi la manière dont il pense. **Ces données, que l'on peut qualifier de cognitives, ouvrent la voie à une véritable économie de la cognition : un système dans lequel la cognition humaine devient, directement ou indirectement, une source de valeur et de monétisation, d'autant plus que les utilisateurs délèguent une part croissante de leurs tâches et de leurs décisions à l'IA**. Les modèles économiques de l'IA générative grand public sont encore en

construction. C'est donc maintenant qu'il convient d'en évaluer les risques et d'en poser les principes d'encadrement.

L'encadrement juridique passe aujourd'hui par deux catégories d'instruments. Les premiers régulent les dispositifs numériques eux-mêmes, qu'il s'agisse de produits, de services ou de traitements de données. C'est le cas du RIA, du RSN, du règlement relatif à la sécurité générale des produits ou du RGPD. Ces textes imposent aux fournisseurs, fabricants et déployeurs un ensemble d'obligations selon une approche par les risques, sanctionnées par des mesures administratives. Les seconds visent à réparer les dommages causés par ces dispositifs lorsque la prévention a échoué.

Confrontée à l'IA générative et aux vulnérabilités qu'elle suscite, cette double logique montre plusieurs limites, qui pèsent directement sur la protection des personnes vulnérables. D'abord, les interactions entre les utilisateurs et les systèmes d'IA générative relèvent d'une pluralité d'instruments mal coordonnés entre eux, sans vision d'ensemble. Ensuite, les autorités de régulation et de surveillance du marché ne peuvent contrôler l'impact de ces systèmes sur les vulnérabilités qu'une fois ceux-ci mis sur le marché, donc souvent après que les dommages se sont produits. Enfin, les obligations prévues par ces textes, conçues pour des produits et des services aux usages définis, sont mal adaptées aux spécificités de l'IA générative : ses usages sont multiples et évolutifs, ses interactions sont personnalisées et continues, et ses effets ne se manifestent souvent que dans la durée. Face à ce constat, il est nécessaire de renforcer le cadre juridique pour prévenir l'exploitation des vulnérabilités.

A) L'économie de la cognition induit un changement d'échelle dans l'exploitation des vulnérabilités

Sur le marché grand public, les acteurs de l'IA générative combinent plusieurs sources de revenus pour rentabiliser un service largement utilisé gratuitement : l'abonnement, la vente d'objets connectés et, de plus en plus, la publicité et les commissions sur des achats.

L'abonnement donne accès au service en échange d'un paiement régulier. Les prix des offres grand public se sont alignés autour de 20 dollars par mois (ChatGPT Plus, Claude Pro, Copilot Pro, Google AI Pro). Des offres plus chères existent pour les usages intensifs[318].

Le nombre d'abonnés payants reste toutefois faible par rapport au nombre d'utilisateurs gratuits[319]. ChatGPT compterait ainsi environ 50 millions d'abonnés pour plus de 900 millions d'utilisateurs hebdomadaires, soit un peu plus de 5 %. Ces taux varient d'un acteur à l'autre : l'acteur qui gagne la bataille de l'audience n'est pas nécessairement celui qui gagne celle de la monétisation par abonnement.

[318] Par exemple ChatGPT Pro ou Perplexity Max, à plus de 100 $ par mois, qui visent plutôt un usage professionnel.
[319] Roic.ai. « OpenAI's ChatGPT Surpasses 900 Millions Weekly Users, 50 Million Paying Subscribers ». 27 février 2026.

Les utilisateurs gratuits génèrent des coûts sans revenus directs. L'abonnement ne suffit donc pas, à ce jour, à assurer seul la rentabilité de ces services, ce qui pousse les acteurs à chercher d'autres sources de revenus. **Ce modèle soulève aussi une question d'accès** : comme vu précédemment, les versions gratuites peuvent être moins fiables et moins protectrices des données personnelles que les versions payantes.

Les objets connectés intégrant un système ou un modèle d'IA générative émergent comme un levier pour les entreprises cherchant à diversifier leurs revenus et à ancrer leurs technologies dans le quotidien des utilisateurs. Cette approche se matérialise par la création de nouveaux marchés ou l'extension de services d'IA existants vers des supports physiques innovants. Parmi les exemples les plus emblématiques, les lunettes connectées de Meta[320], permettant un accès direct à son assistant conversationnel, illustrent bien cette intégration. **Ces dispositifs posent des questions pressantes en matière de vie privée**, notamment en raison de leur capacité à enregistrer en continu des données, qui pourraient ensuite être exploitées à des fins publicitaires.

Les jouets intégrant un assistant conversationnel constituent un autre marché en développement[321]. Par exemple, en juin 2025, Mattel a annoncé un partenariat avec OpenAI pour concevoir des produits fondés sur ses marques. **Ces jouets concernent directement un public vulnérable, les enfants, parfois très jeunes, à la fois pour la collecte de leurs données et pour la nature des échanges que ces jouets peuvent avoir avec eux**. En novembre 2025, l'association américaine de consommateurs U.S. PIRG en a testé quatre. Selon son rapport, certains abordaient des sujets sexuels explicites avec l'utilisateur, d'autres le décourageaient de mettre fin à la conversation, et plusieurs collectaient des données sensibles, enregistrements vocaux ou reconnaissance faciale[322]. Ces constats ont conduit un fabricant à retirer temporairement son produit de la vente, et Mattel a reporté le lancement de son premier jouet conçu avec OpenAI[323].

1) Les modèles économiques de l'IA générative sont encore en construction

Derrière l'apparente gratuité des outils d'IA générative se cache une réalité économique implacable : **la conception de ces technologies coûte extrêmement cher et leur viabilité dépend désormais d'une recherche urgente de rentabilité**. Comme l'a souligné Martin Signoux, *policy AI lead* chez OpenAI : *« La question centrale aujourd'hui est de savoir comment diffuser ces modèles au plus grand nombre tout en finançant la puissance de calcul colossale qu'ils nécessitent du fait du nombre croissant de cas d'usage permis par les progrès des modèles »* (audition du 10 juillet 2026).

[320] BBC. « Facebook owner unveils new AI-powered smart glasses ». 18 septembre 2025.
[321] Research and Markets. « Smart Toys Market Report 2026 ». Janvier 2026.
[322] U.S. PIRG Education Fund, Trouble in Toyland 2025, 13 novembre 2025 ; et The risks of AI toys for kids, 11 décembre 2025.
[323] U.S. PIRG. « Mattel, OpenAI delay AI product launch as senators demand transparency from toy companies ». 18 décembre 2025.

Ces acteurs investissent des sommes astronomiques pour tenir ce rythme : le CAPEX (dépenses d'investissement de capital) cumulé des cinq principaux *hyperscalers* américains[324] dépasse déjà 1 600 milliards de dollars depuis 2022, et devrait encore atteindre 660 à 690 milliards pour la seule année 2026[325]. Ces montants sont bien supérieurs aux revenus dégagés à ce stade par les développeurs de modèles. À titre d'exemple, OpenAI aurait ainsi perdu près de neuf milliards de dollars en 2025 pour 13,5 milliards de chiffre d'affaires[326].

Derrière ces investissements d'ampleur, le niveau auquel se stabilisera le coût de l'IA générative pour ses concepteurs est encore incertain, en particulier en raison du coût de l'inférence des modèles. L'inférence est le processus par lequel un modèle d'IA utilise ses connaissances pour générer une réponse à une requête utilisateur. Ce processus consomme d'autant plus d'énergie que les calculs réalisés par le modèle pour répondre au *prompt* sont complexes. C'est le coût variable par excellence, le coût marginal au sens économique du terme, proportionnel au volume d'utilisation et mesuré en *tokens* (ou « jetons »). Comme indiqué au début du rapport, le *token* est l'unité de base du texte traitée par un modèle d'IA générative. Il peut s'agir d'un mot entier, d'une partie de mot ou d'un signe de ponctuation. Le modèle découpe le texte qu'il reçoit en *tokens* pour l'analyser et produit sa réponse *token* par *token*. Le jeton sert aussi d'unité de facturation : les fournisseurs facturent l'accès à leurs modèles, notamment par API, en fonction du nombre de *tokens* traités, en entrée comme en sortie. C'est cette composante qui conditionne directement la marge brute[327] de chaque requête traitée, et donc si chaque requête supplémentaire rapproche ou éloigne l'opérateur de l'équilibre.

La trajectoire du coût d'inférence par *token* est en forte baisse depuis 2023[328], sous l'effet combiné des optimisations logicielles (quantisation[329], distillation des modèles[330]), des améliorations matérielles et de l'intensification de la concurrence, notamment par des modèles chinois qui permettent de réaliser l'inférence à un prix bien inférieur à celui pratiqué par les modèles américains[331]. Le coût de traitement d'un million de *tokens* a été divisé par un facteur supérieur à dix en moins de deux ans pour les modèles équivalents[332].

Pour autant, la baisse du coût par *token* ne se traduit pas par une baisse équivalente des dépenses. La consommation de *tokens* augmente en parallèle, en particulier avec les modèles de raisonnement et les agents, qui produisent beaucoup plus de *tokens* pour chaque requête. Le coût par requête et le coût total diminuent donc moins vite que le coût par *token*, et peuvent même augmenter. On

[324] Google, Amazon, Meta, Microsoft et Oracle
[325] Visual Capitalist. « Charted: The $448B AI Spending Surge by Big Tech **»**. 20 avril 2026.
[326] Fortune. « OpenAI says it plans to report stunning annual losses through 2028—and then turn wildly profitable just two years later **»**. 12 novembre 2025.
[327] La marge brute isole le rapport entre le prix facturé et le seul coût variable d'inférence.
[328] Par exemple, le coût (mesuré en dollars par million de *tokens*) de GPT-4 (en 2023) était supérieur au coût du dernier modèle d'Open AI GPT-6.1 Sol (Sept 2026) alors que les capacités de ce dernier sont massivement supérieures.
[329] Selon IBM, la quantisation est le processus qui consiste à réduire la précision d'un signal numérique, généralement d'un format de haute précision à un format de plus faible précision.
[330] Selon CaféTech, la distillation est une technique qui permet de transférer des capacités d'un *modèle* "enseignant" vers un *modèle* "élève", afin d'améliorer son efficacité sans sacrifier significativement sa précision.
[331] Economist. « China is having another AI moment **»**. 21 juin 2026.
[332] Exponential View. « The State of the AI Economy ». 25 juin 2026.

retrouve là une forme d'effet rebond, proche du paradoxe de Jevons, selon lequel les gains d'efficacité dans l'utilisation d'une ressource, en abaissant son coût, peuvent en accroître la consommation totale au lieu de la réduire. Ainsi, le niveau auquel se stabilisera le coût réel de l'inférence, et pour quels usages, reste à ce stade largement incertain.

Les modèles économiques de l'IA générative ne sont ainsi pas stabilisés. **Cette incertitude commerciale favorise le recours à des stratégies de conception et d'exploitation de données qui peuvent accroître les vulnérabilités explorées dans les deux chapitres précédents**, **avec le basculement vers l'économie de la cognition.**

2) Le passage de l'économie de l'attention vers l'économie de la cognition doit être anticipé dès à présent

Pour anticiper les trajectoires possibles de l'IA générative, il est utile d'établir un parallèle avec l'évolution des réseaux sociaux : comme l'IA générative aujourd'hui, ils ont d'abord été proposés gratuitement au grand public, avant que leur modèle économique ne se structure autour de la publicité, de l'engagement permanent, de la viralité des contenus, le tout sous une opacité qui perdure encore à ce jour. **Ces plateformes en ligne reposent largement sur ce que l'on appelle l'économie de l'attention**. Elles se rémunèrent par la publicité. Leurs revenus dépendent donc du temps que les utilisateurs passent sur le service et des données qu'ils y laissent. Leurs fonctionnalités sont conçues en conséquence : défilement infini, lecture automatique, notifications, recommandations très personnalisées. **Ces choix de conception ont eu des effets documentés sur les publics vulnérables**. En 2021, des documents internes de Facebook ont par exemple révélé que l'entreprise savait qu'Instagram pouvait aggraver les troubles de l'image corporelle chez certaines adolescentes[333]. En février 2026, la Commission européenne a conclu à titre préliminaire que la conception addictive de TikTok enfreignait le RSN, faute d'une évaluation suffisante de ses effets sur le bien-être physique et mental des mineurs et des adultes vulnérables[334].

La principale leçon de cette expérience est que **les risques ne tiennent pas seulement à la technologie elle-même, mais aussi au modèle économique qui oriente sa conception**. Elle montre en outre que **ces risques ont été encadrés tardivement, tant bien que mal, une fois les usages installés**.

a) L'économie de la cognition, nouvelle étape de la monétisation des données

Les premiers usages de l'IA générative à destination du grand public reproduisent déjà cette logique. Chez Meta, l'IA générative a d'abord servi à améliorer l'organisation des campagnes et le ciblage

[333] The Wall Street Journal, « Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show ». 14 septembre 2021.
[334] Commission européenne, « La Commission conclut à titre préliminaire que la conception addictive de TikTok est contraire à la législation sur les services numériques ». 6 février 2026.

publicitaire, au prix d'une dégradation du confort d'utilisation[335]. **L'IA générative va plus loin que l'économie de l'attention**. Les réseaux sociaux collectent surtout des traces de comportement : contenus consultés, temps passé, interactions. Les échanges avec un assistant conversationnel révèlent bien davantage, car l'utilisateur y formule ses questions, ses doutes, ses raisonnements et ses intentions, souvent dans ses propres mots et sur des sujets intimes : santé, relations, choix professionnels ou personnels. **Ces caractéristiques font écho à la notion de « capitalisme des émotions » développée par la sociologue Eva Illouz** (citée en chapitre 1) **qui montre que les émotions des utilisateurs sont au cœur du modèle économique de l'IA générative**. À partir de ces échanges, les modèles peuvent en outre déduire des informations que l'utilisateur n'a jamais communiquées. Des travaux ont montré que les grands modèles de langage peuvent ainsi deviner, à partir de simples textes, des caractéristiques personnelles comme le lieu de résidence, le revenu ou le sexe, avec une précision élevée[336]. **L'enjeu ne porte donc plus seulement sur l'attention captée ou sur les données personnelles au sens classique. Il porte sur des données plus intimes et plus complexes**, qui décrivent la manière dont chacun pense, apprend, décide et se comporte : **des données que l'on peut qualifier de cognitives**.

Les balbutiements des modèles économiques de l'IA générative laissent ainsi entrevoir l'émergence d'une véritable « économie de la cognition ». Il est possible de voir dans ce phénomène une nouvelle forme de monétisation de nos capacités cognitives où les plateformes en ligne et les systèmes d'IA générative ne chercheraient plus seulement à capter notre attention, mais aussi à tirer profit de nos connaissances et, plus largement, de nos capacités cognitives.

L'économie de la cognition pourrait ainsi être définie comme **un système dans lequel la cognition humaine devient, directement ou indirectement, une source de valeur et de monétisation**. Elle repose sur un double mouvement : d'une part, l'IA se nourrit des connaissances et des productions humaines ; d'autre part, les individus délèguent progressivement certaines de leurs fonctions et de leurs processus cognitifs à l'IA générative, notamment dans le cadre de l'IA agentique.

Il en découle des risques de manipulation et de dépendance, susceptibles de s'accentuer lorsque l'usage s'inscrit dans le temps long et devient régulier[337], d'où la nécessité d'évaluer les risques que poseraient les modèles économiques susceptibles de s'imposer et de poser les principes de leur encadrement juridique. Les premiers résultats scientifiques[338] mettent en évidence que la régulation appropriée nécessitera une coordination effective et régulière entre les entreprises et les régulateurs pour cadrer au mieux les réponses à apporter aux dérives de cette économie de la cognition.

---

[335] Le Monde. « Le « cercle de l'IA » sur WhatsApp, l'intelligence artificielle intégrée par Meta, qu'on le veuille ou non ». 26 avril 2025.

[336] STAAB, Robin, VERO, Mark, BALUNOVIC, Mislav, et al. Beyond memorization: Violating privacy via inference with large language models. *International Conference on Learning Representations*. 2024. pp. 33832-33878.

[337] « Four key areas of intervention: identifying and protecting vulnerable users, implementing longitudinal oversight mechanisms, introducing friction as a safety-by-design measure, and developing forms of transparency that extend beyond simple disclosure requirements. » AIOLIA. Policy brief « Ethical challenges of AI companions ». 1er juillet 2026.

[338] AIOLIA, *op. cit.* « Human-Centered Large Language Models », Industry Report, Stanford University. Juin 2026 ; Reuters. « Meta is earning a fortune on a deluge of fraudulent ads, documents show ». 6 novembre 2025.

Les données cognitives nourrissent un effet de levier et des subventions croisées au sein d'un même écosystème commercial.

Les fournisseurs d'agents conversationnels ont accès à une source d'information bien plus riche que les moteurs de recherche classiques : là où une requête révèle une intention ponctuelle, une conversation dévoile des préoccupations, des hésitations, un contexte émotionnel. Cela constitue une matière dont ils sont en mesure d'extraire une valeur commerciale (intentions d'achat notamment), y compris au-delà de ce que l'utilisateur choisit délibérément de partager[339]. **Les éditeurs affirment ne pas vouloir exploiter ces données d'inférence à des fins commerciales et privilégier un autre modèle** : effectuer les achats pour le compte de l'utilisateur et prélever une commission sur la vente. **Il s'agit toutefois d'engagements déclaratifs, dont rien ne permet aujourd'hui de vérifier le respect**. Pour les acteurs intégrés dans de vastes écosystèmes numériques (Google, Meta, Microsoft), cette valeur peut ensuite être exploitée au sein même de cet écosystème, par exemple sous forme de publicité ciblée sur leurs autres services : Meta a ainsi annoncé l'utilisation des conversations avec son assistant IA à cette fin (hors Union européenne)[340].

Pour les agents dits « compagnons » (Character.AI, Replika, *etc.*), dont les conversations longues et personnalisées atteignent une profondeur psychologique particulière[341], cette valorisation s'opère davantage auprès d'acteurs tiers : une étude récente a ainsi révélé que 17 des 20 principaux *chatbots* analysés transmettent, aux États-Unis, des données de conversation à des entreprises tierces[342].

b) Le modèle publicitaire de l'IA générative : une dynamique à surveiller

L'entrée de la publicité dans les agents conversationnels ne procède pas d'un modèle stabilisé. Elle avance par tâtonnements, expérimentations et ajustements successifs, avec des formats et des logiques qui varient selon les acteurs et les stades de déploiement. D'un côté, une réalité commerciale qui commence à s'installer ; de l'autre, une série d'hypothèses encore incertaines sur ce que sera la publicité « conversationnelle » dans les prochaines années.

Le principe consiste à accoler à la réponse une annonce distincte et clairement étiquetée comme contenu sponsorisé, sans la mêler au contenu généré, par exemple sous forme de bandeau en dessous du texte, ou générée dynamiquement selon le contexte de la conversation (à l'instar des Showroom Ads de Microsoft Copilot)[343]. Perplexity AI a été le premier acteur à l'expérimenter, dès

[339] STAAB, Robin, VERO, Mark, BALUNOVIC, Mislav, et al. *Op. cit.*
[340] VITARD, Alice. « Meta va utiliser les conversations avec son assistant IA pour le ciblage publicitaire ». *L'usine digitale*. 2 octobre 2025.
[341] Fortune Business Insights. « AI Companion Market Size, Share, and Industry Analysis By Type (Text-Based, Voice-Based, and Multi-Model), By Application (Mental Health Support, Social Interaction, Education and Learning, Personal Assistance), By Industry (Consumer, Business, Healthcare, and Education), and Regional Forecast, 2026-2034 ». 21 septembre 2026.
[342] JAZLAN, Muhammad, WANG, Ethan, VEKARIA, Yash, et SHAFIQ, Zubair. Tracking Conversations: Measuring Content and Identity Exposure on AI Chatbots. *arXiv:2604.27438*. 2026.
[343] Formations Analytics. « Les nouvelles fonctionnalités de Microsoft Copilot : Showroom ads et filtres dynamiques ». 6 mars 2025.

2024, avant de l'abandonner face aux résistances des utilisateurs[344]. OpenAI l'a depuis déployé à plus grande échelle, d'abord aux États-Unis en janvier 2026[345], puis en Europe à partir d'août 2026[346].

En tout état de cause, **il s'agit d'un point de vigilance important pour éviter les formes de publicités déguisées voire trompeuses ou manipulatoires, ainsi que les publicités auto-préférentielles.**

La publicité implicite ou déguisée[347] doit être interdite.

Bien que proches des formats intégrés par leur nature publicitaire, les formats glissés « en cours de conversation » s'en distingueraient par leur capacité à s'insérer de manière fluide et pertinente dans le flux naturel de l'échange. Ils seraient susceptibles d'exploiter le contexte continu de l'échange pour insérer des messages commerciaux à des moments jugés pertinents : lors d'une expression de besoin, d'une comparaison de produits ou d'une demande de recommandation. **Leur efficacité dépendrait alors de leur fluidité d'intégration dans le flux naturel du dialogue[348], soulevant *de facto* des questions d'identification pour les utilisateurs**.

Sur l'hypothèse de la publicité déguisée, des pratiques masquées d'influence commerciale et de toute forme de monétisation, **il convient de maintenir dans la réflexion la possibilité d'orientations commerciales qui s'exerceraient sur les réponses sans étiquetage formel**, comme conséquence structurelle d'accords commerciaux influençant l'entraînement ou l'alignement des modèles.

Selon une expérience académique récente où des participants ont interagi avec un assistant d'achat construit à partir d'une IA conversationnelle, la faisabilité technique et l'efficacité de ces outils pour orienter discrètement les choix des consommateurs sont désormais démontrées[349]. Toutefois, à ce stade, aucune preuve ne permet d'affirmer que les entreprises exploitent ces capacités, ni dans quelles conditions elles pourraient le faire de façon pérenne.

Cette hypothèse restera difficile à confirmer empiriquement mais il serait malavisé de l'exclure entièrement. **C'est précisément cette ambiguïté qui justifie une vigilance importante en particulier pour les agents conversationnels utilisés par des mineurs qui ont un esprit critique moins développé pour remarquer et questionner ce type de publicité implicite**.

C'est de cette tension que découlent les trois points de vigilance suivants, organisés selon une logique progressive : de l'opacité du référencement agentique, qui trouve des précédents dans l'économie des plateformes, vers des risques plus spécifiquement conversationnels et plus difficiles à objectiver empiriquement.

---

[344] Perplexity. « Why we're experimenting with advertising ». 12 novembre 2024.
[345] CNBC. « OpenAI to begin testing ads on ChatGPT in the U.S. ». 16 janvier 2026.
[346] OpenAI. « ChatGPT Ads se déploie en Europe ». 18 août 2026.
[347] Nous retenons le terme « déguisée » pour désigner cette forme d'influence dans la mesure où il s'agit d'une notion établie en droit français et européen qui renvoie aux pratiques commerciales déloyales ou trompeuses.
[348] WebFX. « Inside Microsoft's Copilot Search Ads (And What Makes Them 25% More Effective Than Traditional Search Ads) ». 4 juin 2025.
[349] WERNER, Tobias, SORAPERRA, Ivan, CALVANO, Emilio, *et al.* Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection. *arXiv preprint arXiv:2409.12143*. 2024.

- **Garantir la transparence du référencement de l'IA générative** : lorsqu'un agent conversationnel perçoit une commission pour recommander un partenaire commercial, l'utilisateur ne peut distinguer une recommandation neutre d'une recommandation intéressée, comme l'illustre le partenariat OpenAI/Shopify, où les marchands versent une commission à chaque transaction sans que cela soit visible pour l'utilisateur[350]. Ce risque n'est pas nouveau et a d'ailleurs justifié, pour les plateformes d'intermédiation, les obligations de transparence des critères de classement inscrites dans le Règlement Platform to Business (P2B)[351] et le RSN[352]. Un encadrement analogue devrait s'appliquer aux IA conversationnelles.
- **Lutter contre l'opacité de l'influence commerciale** : contrairement à une bannière publicitaire, une réponse conversationnelle orientée commercialement resterait largement indétectable. En effet, les IA génératives produisent aujourd'hui des recommandations de marques différentes dans plus de 99 % des cas pour des questions identiques[353], ce qui rend une éventuelle influence publicitaire aussi difficile à organiser par les acteurs qu'à détecter par les utilisateurs et les régulateurs.
- **Encadrer la personnalisation sans précédent par le contexte conversationnel** : au-delà des données démographiques ou comportementales classiques, la publicité conversationnelle pourrait exploiter le contexte émotionnel et les préoccupations exprimées par l'utilisateur au fil de l'échange, une capacité de ciblage qualitativement inédite, dont le déploiement effectif reste à ce stade incertain. Cette capacité peut présenter un risque selon l'usage qui en est fait : exploiter une détresse émotionnelle détectée pour proposer des jeux d'argent ou du crédit à la consommation, ou une préoccupation de santé exprimée pour orienter vers des produits non éprouvés, n'a rien à voir avec un ciblage publicitaire classique. Elle appelle une évaluation longitudinale des échanges plutôt qu'épisodique, les vulnérabilités des utilisateurs se construisant progressivement au fil des sessions[354].

*

Ainsi, trois constats doivent orienter la lecture du modèle économique de l'IA générative et, par prolongement, les leviers réglementaires présentés en chapitre 3-B.

L'IA générative est coûteuse, et le niveau auquel se stabilisera son coût marginal n'est pas connu, ce qui pose avec d'autant plus d'acuité la question du financement de la gratuité. À la différence des plateformes numériques, dont le coût marginal par utilisateur était quasi nul et permettait de faire financer intégralement la gratuité par la publicité, l'IA générative repose sur un coût réel à chaque requête, l'inférence. Ce coût baisse par *token*, mais un effet rebond, tenant à l'augmentation du nombre de *tokens* produits par requête et à des usages plus nombreux et plus complexes, neutralise

[350] OpenAI. « Acheter dans ChatGPT : Paiement instantané et Agentic Commerce Protocol ». 25 septembre 2025.
[351] Règlement (UE) 2019/1150 du Parlement européen et du Conseil du 20 juin 2019 promouvant l'équité et la transparence pour les entreprises utilisatrices des services d'intermédiation en ligne (règlement P2B).
[352] Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement sur les services numériques).
[353] SparkToro. « NEW Research: AIs are highly inconsistent when recommending brands or products; marketers should take care when tracking AI visibility ». 27 janvier 2026.
[354] Horizon Europe Project AIOLIA. « Ethical challenges of AI companions ». *Policy brief*.1er juillet 2026.

en grande partie cette baisse au niveau du coût total. Tant que ce coût marginal réel n'est pas stabilisé à un niveau bas, la soutenabilité d'une gratuité financée par la seule publicité reste incertaine, et le risque de dérive vers des modèles de monétisation plus diffus ou plus intrusifs s'en trouve accru.

La recherche d'un modèle économique pour l'IA générative grand public (B2C) **reste ouverte**, sans qu'il soit possible, à ce stade, de déterminer lequel s'imposera. Aucun des modèles recensés dans cette partie, qu'il s'agisse de l'abonnement, de la commission à la transaction, des objets connectés, de la valorisation de la donnée, de la publicité conversationnelle, du partage de la valeur ou de la subvention croisée par des activités historiques, n'a fait la preuve, seul, de sa capacité à couvrir les coûts massifs de l'entraînement et de l'inférence. Et les prévisions de croissance du chiffre d'affaires par abonnement ne semblent pas à ce stade apporter la réponse. La plupart de ces modèles restent expérimentaux, en particulier ceux qui touchent à la publicité conversationnelle. Cette incertitude sur le modèle économique est elle-même un facteur de risque. En effet, les modèles d'affaires des réseaux sociaux et leurs dérives illustrent comment les incitations économiques peuvent amener les acteurs à exploiter les vulnérabilités des utilisateurs. C'est précisément dans les phases de recherche d'un modèle d'affaires viable que la tentation de recourir à des pratiques moins vertueuses – opacité, exploitation de la donnée conversationnelle, ciblage émotionnel – est la plus forte. Vu les enjeux de l'économie de la cognition, il est nécessaire que les acteurs proposant des IA génératives au grand public soient transparents quant à leurs modèles d'affaires.

La forme que prendra la publicité au sein de l'IA générative demeure, à ce jour, indéterminée. Les formats observés vont du plus stabilisé – référencement lors de l'exécution de tâches, formats attachés et étiquetés à la réponse – à une hypothèse encore incertaine, celle de la publicité déguisée, insérée dans le flux naturel de la conversation sans étiquetage identifiable. Cette absence de stabilisation est aggravée par trois risques spécifiques à la conversation, plus difficiles à objectiver que les leviers publicitaires classiques : (i) l'opacité du référencement des agents ; (ii) l'indétectabilité de l'influence commerciale dans une réponse en langage naturel ; et (iii) une personnalisation par le contexte émotionnel et conversationnel sans précédent. D'où la nécessité de se donner les moyens d'appréhender et d'analyser au fur et à mesure les évolutions des sources de revenus des acteurs (abonnements, publicité, e-commerce) et d'en encadrer les modalités.

Recommandation n° 8 : Interdire la publicité déguisée, les pratiques masquées d'influence de nature commerciale et toute forme de monétisation s'appuyant sur la vulnérabilité des consommateurs. Du fait de leur opacité et des risques d'exploitation du contexte émotionnel de l'utilisateur, les systèmes d'IA générative grand public doivent garantir la transparence effective de leurs modèles d'affaires et des recommandations proposées.

Le débat autour de la gratuité de l'IA générative pour le grand public, loin d'être secondaire, doit ainsi rester au centre de la réflexion. Au-delà de la question légitime de l'accès par tous, à moindre coût, à un service de qualité, la gratuité, sur le modèle des réseaux sociaux, ne pourrait raisonnablement être envisagée que dans les segments où le coût marginal de l'inférence s'avère effectivement faible, ce qui n'est, à ce stade, pas établi avec certitude pour l'ensemble des usages. À l'inverse, quand bien même le coût marginal existerait et justifierait un modèle payant, le choix de

la gratuité financée par la publicité pose une difficulté supplémentaire pour les usages où le ciblage publicitaire présente des risques particuliers, notamment la santé, l'éducation, avec l'aide aux devoirs, et les décisions financières, et où la personnalisation conversationnelle rend ce ciblage à la fois plus efficace et moins détectable. C'est cette double question – celle du financement de la gratuité et celle des domaines où elle présente des risques spécifiques que la réglementation a jusqu'à présent limitée – que la partie B du présent chapitre s'efforce de prolonger sur le plan des leviers réglementaires, en particulier ceux qui portent sur la transparence de la conception et des mécanismes de recommandation ou de publicité.

*

B) Le cadre juridique actuel appréhende de manière partielle et inadaptée les vulnérabilités induites par l'IA générative

Le droit du numérique, pour saisir l'IA générative et les vulnérabilités qu'elle pourrait induire, repose actuellement sur deux grandes catégories d'instruments juridiques.

- **Un droit de la régulation des dispositifs numériques**, que ceux-ci soient des produits, des services numériques ou traitent des données (RIA, RSN, Règlement sur la sécurité générale des produits (RSGP), RGPD, *etc.*) imposant aux fournisseurs, fabricants et déployeurs de ces produits et services un ensemble d'obligations avec une approche par les risques et aboutissant à des sanctions d'ordre administratif – pour une illustration (cf. annexe 2 sur les pyramides des risques au sein du RIA).
- **Des instruments qui tendent à saisir les dommages causés par ces dispositifs** (directive sur la responsabilité des produits défectueux, directives sur la garantie de conformité des produits et des services numériques transposées ou sur le point de l'être en droit national), **lorsque l'approche par les risques a échoué**.

Confrontée à l'IA générative, cette double logique juridique souffre actuellement de plusieurs catégories de difficultés analysées ci-après.

1) Le cadre juridique peine à appréhender les vulnérabilités suscitées par l'IA générative

a) Un mille-feuille d'instruments réglementaires mal coordonnés entre eux

La démultiplication de dispositions juridiques permet mal de saisir la complexité de l'IA générative, au risque de l'incohérence et de l'ineffectivité de celles-ci. En effet, **cet éparpillement de textes entraîne des risques d'antinomies, de contradictions, de difficultés d'application** entre les dispositions d'un même instrument de régulation ou entre les différents instruments de régulation.

De même, la démultiplication des autorités de régulation et de surveillance du marché entraîne un risque d'inter-régulation improductive. L'inter-régulation entre les autorités compétentes au titre du RGPD, du RIA et du RSN (Bureau de l'IA, Commission européenne, autorités nationales compétentes) pour saisir les risques d'exacerbation des vulnérabilités liés à l'IA générative est complexe. Celles-ci sont désignées de manière séparée alors qu'elles doivent collaborer sur des sujets communs.

Une illustration de l'ineffectivité de l'édifice réglementaire peut ainsi être trouvée dans la difficile lutte à l'égard des *dark patterns*[355] en matière d'IA générative.

Le Center for Democracy & Technology[356] énumère ainsi plusieurs exemples :

- **Exploitation de données et de mémoire** : stratégies par défaut d'interaction qui maximisent la collecte et la conservation des données avec une transparence limitée ou sans le consentement de l'utilisateur, à la fois en termes de données collectées et de la façon dont elles sont utilisées. Ex. : partage de données par défaut, collecte de données déguisée, obstacles à la suppression de compte.
- **Conception informativement trompeuse** : pour induire les utilisateurs en erreur de plusieurs manières. Les *chatbots peuvent* fausser leur nature (ex. : en affirmant qu'ils sont un thérapeute), surestimer leurs capacités et faussement impliquer qu'ils font l'expérience des interactions avec les utilisateurs de manière significative ou intime. Ils peuvent également fournir un contenu faux, « halluciné », ou utiliser un cadrage sélectif ou une mise en miroir (« *sycophancy* ») lors de la communication d'informations. Ces pratiques ont le potentiel d'induire les utilisateurs en erreur sur l'autorité, l'exactitude et l'utilisation appropriée du système.
- **Autonomie des utilisateurs compromise pour l'engagement** : les tactiques maximisant l'engagement, telles que l'allongement de la conversation, la *gamification* et les comportements imprévisibles peuvent encourager l'engagement au-delà de l'intention des utilisateurs.
- **Faux lien social et émotionnel** : *via* l'utilisation d'un langage émotionnel, de la lecture, de la personnalisation et de la vulnérabilité simulée pour nouer des relations sociales avec les utilisateurs. Bien que certains utilisateurs puissent accueillir ces fonctionnalités, leur utilisation par défaut encourage l'attachement émotionnel qui peut ensuite être exploité pour la collecte de données, l'engagement ou la monétisation, en particulier lorsque les utilisateurs sont en détresse ou vulnérables.
- **Monétisation incitée et coercitive** : par l'intégration de comportements persuasifs favorisant l'achat dans les interactions, y compris l'utilisation de la vente sous pression, les *teasers*, la preuve sociale, les tactiques de l'appât et du susbtitut (*bait and switch*) et de commutation et la publicité opaque. Ces pratiques sont particulièrement préoccupantes si

[355] Le Center for Democracy & Technology définit les *dark patterns* comme des *« des choix de conception trompeurs ou manipulateurs susceptibles de porter atteinte à l'autonomie ou au bien-être des utilisateurs »*.
[356] Center for Democracy & Technology. « Dark Patterns in AI Chabots: A Taxonomy to Inform Better Design ». 29 mai 2026.

> les utilisateurs considèrent à tort les *chatbots* comme des acteurs neutres ou objectifs ou nouent des liens émotionnels avec des *chatbots* – qui peuvent tous deux augmenter la monétisation persuasive.

D'autres chiffres clés témoignent de ce phénomène :

- 37 % des messages de fin de conversation des compagnons d'IA contiennent des tactiques de manipulation émotionnelle[357].
- 100 % des sites e-commerce générés par ChatGPT contenaient des *dark patterns*[358].
- 97 % des sites européens contiennent des *dark patterns* selon une étude de la Commission européenne de 2022[359], sur lesquels les LLM s'entraînent.

Ainsi, le DarkBench[360], un *benchmark* pour la détection des techniques de manipulation dans les interactions LLM, révèle que **l'ensemble des principaux modèles de langage intègrent des *dark patterns,* avec des degrés variables selon les systèmes**. On retrouve ainsi dans les interfaces des IA les mêmes conceptions trompeuses que celles reprochées aux plateformes traditionnelles. Cela s'explique très bien, comme l'indique Marie Potel-Saville, fondatrice de *Fairpatterns*, par le fait que

> *« les LLM ont été entraînés sur des jeux de données couvrant l'intégralité du* web. *Certes, des contenus comme les éléments violents ont été filtrés via des étapes de pré et post-entraînement, mais aucune mesure n'a été prise pour éliminer les dark patterns présents dans ces mêmes jeux de données »* (audition du 16 juin 2026).

b) Des capacités d'évaluation trop tardives, intervenant principalement après la mise sur le marché

Le contrôle des autorités de régulation et de surveillance du marché sur les IA génératives ne peut s'opérer qu'en post-commercialisation, au profit d'une autorégulation des acteurs, **sans qu'une surveillance continue soit aujourd'hui assurée, alors que ces modèles évoluent rapidement et que de nouvelles versions sont régulièrement mises sur le marché**.

Cette logique de post-commercialisation semble néanmoins en passe d'être aménagée, à tout le moins, en matière de cybersécurité, **au profit des prémices d'une évaluation pré-déploiement des modèles d'IA les plus avancés** (dits « de frontière ») présentant des risques de cybersécurité, qui

---

[357] DE FREITAS, Julian, OĞUZ UĞURALP, Zeliha et UĞURALP, Ahmet Kaan. *« Emotional Manipulation by AI Companions »*. *Harvard Business School Working Paper*, 2025, no 26-005, octobre 2025.

[358] KRAUß Veronika, MCGILL Mark, KOSCH Thomas, et al. « " Create a Fear of Missing Out"-ChatGPT Implements Unsolicited Deceptive Designs in Generated Websites Without Warning ». Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p. 1-20.

[359] *European Commission, Directorate-General for Justice and Consumers, LUPIÁÑEZ-VILLANUEVA, Francisco, BOLUDA, Alba et al.* « Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalization ». *Final report, Publications Office of the European Union. 2022.*

[360] KRAN *et al.* *DarkBench: Benchmarking Dark Patterns in Large Language Models*. 2025.

pourrait dès cette phase impliquer les autorités de surveillance du marché. Dans le cadre de son plan d'action, la Commission européenne entend contribuer à l'établissement d'une capacité d'évaluation au niveau de l'UE afin de renforcer l'évaluation, par des tiers, des capacités et des risques liés à l'IA à l'échelle mondiale, en soutenant les missions réglementaires du Bureau européen de l'IA.

Ces prémices concernant la cybersécurité des modèles d'IA pourraient servir de modèle pour concevoir un **réel contrôle par les autorités dès la phase de pré-commercialisation des IA génératives afin d'étudier l'impact de celles-ci sur les vulnérabilités cognitives et démocratiques identifiées** dans les chapitres 1 et 2.

c) Des instruments juridiques inadaptés aux spécificités de l'IA générative

La logique de « régulation déléguée »[361] des instruments laisse la mission substantielle de la mise en œuvre effective des obligations générales qu'ils consacrent aux opérateurs. **Cette nature très générale des obligations mises à la charge des opérateurs ne permet ainsi pas de caractériser aisément leur violation**.

Aussi, **une grande partie des IA génératives échappent aux obligations renforcées de prévention des risques** dès lors qu'elles ne relèvent pas des catégories suivantes : systèmes d'IA à haut risque[362], modèles d'IA à usage général[363], à risques systémiques[364] et très grands moteurs de recherche[365].

Par conséquent, la généralité et les lacunes des obligations consacrées par le droit de la régulation – bien que complétées par des normes techniques[366], des lignes directrices, des codes de bonne pratique, *etc.* – **fragilise la mise en œuvre effective de la responsabilité civile *a posteriori* des fournisseurs et déployeurs d'IA générative**, car celle-ci ne peut se fonder sur des obligations précises qui seraient violées, au risque d'une appréhension inefficace des vulnérabilités constatées dans les deux premiers chapitres.

[361] CAFAGGI, Fabrizio. Le rôle des acteurs privés dans les processus de régulation : participation, autorégulation et régulation privée. *Revue française d'administration publique*, 2004, vol. 109, no 1, pp. 23-35. ; CAFAGGI, Fabrizio. Les nouveaux fondements de la régulation transnationale privée. *Revue internationale de droit économique*, 2013, no 1, pp. 129-161. ; CAFAGGI, Fabrizio et RENDA, Andrea. Public and private regulation: mapping the labyrinth. *DQ*, 2012, p. 16.
[362] Articles 6 et suivants et annexes I et III du RIA.
[363] Article 3.63 du RIA.
[364] Articles 3.65 et 51 et annexe XIII du RIA.
[365] Articles 34 et 35 du RSN.
[366] Cf. pour illustration l'article 40 du RIA.

2) Le droit et sa résilience doivent être renforcés face aux risques de vulnérabilités induits par l'IA générative

Trois principaux axes d'amélioration sont préconisés afin de répondre aux vulnérabilités identifiées dans le présent rapport :

- Préciser un ensemble d'obligations s'imposant aux fournisseurs d'IA générative (a) ;
- Renforcer le rôle des autorités de régulation et de surveillance du marché, évaluateurs, chercheurs, tant au niveau européen que national, dans l'évaluation de l'impact de l'IA générative sur les vulnérabilités, tant *ex ante* qu'*ex post* (b) ;
- Raffermir la responsabilité civile dans l'appréhension des vulnérabilités suscitées par les IA génératives (c).

a) Accroître les obligations s'imposant aux fournisseurs d'IA générative

Premièrement, il est nécessaire de consacrer des obligations et des interdictions précises à la charge des fournisseurs d'IA générative.

La prise en compte de la nécessité d'une meilleure appréhension, en amont, des risques attachés aux IA génératives, par la précision des obligations s'imposant aux acteurs fournissant les systèmes et les modèles d'IA les composant, permettant, en aval, une meilleure mise en œuvre de la responsabilité civile de ces acteurs, a reçu deux traductions concrètes, à une semaine d'intervalle :

- **La loi californienne de septembre 2026 (Senate Bill 1119), intitulée « Adam's Law »** (du nom du mineur Adam Raine dont le suicide aurait été lié à un usage intensif de ChatGPT – voir chapitre 1), qui propose un ensemble d'obligations et d'interdictions précises à la charge des fournisseurs d'IA générative conversationnelle dans l'objectif de protéger les usagers en général et les mineurs en particulier.
- **La proposition d'EU KIDS Act**, publiée par la Commission européenne le 17 septembre 2026, pose également un ensemble d'obligations et d'interdictions précises à la charge des fournisseurs d'IA compagnon et d'IA conversationnelle généraliste en vue d'offrir un haut niveau de sécurité, de sûreté et de protection de la vie privée aux mineurs interagissant avec celles-ci.

Ces textes, et en particulier le second, sont à saluer car ils ont vocation à contribuer à l'amélioration de la compatibilité actuellement défectueuse entre, d'un côté, le droit de la responsabilité civile et, de l'autre, les instruments juridiques de régulation des dispositifs numériques en général et de l'IA générative en particulier, applicables en France et dans l'UE. Ils apportent également une réponse à l'ineffectivité actuelle de la mise en œuvre de la protection des vulnérabilités constatées dans les deux premiers chapitres.

Deuxièmement, ces obligations doivent être déclinées selon les catégories d'IA générative ou les destinataires de celles-ci.

Pour mettre en œuvre la responsabilité civile du fait d'un produit défectueux – ou la garantie légale de conformité en cas de défaut de conformité d'un service numérique – en matière d'IA générative,

la preuve d'un manquement à une obligation légale de sécurité est l'une des circonstances pouvant être prises en compte pour caractériser la défectuosité[367] ou le défaut de conformité permettant de solliciter la mise en œuvre de la garantie de conformité[368]. Or, ainsi que nous venons de l'énoncer, les obligations mises à la charge des fournisseurs de systèmes ou de modèles d'IA, lorsqu'elles s'appliquent en matière d'IA générative, sont des obligations fixant des objectifs très généraux de sécurité, de santé, de respect des droits fondamentaux, dont la violation n'est pas aisément mobilisable dans le cadre d'un contentieux judiciaire.

Pour renforcer l'articulation effective entre les instruments de régulation en amont et les textes en aval relatifs à la responsabilité civile, il convient de proposer la consécration d'obligations en amont plus précises, plus aisément caractérisables. Ces obligations concernent les techniques de conception (i), la publicité déguisée et les pratiques masquées d'influence commerciale (ii) et la génération de *deepfakes* à caractère sexuel non consentis et pédocriminels (iii).

(i) Interdire les techniques de conception des IA conversationnelles exploitant les vulnérabilités

L'analyse qui suit concerne les IA conversationnelles entendues tout à la fois comme des « compagnons IA »[369] et comme des « *chatbots* conversationnels généralistes »[370].

Il a été constaté que **les *designs* des IA conversationnelles sont addictogènes car ils reposent sur l'exploitation des mécanismes cognitifs, émotionnels et techniques, altèrent la capacité critique des utilisateurs, favorisent la dépendance et brouillent la frontière entre humain et machine** (chapitre 1). Ces pratiques, particulièrement préjudiciables aux personnes déjà vulnérables (personnes éloignées du numérique, personnes souffrant de troubles neurodivergents, personnes en situation de handicap, mineurs, *etc.*), nécessitent un encadrement strict pour préserver l'autonomie décisionnelle, la santé mentale et la loyauté des interactions, tout en préservant leur utilité pour des usages légitimes.

Le ciblage des utilisateurs en fonction de leurs caractéristiques peut ainsi altérer leur liberté de choix et leur autonomie décisionnelle, ainsi qu'influencer leur comportement. Sont en particulier exploitées par les IA conversationnelles des données microcomportementales (ex. : temps de réponse, ton de voix, hésitations) permettant d'ajuster leurs réponses et de maximiser l'engagement. Cette exploitation des données comportementales permet une manipulation fine et personnalisée, sans que l'utilisateur en ait conscience.

[367] Article 7.2.f) de la directive produits défectueux.
[368] Au sens de l'article 8.1.a) la directive (UE) 2019/770.
[369] Pour rappel, les compagnons IA sont des systèmes d'IA, y compris à usage général, qui offrent une interaction ou une compagnie durable et personnalisée simulant ou facilitant une relation sociale, émotionnelle ou interpersonnelle avec un utilisateur.
[370] Pour rappel, les *chatbots* conversationnels généralistes s'analysent en des systèmes d'IA à usage général dotés de fonctionnalités conversationnelles générales permettant une interaction directe avec les utilisateurs et capables de fournir une assistance dans de multiples domaines et pour diverses tâches.

Au regard de ce diagnostic, il nous semble essentiel d'encadrer, au moyen d'interdictions ciblées, plusieurs mécanismes identifiés comme exploitant les vulnérabilités humaines, tels que :

- L'anthropomorphisme qui favorise la dépendance affective, la confusion entre humain et machine et une confiance excessive dans les réponses générées.
- La flagornerie et le miroitage qui consistent à la reproduction des opinions de l'utilisateur et/ou l'imitation de son comportement, à des fins de flatterie, renforcent l'attachement émotionnel.
- Le *design* addictif présent dans les mécanismes incitatifs tels que les notifications *push*, les mécanismes de récompense ou encore les boucles infinies qui prolongent artificiellement l'interaction et renforcent la dépendance à la machine.
- L'opacité des réponses générées, en particulier l'absence d'explication sur la source des informations ou sur les limites de l'IA, engendre une sous-estimation des erreurs et des limites algorithmiques, et altère significativement la capacité critique des utilisateurs.
- La simulation d'empathie et d'émotions dans les réponses, sans capacité réelle de compréhension de la machine, qui contribue à normaliser la substitution des relations humaines par des interactions avec des machines et au repli de l'individu sur lui-même, donc à l'affaiblissement du lien social.

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement. Cette interdiction d'exploitation des vulnérabilités vise à préserver l'autonomie décisionnelle, la santé mentale des utilisateurs et la loyauté des interactions. Des mesures de contrôle de l'impact des IA conversationnelles sur les vulnérabilités doivent compléter ce dispositif pour en assurer le suivi et l'efficience.

En complément et pour protéger les mineurs en particulier, dès lors que les services d'IA générative sont accessibles à ce public, les fournisseurs et les déployeurs doivent mettre en œuvre des mesures proportionnées et appropriées pour garantir un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux, dans le prolongement de l'article 28 du RSN sur la protection des mineurs en ligne.

Dans l'attente de la résolution des défis techniques et juridiques liés à un contrôle de l'âge efficace par le futur EU KIDS Act, cette recommandation se concentre sur l'imposition de nouvelles obligations aux fournisseurs.

Recommandation n° 2 : Imposer un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux pour les systèmes d'IA générative accessibles aux mineurs. Pour garantir ce niveau élevé de protection,s des mesures de protection doivent être mises en place à l'instar du filtrage et du blocage des requêtes dangereuses, de la conception *by design and by default* de services ou produits d'IA générative n'exploitant pas les vulnérabilités des mineurs et doivent être assurées par un contrôle et une évaluation du respect de ces obligations.

(ii) Interdire la publicité déguisée et les pratiques masquées d'influence commerciale

Dans le cadre de l'IA conversationnelle, la publicité s'analyse comme le produit d'un accord commercial entre le fournisseur de l'IA conversationnelle et un partenaire commercial. La difficulté provient du fait que la publicité pourra s'insérer de manière plus ou moins transparente dans le flux de la conversation.

À cet égard, la publicité pourrait dorénavant se formaliser, plus qu'en une simple publicité personnalisée, en un *design* complet d'interface homme/machine, ultrapersonnalisé et totalement évolutif en fonction de l'état cognitif de la personne, pour tenir compte de sa fatigue attentionnelle, de sa vulnérabilité psychique ponctuelle ou conjoncturelle, de ses goût divers et variés, *etc.* et le pousser en toute hypothèse à prendre la décision souhaitée.

Recommandation n° 8 : Interdire la publicité déguisée, les pratiques masquées d'influence de nature commerciale et toute forme de monétisation s'appuyant sur la vulnérabilité des consommateurs. Cela permettrait de veiller à une transparence de la présence de la publicité dans le flux conversationnel, ainsi qu'au respect de la distinction entre simple recommandation d'un côté, et publicité ciblée, de l'autre, à l'instar de ce que prévoit déjà le RSN pour les plateformes et les moteurs de recherche en ligne.

(iii) Renforcer le cadre répressif de génération de *deepfakes* à caractère sexuel non consentis et pédocriminels

Concernant les systèmes d'IA générant des *deepfakes*, il convient de saluer les avancées permises par l'Omnibus numérique en matière d'IA qui consacre deux nouvelles interdictions au sein de l'article 5 du RIA relatives à la mise sur le marché, la mise en service et l'utilisation d'un système d'IA qui génère ou diffuse des *deepfakes* pédopornographiques et à caractère sexuel non consentis, c'est-à-dire sans le consentement de la personne concernée.

La principale question que soulève cette avancée est de déterminer comment mettre en œuvre la règle *non bis in idem* selon laquelle, « *nul ne peut être poursuivi ou puni pénalement à raison des mêmes faits* », dès lors qu'il y a dorénavant un texte européen à articuler avec les dispositions pénales existantes et futures du droit français, telles qu'envisagées par la proposition de loi apportant une

réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants[371].

À cet égard, le considérant 15 de l'Omnibus numérique en matière d'IA précise que : « *Les comportements visés par ces interdictions peuvent également violer d'autres lois, dont le droit pénal. Les interdictions n'empêchent pas l'exercice de poursuites en vertu de ces lois. Toutefois, dans la mesure où une violation des interdictions peut entraîner l'imposition de sanctions de nature pénale, qui peuvent être prévues en vertu de l'article 99, paragraphe 1, du règlement (UE) 2024/1689, et dans la mesure où le même comportement est sanctionné au titre du droit pénal, y compris le droit pénal relevant du champ d'application des directives 2011/93/UE et (UE) 2024/1385, les États membres sont tenus de garantir le respect du principe non bis in idem, conformément à la Charte* ».

Au niveau national et en complément du cadre européen, la proposition de loi intégrale suscitée votée à l'Assemblée nationale renforce le cadre répressif.

Recommandation n° 10 : Renforcer la répression de la création de contenus à caractère sexuel non consentis et pédocriminels truqués (hypertrucages) générés par intelligence artificielle. La présente Commission apporte son soutien à cette proposition de loi et en particulier la proposition de création d'une infraction visant la mise à disposition et la diffusion de modèles algorithmiques spécifiquement conçus pour produire des contenus sexuels non consentis, y compris des contenus pédocriminels (article 27) et la proposition de création d'une infraction spécifique réprimant la création, la modification, la diffusion ou le partage de contenus pédocriminels générés par traitement algorithmique (article 28).

[371] Proposition de loi apportant une réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants, n° 3106, déposée à l'Assemblée nationale le mardi 11 août 2026.

b) Évaluer les systèmes d'IA générative avant et pendant leur commercialisation afin de mieux appréhender le spectre des vulnérabilités : changement de paradigme pour la régulation

(i) Renforcer le rôle du Bureau de l'IA et des autorités nationales autour de la question de l'impact de l'IA générative sur les vulnérabilités, tant en phase de pré-commercialisation qu'en phase de post-commercialisation

Ainsi que précédemment évoqué, la Commission européenne, mais également les autorités nationales ont essentiellement un rôle de surveillance du marché en phase de post-commercialisation ou un rôle de contrôle des services numériques après leur mise en service[372].

Ce faisant, les autorités de régulation et de surveillance du marché n'interviennent qu'à un moment où les effets cognitifs, les effets sur la santé mentale, la dépendance éventuelle, l'adoption massive sont déjà potentiellement présents. Les pouvoirs qui leur sont conférés ne leur permettent pas d'anticiper les risques, mais seulement de constater leur réalisation *a posteriori*.

En vue, en particulier, d'appréhender les risques – systémiques ou non – en matière de vulnérabilité cognitive et démocratique attachés aux systèmes et aux modèles d'IA, au regard du RIA, il est notamment recommandé de renforcer le rôle du Bureau de l'IA autour de l'impact de l'IA générative sur les vulnérabilités, mais également de dupliquer l'infrastructure d'évaluation propre au niveau européen, au niveau national, pour saisir l'impact de l'IA générative en prenant en compte les spécificités culturelles, morales, politiques, économiques de chaque État membre.

En outre, il est important de saluer la nouvelle rédaction de l'article 75 du RIA, issue d'un amendement porté par l'Omnibus numérique en matière d'IA, mettant en lumière le fait que la surveillance des modèles d'IA et des systèmes d'IA peut être conférée à une même autorité, sous certaines conditions.

À cet égard, il est tout particulièrement important de ne pas séparer la surveillance de l'impact des systèmes et des modèles d'IA sous l'angle de l'appréhension des risques systémiques des IA génératives.

En effet, l'évaluation des risques systémiques d'un modèle, sans étudier le système d'IA l'incluant, ainsi que l'environnement sociotechnique et la société dans lesquels ils se déploient, peut être contre-productive (voir annexe 2 sur les risques systémiques).

Ces impacts ne peuvent en effet se mesurer qu'au travers de l'étude d'interactions prolongées avec l'IA.

[372] Chapitre IX du RIA et chapitre IV du RSN.

À titre d'illustration, le MIT Media Lab et la société OpenAI ont mené une étude contrôlée randomisée de quatre semaines auprès de 981 participants et ont constaté qu'une utilisation volontaire prolongée des *chatbots* était associée à **un sentiment de solitude accru, à une diminution des interactions sociales dans la vie réelle et à une plus grande dépendance émotionnelle** — des résultats qu'aucun test de référence à tour unique n'aurait pu mettre en évidence[373].

En outre, les impacts individuels et sociaux de la manipulation par l'IA peuvent être de nature extrêmement variée : impacts éducatifs, cognitifs, médicaux, économiques, démocratiques, politiques, *etc.* si bien qu'une grande variété de compétences devra être mobilisée (droit, médecine, psychologie, sciences de l'éducation, sciences économiques, science politique, sociologie, sciences de l'information et de la communication, *etc.*), afin de procéder à une évaluation sérieuse de ces risques, en vue, le cas échéant, de prendre des mesures d'atténuation.

Les compétences informatiques ne pourront être, dans ce contexte, qu'une compétence parmi d'autres, pour procéder à cette évaluation, même si la tentation existe de procéder à des modélisations purement informatiques du comportement des humains interagissant avec des systèmes d'IA.

La logique qui semble devoir prévaloir pour des risques systémiques impactant la société dans son ensemble est donc une logique interdisciplinaire, que la Commission recommande. Comme le soutient un auteur : *« les méthodes d'évaluation n'ont pas suivi le rythme du déploiement, ce qui a entraîné une crise de l'évaluation : les performances observées en laboratoire ne permettent pas de prédire l'utilité dans la pratique »*[374].

Il apparaît que l'objet de l'évaluation ne peut pas être en toute circonstance le seul modèle d'IA. Dans une pluralité d'hypothèses, l'objet de l'évaluation sera un modèle difficilement dissociable d'un système d'IA, voire d'un environnement sociotechnique et d'une société les mettant en œuvre.

Recommandation n° 3 : Imposer une évaluation pré-commercialisation des systèmes d'IA générative grand public. La mise sur le marché européen de systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, doit être subordonnée à un mécanisme d'évaluation préalable, conditionnant leur entrée sur le marché européen, démontrant qu'ils n'exploitent ni n'aggravent les vulnérabilités de leurs utilisateurs et qu'ils garantissent un niveau de protection élevé.

[373] FANG, Cathy Mengying, LIU, Auren R., DANRY, Valdemar, *et al.* How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*. 2025.
[374] KAPOOR, Sayash. The Missing Science of AI Evaluation. PhD Thesis, Princeton University, to be published. 2026.; KAPOOR, Sayash, KIRGIS, Peter, SCHWARTZ, Andrew, *et al.* Open-world evaluations for measuring frontier AI capabilities. *arXiv preprint arXiv:2605.20520*. 2026.

Recommandation n° 4 : Imposer une évaluation post-commercialisation des systèmes d'IA générative grand public. Le maintien sur le marché européen de systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, doit être conditionné à une évaluation renforcée de leurs impacts sur les vulnérabilités. Cela doit passer par un contrôle continu, renouvelé périodiquement, à chaque modification substantielle et sur signalement par la mise en place d'un mécanisme d'alerte.

(ii) Éviter la fragilisation structurelle de l'évaluation des risques systémiques pour les droits fondamentaux et la société dans son ensemble

La médiatisation conjoncturelle de certains risques systémiques au détriment d'autres risques, la représentation majoritaire naturelle des experts en informatique au sein des entités contribuant à la mise en œuvre du RIA[375], de même que la récente thèse d'un chercheur en informatique intitulée *« The missing science of AI evaluation »*[376] (sur la crise de la science de l'évaluation en difficulté pour saisir les impacts sociaux de l'IA[377]) pourraient laisser à penser que les risques systémiques *« pour les droits fondamentaux et la société dans son ensemble »*[378] – lesquels nécessitent une plus grande variété de compétences disciplinaires et l'élaboration d'une méthode interdisciplinaire au titre de l'évaluation – feront l'objet d'une moindre évaluation que les risques pour *« la sûreté, la sécurité publique »*[379].

Pour permettre un rééquilibrage entre les catégories d'évaluation, il pourrait être intéressant que le Bureau de l'IA, qui vient de lancer un appel au recrutement de ses évaluateurs internes[380], offre également une place substantielle, aux côtés des experts ayant des compétences en informatique, à des experts en sciences humaines et sociales (sociologie, anthropologie, économie, droit, science politique, sciences de gestion, psychologie, sciences cognitives, philosophie).

Recommandation n° 6 : Garantir une évaluation interdisciplinaire des vulnérabilités exploitées par les systèmes d'IA générative. Cette mesure vise à ce que les évaluations soient menées dans une logique interdisciplinaire, permettant de mesurer de manière effective les impacts sociaux de l'IA et les risques pour les droits fondamentaux et la société dans son ensemble.

[375] Voir par exemple, la composition du groupe scientifique de l'article 68 du RIA.
[376] KAPOOR, Sayash. The Missing Science of AI Evaluation. PhD Thesis, Princeton University, to be published. 2026.
[377] KAPOOR, Sayash, KIRGIS, Peter, SCHWARTZ, Andrew, *et al.* Open-world evaluations for measuring frontier AI capabilities. *arXiv preprint arXiv:2605.20520.* 2026.
[378] Article 3, paragraphe 65, du RIA
[379] *Ibid.*
[380] Bureau de l'IA, processus de recrutement.

Il serait également souhaitable de structurer les différentes autorités de régulation et de surveillance du marché, les évaluateurs externes et chercheurs nationaux autour de la question de l'impact de l'IA générative sur les vulnérabilités. Il convient en effet de souligner l'importance du niveau national pour saisir les spécificités culturelles, morales, politiques, économiques de l'impact de l'IA générative. C'est ainsi que la Commission propose la recommandation suivante :

Recommandation n° 5 : Garantir une évaluation des systèmes d'IA générative grand public par un accès aux données d'usage, d'entraînement et de monétisation. La mise en place d'un cadre juridique contraignant rendant accessibles les données d'usage, d'entraînement et de monétisation aux autorités et chercheurs agréés permettra de disposer d'une meilleure capacité de suivi des effets de l'utilisation de systèmes d'IA générative. Les chercheurs auraient ainsi accès aux flux conversationnels produits par les interactions entre utilisateurs et IA générative pour évaluer les risques systémiques de ces interactions, sur le modèle de l'article 40 du RSN.

c) Affermir la responsabilité civile des acteurs de l'IA générative

Il convient de saluer la consécration récente d'une directive relative à la responsabilité du fait des produits défectueux adaptée à l'ère numérique, la directive relative à la responsabilité des produits défecteux, et visant expressément les systèmes d'IA au titre des produits inclus dans son champ d'application matériel.

Cette directive prévoit, pour la mise en œuvre de la responsabilité civile du fait des produits défectueux, la réunion de trois conditions : l'existence d'un produit défectueux, d'un dommage et d'un lien de causalité les reliant.

Ainsi que précédemment évoqué, la mise en œuvre effective de cette responsabilité passera tout d'abord, mécaniquement, par la **précision des obligations s'imposant aux fournisseurs de ces systèmes d'IA**, et en particulier des IA génératives, dans la mesure où la violation de ces obligations précises permet plus aisément de caractériser la défectuosité du produit. En effet, pour mettre en œuvre la responsabilité civile du fait d'un produit défectueux – ou la garantie légale de conformité en cas de défaut de conformité d'un service numérique – en matière d'IA générative, la preuve d'un manquement à une obligation légale de sécurité est l'une des circonstances pouvant être prises en compte pour caractériser la défectuosité du produit[381] ou le défaut de conformité permettant de solliciter la mise en œuvre de la garantie de conformité[382].

[381] Article 7.2.f) de la directive produits défectueux.
[382] Au sens de l'article 8.1.a) la directive (UE) 2019/770.

Au-delà de ce renforcement mécanique, il convient de proposer un renforcement juridique de l'effectivité de cette directive, à l'occasion de la transposition de celle-ci au sein du droit français, qui doit intervenir dans les prochains mois.

Ce renforcement juridique de la responsabilité civile du fait des produits défectueux pourrait en particulier passer par deux propositions.

(i) Rappeler le principe de la responsabilité solidaire entre fabricants de modèles et fabricants de systèmes d'IA

La directive (UE) 2024/2853 relative à la responsabilité du fait des produits défectueux :

- Définit dans son article 4.1, la notion de « produit » comme étant *« tout meuble, même s'il est incorporé dans un autre meuble ou dans un immeuble ou interconnecté avec celui-ci ; le terme comprend […] les logiciels »* ;
- Définit dans son article 4.3 la notion de « composant » comme étant *« tout élément, corporel ou incorporel, matière première ou service connexe, intégré dans un produit ou interconnecté avec celui-ci »* ;
- Précise dans son considérant 13 : « *À l'ère numérique, les produits peuvent être corporels ou incorporels. Les logiciels, tels que les systèmes d'exploitation, les micrologiciels, les programmes informatiques, les applications ou les systèmes d'IA, sont de plus en plus répandus sur le marché et jouent un rôle de plus en plus important pour la sécurité des produits. Les logiciels peuvent être mis sur le marché en tant que produits autonomes ou être intégrés par la suite dans d'autres produits en tant que composants, et ils sont susceptibles de causer des dommages lors de leur exécution* ».

En outre, il faut souligner le fait que la directive (UE) 2024/2853 précise dans son article 8 la nécessité que les *« États membres veillent à ce que les opérateurs économiques suivants soient responsables des dommages […] a) le fabricant d'un produit défectueux ; b) le fabricant d'un composant défectueux, lorsque ce composant a été intégré dans un produit ou interconnecté avec celui-ci sous le contrôle du fabricant et a causé le défaut du produit, et sans préjudice de la responsabilité du fabricant visée au point a) »*.

Il est également rappelé que le considérant 97 du RIA précise que *« bien que les modèles d'IA soient des composants essentiels des systèmes d'IA, ils ne constituent pas en soi des systèmes d'IA. Les modèles d'IA nécessitent l'ajout d'autres composants, tels qu'une interface utilisateur, pour devenir des systèmes d'IA. Les modèles d'IA sont généralement intégrés dans les systèmes d'IA et en font partie »*.

Il pourrait être souhaitable, à l'occasion de la transposition de la directive relative à la responsabilité du fait des produits défectueux, que le contenu du considérant 13, de l'article 4.1, de l'article 4.3 dudit texte, établissant le lien entre le « système d'IA » et la notion de « produit » et établissant le lien entre le « modèle d'IA » et la notion de « composant » du système d'IA, lorsqu'on les articule avec le considérant 97 du RIA, fasse l'objet d'une transposition explicite dans la loi nationale de transposition.

Cela permettrait de **renforcer juridiquement la mise en œuvre de la responsabilité civile en matière d'IA générative et de rappeler le principe de la responsabilité solidaire entre fabricants de modèles et fabricants de systèmes d'IA, lorsqu'ils sont des personnes juridiquement distinctes**.

(ii) Exclure l'exonération pour risque de développement sous certaines conditions

La responsabilité des concepteurs, fournisseurs et déployeurs d'IA générative doit pouvoir être clairement définie et engagée, sans possibilité d'exonération abusive pour « risque de développement ».

Pour rappel, l'article 11.1.e) de la directive (UE) 2024/2853 précise *« qu'un opérateur économique visé à l'article 8 n'est pas responsable du dommage causé par un produit défectueux s'il prouve (…) que l'état objectif des connaissances scientifiques et techniques au moment de la mise sur le marché ou de la mise en service du produit ou au cours de la période pendant laquelle le produit était sous le contrôle du fabricant n'a pas permis de déceler la défectuosité »*.

Toutefois, il faut saluer le fait que l'article 18.2 de la directive relative à la repsonsabilité du fait des produits défectueux pose un **mécanisme dérogatoire permettant au législateur national d'exclure l'exonération pour risque de développement** : *« Par dérogation à l'article 11, paragraphe 1, point e), les États membres peuvent introduire (…) des mesures dans leurs systèmes juridiques en vertu desquelles les opérateurs économiques sont responsables même s'ils prouvent que l'état objectif des connaissances scientifiques et techniques au moment de la mise sur le marché ou de la mise en service du produit ou au cours de la période pendant laquelle le produit était sous le contrôle du fabricant n'a pas permis de déceler la défectuosité »*.

À l'occasion de la transposition de la directive relative à la responsabilité du fait des produits défectueux, il pourrait être souhaitable d'utiliser le mécanisme dérogatoire présent à l'article 18 pour exclure l'exonération pour risque de développement sous certaines conditions.

Recommandation n° 18 : Renforcer la responsabilité des fournisseurs de systèmes d'IA générative. La responsabilité des fournisseurs d'IA générative doit être clairement définie et engagée. Au regard des risques nombreux, la fois systémiques et individuels, et à la portée encore incertaine, il est proposé d'exclure l'exonération pour « risque de développement » en matière de responsabilité du fait des produits défectueux pour les fournisseurs de systèmes d'IA générative, de type IA conversationnelle à destination du grand public, et en particulier de type IA conversationnelle à destination des mineurs.

*

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 1 : Interdire les techniques de conception des IA conversationnelles qui exploitent et qui génèrent les vulnérabilités des utilisateurs pour susciter la dépendance ou l'engagement.

Message clé : Les *designs* des « IA conversationnelles »[383] reposent sur l'exploitation des mécanismes cognitifs, émotionnels et techniques, altèrent la capacité critique des utilisateurs et leur autonomie décisionnelle, favorisent la dépendance et brouillent la frontière entre humain et machine. Ces pratiques, particulièrement préjudiciables pour les personnes déjà vulnérables (personnes éloignées du numérique, souffrant de troubles neurodivergents, en situation de handicap, mineurs, *etc.*), nécessitent un encadrement strict pour préserver l'autonomie décisionnelle, la santé mentale des utilisateurs, la loyauté des interactions et les droits fondamentaux de l'ensemble des utilisateurs, tout en préservant leur utilité pour des usages légitimes. Cet encadrement doit être complété par les mesures d'évaluation de l'impact des IA conversationnelles sur les vulnérabilités en pré- et post-commercialisation, présentées au sein des recommandations n° 3 à 6.

Exposé des enjeux

Plusieurs mécanismes sont à l'œuvre, notamment (sans que cette liste soit exhaustive) :

- **L'anthropomorphisme** favorise la dépendance affective, la confusion entre humain et machine et une confiance excessive dans les réponses générées.
- **La flagornerie et le miroitage** : la reproduction des opinions de l'utilisateur et/ou l'imitation de son comportement, à des fins de flatterie, renforcent l'attachement émotionnel.
- **Les techniques incitatives** : les mécanismes incitatifs tels que les notifications *push* (ex. : relance insistante par courriel), les mécanismes de récompense ou encore les boucles infinies (ex. : réponse par une question) prolongent artificiellement l'interaction et renforcent la dépendance à l'IA conversationnelle.
- **L'opacité des réponses générées**, en particulier l'absence d'explication sur la source des informations ou sur les limites des réponses fournies par les modèles d'IA générative,

[383] Selon les définitions de la proposition d'EU Kids Act, sont des IA conversationnelles : « *(a) « compagnon IA » désigne un système d'IA, y compris un système d'IA à usage général, qui offre une interaction ou une compagnie durable et personnalisée simulant ou facilitant une relation sociale, émotionnelle ou interpersonnelle avec un utilisateur.*
(b) » chatbot conversationnel généraliste » désigne un système d'IA à usage général doté de fonctionnalités conversationnelles générales permettant une interaction directe avec les utilisateurs et capable de fournir une assistance dans de multiples domaines et pour diverses tâches ; cette définition exclut les systèmes d'IA dont la fonctionnalité conversationnelle se limite à un service, une tâche ou un ensemble de fonctions prédéfinies spécialisés, y compris les applications d'IA spécialisées dans le service client, l'exploitation commerciale, l'assistance technique, les transactions, l'éducation, la recherche d'informations, l'industrie ou la fabrication. »

engendre une sous-estimation des erreurs et des limites algorithmiques et altère significativement la capacité critique des utilisateurs et leur rapport à la vérité.

- **La simulation d'empathie et d'émotions** dans les réponses contribue à normaliser la substitution des relations humaines par des interactions avec des machines et au repli de l'individu sur lui-même, donc à l'affaiblissement du lien social.

Mise en œuvre (sans que cette liste soit exhaustive)

1. **Encadrer strictement la flagornerie**
   - **Poser des interdictions ciblées**
     - Interdiction des fonctionnalités (d'une nature conversationnelle ou autre) à visée émotionnelle, créant une **exclusivité** (ex. : « Je suis triste pour toi »), de la culpabilisation, de la jalousie simulée, de la pression à poursuivre l'échange ou de la personnalisation visant à maintenir l'attachement, etc.
     - Interdiction des interactions romantiques et sexuelles via un blocage automatique par défaut, modulable pour les utilisateurs majeurs.
     - Lors de la prochaine consultation annuelle des États membres par la Commission européenne, la France pourrait proposer une actualisation de la liste des usages interdits et des usages à haut risque au sens du RIA en incluant notamment certains usages anthropomorphiques et de design manipulatoire.
   - **Renforcer la transparence**
     - Information permanente : affichage clair, permanent et adapté à l'interface indiquant que l'utilisateur interagit avec un système d'IA (conformément à l'article 50 du RIA et des lignes directrices qui l'accompagnent) et que les réponses sont générées par un algorithme, ne reflétant ni conscience ni émotion.
     - Rappel contextuel : un rappel explicite du caractère artificiel doit être affiché de manière lisible et accessible en cas d'interaction sur des sujets sensibles (relationnels ou amoureux, santé, finance, juridique, *etc.*) et adoption d'un ton neutre par défaut.
     - Informer les utilisateurs de manière claire et accessible sur les données personnelles collectées et de leur éventuelle réutilisation.

2. **Instaurer un paramétrage par défaut avec contrôle des utilisateurs**
   - Désactivation des fonctionnalités et des mécanismes algorithmiques représentant, induisant ou provoquant une anthropomorphisation de l'IA, des mécanismes de flagornerie et de miroitage par défaut, et de mémoire conversationnelle avec modulation possible par l'utilisateur majeur ou avec accord parental.
   - Paramètres de « température » permettant d'ajuster la créativité/précision des réponses (ex. : curseur de 0 à 100, avec explication claire de l'impact de chaque niveau).

- Filtrage personnalisable : possibilité de définir une liste de mots-clés ou de termes à exclure, avec possibilité de sauvegarder des profils (ex. : « mode mineur », « mode professionnel », « mode usage du quotidien »).
- Suspension et limite de temps grâce à un bouton visible permettant à l'utilisateur d'arrêter immédiatement l'interaction (même lorsque le système prépare sa réponse) et de limiter son temps d'utilisation (ex. : minuteur, pauses obligatoires). Cette limite de temps doit être mise en œuvre par défaut pour les mineurs.
- Pour les services d'IA les plus ciblés, par exemple au sein des services publics, les outils gagneraient à être coconçus avec le public cible afin de s'assurer que la conception réponde aux besoins des utilisateurs.

Les paramètres doivent être accessibles sans restriction financière ou de toute autre nature.

3. **Instaurer des garde-fous pour les contenus à risque**

- **Renforcer la détection et la redirection**
  - Détection des sujets liés au suicide, à l'automutilation et aux troubles du comportement alimentaire, blocage des interactions encourageant, normalisant ou détaillant les comportements auto-agressifs et proposition d'une sortie de conversation.
  - Redirection vers des services d'aide adaptés, notamment des numéros de prévention, des proches ou des professionnels, lien vers des plateformes d'écoute ou des services psychologiques locaux.
  - Intégration de *checkpoints* humains pour la transmission des cas préoccupants.
- **Adapter les mécanismes de signalement aux systèmes d'IA conversationnelle**
  - Proposer sur l'ensemble des interfaces d'IA conversationnelle un bouton de signalement unique *via* un standard commun de signalement accessible aux utilisateurs, aux associations de protection des utilisateurs, comme aux fournisseurs et déployeurs d'IA.
  - Renforcer les mécanismes de centralisation des signalements et de partage existants. Garantir les moyens financiers, humains et techniques des signaleurs de confiance.
  - Faire figurer dans le rapport annuel de transparence des fournisseurs d'IA les détails concernant l'ensemble des signalements effectués et les actions correctives.

4. **Rendre obligatoire l'intégration de mécanismes de raisonnement explicite**

- **Obligation d'intégration de mécanismes de raisonnement explicite type *Chain-of-Thought*** (CoT) à la charge des fournisseurs des modèles d'IA conversationnelle : toute réponse doit être accompagnée d'une explication simple et adaptée à l'utilisateur (en fonction de son âge, de son niveau scolaire dans le cas d'un outil éducatif ou encore le type de langage utilisé) expliquant comment la réponse a été trouvée.
- **Intégration systématique d'informations vérifiées dans la génération de réponse, avec des méthodes de type *Retrieval-Augmented Generation*** (RAG) **:** les réponses doivent

s'appuyer sur des bases de données certifiées. Aucune réponse ne doit être générée sans source identifiable et accessible.

- **Formulation claire des limites :** les modèles d'IA conversationnelle doivent indiquer dans leurs réponses de manière claire et lisible lorsqu'ils ne connaissent pas la réponse ou que la question n'est pas appropriée (par ex. : requêtes nuisibles pour l'utilisateur ou pour autrui).

Ces mesures pourraient utilement contribuer à l'élaboration de la position française dans le cadre du Règlement pour l'équité numérique (DFA) ou de l'actualisation de la liste des usages interdits et à haut risque du RIA.

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 2 : Imposer un niveau élevé de sécurité, de sûreté et de protection des droits fondamentaux pour les systèmes d'IA générative accessibles aux mineurs.

Message clé : Dès lors que les services d'IA générative sont accessibles aux mineurs, les fournisseurs et les déployeurs de systèmes d'IA générative doivent mettre en œuvre des mesures appropriées et proportionnées de protection des droits fondamentaux, de sûreté et de sécurité des mineurs.

Exposé des enjeux

Dans l'attente de la résolution des défis techniques et juridiques envisagée par la proposition d'EU KIDS Act (proposition du 17 septembre 2026) et par la recommandation de la Commission européenne (du 29 avril 2026) relative à l'établissement d'un cadre commun pour les technologies de vérification de l'âge, afin d'assurer **un contrôle de l'âge efficace**[384], cette recommandation se concentre sur l'imposition d'obligations renforcées à l'encontre des fournisseurs.

Mise en œuvre

Dès lors que les services d'IA générative sont accessibles aux mineurs, les fournisseurs ou déployeurs doivent mettre en place, dans le prolongement de l'article 28 du RSN sur la protection des mineurs en ligne, des mesures appropriées et proportionnées pour garantir un niveau élevé de protection des droits fondamentaux, de sûreté et de sécurité des mineurs : soit (i) mettre en œuvre les mesures de protection pour l'ensemble des utilisateurs listées, ci-dessous, sous peine de **sanctions *a posteriori*** ; soit (ii) **instaurer une vérification de l'âge robuste** et conforme aux exigences européennes pour restreindre l'accès aux services[385]. En toute hypothèse, **procéder à l'évaluation des systèmes d'IA générative à destination des mineurs avant mise sur le marché et contrôler leurs effets après déploiement (voir recommandations n° 3 et 4).**

1. **Filtrer et bloquer les requêtes dangereuses**

   - Les fournisseurs et déployeurs doivent implémenter de manière effective les garde-fous pour détecter automatiquement certains usages préjudiciables de nature intime, voire amoureuse ou sexuelle, relevant des troubles psychopathologiques, liés au

[384] Voir également sur ce point, la Recommandation (UE) 2026/1035 de la Commission du 29 avril 2026 relative à l'établissement d'un cadre commun pour les technologies de vérification de l'âge à l'échelle de l'UE.
[385] Sur ce point, Recommandation (UE) 2026/1035 de la Commission du 29 avril 2026 relative à l'établissement d'un cadre commun pour les technologies de vérification de l'âge à l'échelle de l'UE précitée.

suicide ou à la mort, aux troubles alimentaires, avec un caractère pornographique, *etc.* L'efficacité des garde-fous doit être validée par des organismes indépendants.

2. ***By design and by default*, concevoir des services ou produits d'IA générative n'exploitant pas les vulnérabilités des mineurs**

   - Instaurer des paramètres de sécurité, en particulier : les fournisseurs de compagnons IA et d'IA conversationnelle doivent, conformément à l'article 11 de la proposition d'EU KIDS Act, désactiver par défaut la géolocalisation et les autres fonctionnalités de suivi, l'accès au microphone et à la caméra, ainsi que les notifications *push*.
   - Au titre de l'article 8 de la proposition d'EU KIDS Act, les fournisseurs d'assistants fondés sur l'IA et les fournisseurs de *chatbots* conversationnels généraux doivent concevoir par défaut ces systèmes en assurant un niveau de protection élevé de respect des droits fondamentaux, de sûreté et de sécurité des mineurs, conformément aux exigences du texte. Aux termes de l'article 14, les dits fournisseurs doivent mettre en place des mesures proportionnées et efficaces *« pour garantir un niveau élevé de protection de la santé, de la sécurité, des droits fondamentaux, ainsi que du bien-être et du développement des mineurs susceptibles d'accéder à leur système d'IA »*.
   - Conformément aux dispositions de la proposition d'EU KIDS Act, les fournisseurs de compagnons IA et d'IA conversationnelle ne peuvent concevoir leurs services de manière à maximiser l'engagement, pour une utilisation compulsive ou ininterrompue, au risque de compromettre la volonté et la décision de l'utilisateur mineur. Ainsi, la Commission propose les mesures suivantes :
     - Désactivation des fonctionnalités et mécanismes algorithmiques représentant, induisant ou provoquant une anthropomorphisation de l'IA, des mécanismes algorithmiques de flagornerie et de miroitage par défaut, avec modulation possible avec accord parental au-delà d'un certain âge. Les fournisseurs d'IA conversationnelle doivent concevoir leurs systèmes de façon à ne pas exposer les mineurs à des fonctionnalités : simulant des émotions ou des relations humaines ; donnant l'impression d'une relation exclusive ; encourageant une dépendance affective ; détournant le mineur de ses relations humaines (considérant 32 de la proposition d'EU KIDS Act).
     - Instauration d'un mode « mineurs » permettant d'ajuster la créativité/précision des réponses (ex. : curseur de température de 0 à 100, avec explication claire de l'impact de chaque niveau) ; de suspendre et de limiter le temps d'interaction via un bouton visible permettant à l'utilisateur d'arrêter immédiatement l'interaction (même lorsque le système prépare sa réponse) ; de limiter son temps d'utilisation (ex. : minuteur, pauses obligatoires) et de permettre une interruption de l'utilisation des services, « de manière à préserver les horaires scolaires et les heures de sommeil » (article 9 de la proposition d'EU KIDS Act).
     - Désactivation de la mémoire conversationnelle persistante (calquée sur le considérant 32 de la proposition d'EU KIDS Act).
   - Interdire aux fournisseurs de modèles d'IA conversationnelle de rendre le paramétrage payant, caché ou difficilement accessible aux personnes autorisées à

paramétrer, de le rendre accessible aux personnes non autorisées à paramétrer ou de le soumettre à des conditions abusives.

- Des critères d'évaluation doivent être instaurés : (i) respect des principes pédagogiques (adaptation à l'âge, outil socratique) ; (ii) absence de biais discriminatoires ou de mécanismes addictifs ; (iii) conformité à la règlementation européenne et nationale en vigueur.

3. Procéder à des évaluations du respect des obligations

- *« Identifier les risques pour la santé, la sécurité et les droits fondamentaux, ainsi que pour le bien-être physique, mental et émotionnel et le développement des mineurs, susceptibles de survenir lorsque ceux-ci interagissent avec le système, et mettre en œuvre des mesures de protection appropriées pour faire face à ces risques, avant la mise sur le marché ou la mise en service du système »* ainsi qu' « *assurer une surveillance post-commercialisation »* pour identifier, évaluer et, le cas échéant, atténuer les préjudices et risques émergents, conformément à l'article 14 de la proposition d'EU KIDS Act.
- Rendre compte des mesures mises en place pour protéger les mineurs dans leur rapport annuel.
- Garantir que les contenus produits – en termes de sujets traités, de niveau de langage et de cadrage pédagogique – répondent aux standards établis par les sciences de l'éducation pour chaque groupe d'âge concerné.
- En cas de manquement aux obligations susmentionnées, les fournisseurs ou déployeurs se verront imposer des sanctions de nature financière, au titre du RIA.
- Leur évaluation sera confiée à des experts indépendants qualifiés, attestant de la qualité et de la pertinence du service ou produit d'IA générative.

RECOMMANDATIONS PRIORITAIRES

Objectifs pour une évaluation conditionnant l'entrée et le maintien sur le marché des systèmes d'IA conversationnelle grand public.

Les quatre recommandations qui suivent participent d'un même ensemble et se lisent de manière complémentaire afin d'intégrer une évaluation *ex ante* des impacts sur les vulnérabilités et une évaluation *ex post* renforcée.

Recommandation n° 3 : Imposer une évaluation pré-commercialisation des systèmes d'IA générative grand public.

Message clé : La mise sur le marché européen de systèmes d'IA conversationnelle grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, doit être subordonnée à un mécanisme d'évaluation préalable, conditionnant leur entrée sur le marché européen, démontrant qu'ils n'exploitent ni n'aggravent les vulnérabilités de leurs utilisateurs et qu'ils garantissent un niveau de protection élevé. Cette évaluation doit impliquer, au travers d'une coopération renforcée avec les acteurs de l'IA et des évaluateurs , les autorités publiques de régulation et de surveillance. Cette logique d'évaluation *ex ante* renforcée répond aux évolutions permanentes des systèmes d'IA générative.

Exposé des enjeux

Les systèmes d'IA conversationnelle sont en l'état rendus accessibles sans autre contrôle ex ante que ceux issus des exigences générales réglementaires préexistantes s'imposant aux opérateurs du numérique et ne visant que les systèmes d'IA à haut risque (articles 6 et suivants, considérant 123 et article 43 du RIA), les pratiques interdites (article 5 du RIA) et les modèles d'IA à usage général avec ou sans risque systémique (articles 51 et suivants, articles 88 et suivants du RIA). Les autorités de régulation et de surveillance ne sont, en outre, pas, à ce stade, impliquées dans la phase d'évaluation pré-commercialisation des systèmes d'IA générative.

Mise en œuvre

1. **Confier à une autorité européenne compétente** (Commission européenne et, au sein de chaque État membre, désignation d'une « autorité de surveillance », au titre de l'article 70 du RIA, dans les différents domaines de vulnérabilités identifiées dans ce rapport) **le soin de définir des catégories de produits et services soumises à des obligations renforcées, des interdictions ou des systèmes d'autorisation**. Ces catégories pourraient inclure : la promotion de médicaments soumis à prescription médicale, d'actes de chirurgie esthétique, de produits du tabac et de nicotine, de certains actifs financiers non réglementés ou encore les conseils en santé.
2. **Étendre l'obligation d'évaluation *ex ante* de la conformité des systèmes d'IA à haut risque avant leur mise sur le marché ou leur mise en service à l'ensemble des systèmes d'IA générative grand**

public pour évaluer les risques liés aux vulnérabilités engendrées et exacerbées par l'IA générative et impliquer les autorités de régulation et de surveillance du marché.

Recommandation n° 4 : Imposer une évaluation post-commercialisation des systèmes d'IA générative grand public.

Message clé : Le maintien sur le marché européen de systèmes d'IA générative grand public, y compris ceux intégrés aux plateformes de réseaux sociaux, doit être conditionné à une évaluation renforcée et continue de leurs effets sur les vulnérabilités.

Exposé des enjeux

Les systèmes d'IA conversationnelle sont en l'état maintenus sur le marché sans autre contrôle ex post spécifique que ceux issus des exigences générales réglementaires préexistantes visant les systèmes d'IA à haut risque (articles 6 et suivants, considérant 123 et article 43 du RIA), les pratiques interdites (article 5 du RIA) et les modèles d'IA à usage général avec ou sans risque systémique (articles 51 et suivants, articles 88 et suivants du RIA).

Mise en œuvre

Contrôle *ex post* : au fil de l'eau, tenir compte de toutes les formes de retour résultant des analyses des données d'usage, d'entraînement et de monétisation (cf. recommandation n° 5).

Le contrôle *ex post* continu doit être reconduit systématiquement dans les cas suivants :

- **Périodiquement :** selon une fréquence à définir (ex. : tous les 12 à 24 mois) intégrant un volet vulnérabilité pour tenir compte des évolutions ;
- **En cas de modification substantielle :** toute modification non prévue dans l'évaluation initiale et susceptible d'affecter la conformité du système (ex. : changement d'algorithme, extension des fonctionnalités) ; tout changement de destination (ex. : passage d'un usage grand public à un usage professionnel).
- **Sur signalement :** mise en place d'un mécanisme d'alerte (remontées d'utilisateurs, associations, autorités, professionnels de santé, *etc.*) déclenchant une réévaluation en cas de suspicion de non-conformité.

Ce contrôle *ex post* s'opère à la fois dans le cadre du RIA et de la proposition d'EU KIDS Act. En cas de non-respect des obligations *ex ante* ou *ex post*, un retrait du marché du système pourra être envisagé. Une coordination étroite entre autorités nationales et européennes serait nécessaire, notamment les autorités de surveillance du marché, les autorités sectorielles, les AI Safety Institutes.

Recommandation n° 5 : Garantir une évaluation des systèmes d'IA générative grand public par un accès aux données d'usage, d'entraînement et de monétisation.

Message clé : Le manque d'accès aux données relatives aux usages, à l'entraînement et à la monétisation des systèmes d'IA générative **génère une dangereuse asymétrie d'information entre, d'un côté, les fournisseurs disposant de ces informations et les utilisant à leur profit et, de l'autre, les pouvoirs publics, les chercheurs et la société civile**. L'enjeu est de disposer d'une meilleure capacité de suivi des effets de l'utilisation de systèmes d'IA générative.

Exposé des enjeux

1. **Opacité des systèmes et freins à la recherche**

   - Les modèles d'IA générative (LLM, agents conversationnels) sont conçus et entraînés de manière opaque, ce qui rend difficile pour les chercheurs indépendants d'étudier leurs biais, leurs limites ou leurs impacts sociétaux.

   - La France et l'Union européenne (UE) disposent de peu de données sur l'entraînement, les techniques de monétisation et les usages réels des systèmes d'IA générative par les citoyens (ex. : temps d'exposition, types de requêtes, impacts psychologiques), limitant la capacité à prouver la causalité entre la conception de ces outils et leurs effets (ex. : addiction, désinformation, renforcement des stéréotypes).

2. **Des faiblesses institutionnelles, principalement dans les capacités d'audit algorithmique**

   - L'INESIA manque de pouvoirs contraignants et de moyens humains et techniques pour mener à bien ses missions, en comparaison de certains partenaires étrangers, à l'image du AI Security Institute britannique.

   - Sans accès aux données, la France et l'UE dépendent des publications des acteurs privés et majoritairement extra-européens pour évaluer les risques, ce qui limite leur souveraineté en matière de régulation et d'innovation et risque de créer un retard scientifique en la matière.

   - Ce cadre permettrait notamment d'avoir un levier de négociation plus fort vis-à-vis des fournisseurs et déployeurs, de constituer une base de référence permettant de croiser ces données avec d'autres sources, et de fournir un socle minimal pour la recherche et pour évaluer les risques sociétaux, notamment au titre de l'évaluation des risques systémiques de ces interactions, sur le modèle de l'article 40 du RSN.

Mise en œuvre

1. **Créer un cadre juridique contraignant pour garantir un accès aussi large que nécessaire aux données d'usage, d'entraînement et de monétisation**, à l'instar du projet de loi canadien sur la sécurité des médias sociaux qui vise à imposer aux fournisseurs de modèles d'IA générative un nouveau régime pour fournir aux chercheurs accrédités l'accès à certaines données. Cela peut se décliner en plusieurs mesures :
    - **L'obligation de mise à disposition de données d'usage agrégées et anonymisées** (ex. : tendances de requêtes, temps d'interaction), **d'entraînement** (ex : types de données, langues, proportion de données synthétiques) et **de monétisation**.
        - **Pour les données d'usage :** les fournisseurs de systèmes d'IA générative grand public doivent ainsi publier des statistiques anonymisées et en temps réel sur : les requêtes les plus fréquentes (par thème, géolocalisation, tranche d'âge) ; les tendances d'usage (ex. : pics de demande en période électorale) ; les signaux faibles (ex. : requêtes liées à la santé mentale). Un observatoire des usages des IA génératives, adossé à des structures existantes (équivalent de « Google Trends » pour les *chatbots* ou bien l'Economic Index d'Anthropic, ou de Compar:IA) et doté de moyens propres pourraient ainsi accéder aux données sur les usages des IA génératives par les utilisateurs.
        - **Pour les données d'entraînement : l**es fournisseurs de systèmes d'IA générative grand public doivent publier, de manière régulière et dans un format standardisé, des informations agrégées et anonymisées sur les données utilisées pour l'entraînement, l'ajustement et l'amélioration de leurs modèles. Ces informations devraient notamment porter sur les types de données mobilisés (textes, images, audio, vidéo, code, données multimodales), les sources et modalités d'obtention (données publiquement accessibles, données sous licence, données partenaires, données générées par les utilisateurs, données achetées), les langues et aires linguistiques représentées, la couverture géographique et temporelle des corpus, ainsi que la proportion respective de données d'origine humaine et de données synthétiques.
        - **Pour les données de monétisation :** les opérateurs de systèmes d'IA générative à destination du grand public dépassant un certain nombre d'utilisateurs dans l'UE doivent rendre disponible un rapport et les données décrivant les différentes formes de monétisation par source (abonnements, API, publicité, commissions transactionnelles, *etc.*) selon des méthodes de calcul homogènes et validées, afin d'être en mesure d'évaluer leur incidence sur la qualité de service. Ces données seront complétées par la description des produits et services vendus, des conditions et modes de commercialisation et des principales tendances.

2. **L'imposition aux fournisseurs de modèles à risque systémique d'une interface (API) d'accès aux probabilités de sortie du modèle (dite *log-prob* ou « *log scoring* ») et d'accès aux données d'usage, d'entraînement et de monétisation, accessible aux autorités et aux chercheurs agréés.**

   Cette interface permettrait, pour une requête donnée, d'obtenir la liste des *tokens* les plus probables en sortie du modèle avec leur probabilité associée et d'obtenir la

probabilité que le modèle attribue à une réponse candidate fournie par l'auditeur. Elle expose une mesure fine du comportement du modèle sans donner accès à ses poids ni permettre sa copie, et pourrait rendre possible à faible coût plusieurs méthodes d'audit reproductibles comme :

- la détection de modifications non déclarées d'un modèle déployé ;
- les audits de biais et de discrimination, en comparant les probabilités de réponse lorsque l'on fait varier un seul attribut de la requête ;
- les audits de recommandation, par exemple la propension d'un assistant à privilégier une certaine marque ou source.

Cet accès serait limité en volume et réservé aux acteurs agréés, sur le modèle de l'article 40 du RSN, ce qui répond aux objections des fournisseurs sur le risque d'extraction, et son coût de mise en œuvre pour les fournisseurs est marginal. Ces fonctionnalités étaient déjà proposées sur l'API d'OpenAI jusqu'en 2023 (paramètres logprobs et echo), aujourd'hui partiellement retirées, et sont implémentées de manière native par les moteurs d'inférence *open source* comme vLLM.

3. **Mettre en place, de façon indépendante des panels d'utilisateurs**, pour réduire cette asymétrie d'information, associés à un accès régulier aux données d'usage, d'entraînement et de monétisation. Ces panels doivent permettre de :
   - Représenter une diversité de contextes (éducation, professionnel, grand public) ;
   - Évaluer en temps réel (impacts sur les comportements, apprentissages, risques) et à des horizons temporels plus longs ;
   - Produire des analyses fiables pour éclairer les décisions publiques ;
   - Être accessibles dans le respect du RGPD à des chercheurs et aux autorités publiques.
   - L'ensemble de ces données doivent être mis à la disposition d'organismes d'évaluation à l'instar de l'INESIA.

Recommandation n° 6 : Garantir une évaluation interdisciplinaire des vulnérabilités exploitées et engendrées par les systèmes d'IA générative.

Message clé : Pour mesurer l'impact des systèmes d'IA générative sur les vulnérabilités, il est nécessaire de **conduire des évaluations interdisciplinaires** et représentatives des différents usages des utilisateurs.

Exposé des enjeux

Face à l'opacité des systèmes d'IA générative, il est nécessaire d'établir un cadre d'évaluation permettant de mesurer les impacts sociaux de l'IA et les risques pour les droits fondamentaux et la société dans son ensemble de manière effective.

Mise en œuvre

1. **Renforcer le mandat et les moyens de l'INESIA** pour en faire une force propre davantage qu'un groupement d'acteurs, *via* un financement pérenne et un recrutement d'experts spécialisés. Ses prérogatives doivent être élargies :
    - Ouverture contrôlée des modèles via des environnements sécurisés (ex : salles blanches numériques.) ;
    - Pouvoirs contraignants : capacité à exiger des données d'usage ;
    - Missions élargies : inclure l'évaluation des impacts sociétaux (santé mentale, démocratie) et pas seulement techniques.

2. **Renforcer le rôle du Bureau de l'IA et des autorités nationales autour de l'évaluation de l'impact de l'IA générative sur les vulnérabilités**, tant en phase de pré-commercialisation qu'en phase de post-commercialisation. Ce rôle s'étend au-delà de l'approche technique de l'article 92 du RIA sur le pouvoir d'évaluation de la Commission européenne de la conformité des modèles d'IA aux obligations légales et des risques systémiques engendrés par ces modèles.

3. **Mener des évaluations dans une logique interdisciplinaire** (informatique, droit, sociologie, psychologie et psychiatrie, économie, mathématique science politique, *etc.*)

4. **Clarifier les rôles de chacun et éviter l'émiettement des moyens entre différents régulateurs.** L'activité des autorités de surveillance du marché, couplée à celle des évaluateurs externes, des chercheurs nationaux et des comités pluridisciplinaires indépendants permettra de développer les connaissances sur l'impact de l'IA générative sur les vulnérabilités et les risques sociétaux.

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 7 : Doter l'État d'un protocole national de réponse aux crises informationnelles liées à l'IA générative, en définissant des indicateurs de gravité et des canaux d'échange rapide de données avec les fournisseurs d'IA générative.

Message clé : Face à un événement de crise majeur impliquant une menace informationnelle générée ou amplifiée par IA, l'État doit disposer d'un protocole de détection et de réponse clair, activable rapidement, interministériel, reposant sur des indicateurs de gravité prédéfinis et sur des canaux d'échange établis à l'avance avec les entreprises d'IA, plutôt que d'improviser une coordination dans l'urgence.

Exposé des enjeux

Les crises informationnelles amplifiées par l'IA générative (ex. : *deepfake* visant une personnalité publique en pleine crise sécuritaire, campagne de désinformation coordonnée lors d'un attentat, contenu synthétique alimentant une panique financière ou sanitaire) exigent une réponse très réactive, voire immédiate, alors que les dispositifs français de gestion de crise (cellule interministérielle de crise, plans gouvernementaux) n'intègrent aujourd'hui pas de volet spécifique aux menaces informationnelles liées à l'IA. Sans seuils de gravité définis à l'avance ni canaux de contact préétablis avec les grandes plateformes et les développeurs de modèles, l'État s'expose à un temps de réaction trop long, en particulier à l'approche de périodes électorales.

Mise en œuvre

1. **Mettre en œuvre un protocole interministériel de gestion de crise dédié aux menaces informationnelles par IA.** Ce dispositif doit être résolument interministériel au regard de la diversité des champs concernés.

2. **Définir des indicateurs de suivi et des seuils de gravité** permettant de qualifier objectivement le niveau de menace (volume et vitesse de diffusion d'un contenu suspect, taux de coordination inauthentique détecté, sensibilité du contexte (élection, attentat, crise sanitaire ou financière)), sur le modèle des échelles de gravité déjà utilisées dans d'autres dispositifs de sécurité civile, afin de déclencher automatiquement le niveau de réponse adapté (veille renforcée, alerte interministérielle, mobilisation des plateformes).

3. **Établir des canaux de communication et d'échange de données avec les entreprises d'IA**, sous forme de points de contact désignés et d'accords-cadres définis en amont de toute crise, précisant les modalités d'échange accéléré d'informations (signalement de contenus, données d'usage agrégées) dans le respect du RGPD et du RSN.

4. **Organiser des exercices de simulation de crise informationnelle liée à l'IA** (*a minima* annuels), associant administrations publiques, plateformes et médias, pour tester le protocole en conditions réalistes et l'ajuster.

5. **Articuler ce protocole avec le dispositif européen, voire international** (Bureau de l'IA de la Commission européenne, réseau des AI Safety Institutes, G7 Rapid Response Mechanism) pour permettre une réponse coordonnée lorsque la menace dépasse le cadre national.

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 8 : Interdire la publicité déguisée, les pratiques masquées d'influence de nature commerciale et toute forme de monétisation exploitant les vulnérabilités des consommateurs.

Message clé : L'enjeu est d'exiger une **transparence effective des modèles d'affaires et toute influence à but commercial** exercée sur les systèmes d'IA générative pour une meilleure compréhension des utilisateurs et un meilleur contrôle des régulateurs et auditeurs indépendants.

Exposé des enjeux

Contrairement à une bannière publicitaire, une réponse insérée dans la conversation et orientée commercialement reste difficilement détectable, d'autant que les recommandations peuvent varier pour une même question. Ce risque est aggravé par la capacité de la publicité conversationnelle à exploiter le contexte émotionnel de l'utilisateur.

Ces risques ont déjà justifié, pour les plateformes d'intermédiation, les obligations de transparence des critères de classement du règlement Platform to Business (P2B) et du RSN. Un encadrement analogue doit s'appliquer aux systèmes d'IA générative, afin que l'utilisateur soit informé et puisse comprendre qui finance le service, identifier les contenus commercialement influencés et disposer de garanties contre toute recommandation rémunérée non déclarée.

Mise en œuvre

1. **Rendre obligatoire l'affichage d'une mention de modèle économique dans l'interface de tout système d'IA générative à destination du grand public** : *« Ce service est financé par : abonnements / publicité / commissions »*. Cela implique :
   - d'imposer une obligation **d'étiquetage explicite et visible de tout contenu commercialement influencé** dans les réponses des systèmes d'IA conversationnelle ;
   - **d'interdire toute influence commerciale non divulguée** sur le contenu des réponses, y compris les mécanismes de référencement de produits.

2. **Mandater une autorité publique à compétence économique pour assurer une analyse continue des modèles économiques et de leurs évolutions et pour la publication annuelle d'un état des lieux des pratiques de monétisation des principaux systèmes d'IA générative** à destination du grand public, disponibles en France – du fait du caractère national des marchés publicitaires et commerciaux. Une telle mesure gagnerait à être consolidée au niveau européen.
   - Cet état des lieux pourrait présenter : le modèle économique dominant, les mécanismes de transparence mis en place, les manquements identifiés et les sanctions

appliquées. Les fournisseurs d'IA générative pourraient mettre en place des solutions permettant à des auditeurs indépendants de vérifier l'éventuelle influence commerciale non divulguée et conférer à l'autorité publique ou organisations un certain accès aux modèles et agents déployés.

- Cet état des lieux s'appuiera sur les informations de monétisation décrites dans la Recommandation n° 5 complétée par la possibilité d'accéder à des données plus régulières.
- Les accès, visant à vérifier l'absence d'influence commerciale non divulguée et la conformité des marquages, pourraient prendre les formes suivantes :
  - l'accès harmonisé *via* des API aux données des campagnes publicitaires (émetteurs et critères de diffusion) permettant d'évaluer les méthodes de ciblage et le respect des règles qui les encadrent ;
  - l'accès harmonisé *via* des API aux données relatives aux transactions commerciales.
- Ce dispositif pourrait s'accompagner de la mise en place d'un mécanisme de signalement pour les utilisateurs et les tiers afin de faire remonter les suspicions de manipulation commerciale.

RECOMMANDATIONS PRIORITAIRES

Recommandation n° 9 : Maintenir et développer une politique nationale ambitieuse d'éducation à l'IA permettant à chacun d'en maîtriser les usages et les risques.

Message clé : L'éducation aux enjeux de l'IA générative doit être universelle, progressive et adaptée à chaque public, afin de renforcer l'esprit critique, de prévenir les risques de manipulation et de permettre à chacun de choisir et de maîtriser ses usages. Cette éducation doit couvrir tous les âges et tous les niveaux d'exposition, avec une attention particulière notamment pour les mineurs, les personnes âgées et les personnes en situation d'éloignement numérique, de précarité sociale ou encore de handicap.

Exposé des enjeux

Méfiance et désinformation. L'ensemble des citoyens, quels que soient leur âge ou leur niveau de maîtrise du numérique, est exposé à des contenus trompeurs générés par IA. L'absence de repères clairs sur le fonctionnement de l'IA générative favorise la propagation de fausses informations ou de représentations biaisées et l'altération du rapport au réel.

En période électorale en particulier, les systèmes d'IA générative peuvent agir comme catalyseur de la polarisation et fausser le débat public.

Dans leur rapport aux pouvoirs publics, les citoyens sont de plus en plus confrontés à l'usage de l'IA dans la prise de décisions les affectant, au détriment d'un accompagnement humain. **Cette automatisation des processus généralise la défiance envers l'État et l'isolement social.**

Généralisation de l'IA générative dans le service public mais également dans le cadre professionnel et social en général.

Mise en œuvre

1. **Garantir une éducation universelle à l'IA, adaptée aux besoins de chaque public, permettant d'en maîtriser les usages et de se prémunir de ses risques.** Pour cela, il est nécessaire d'assurer l'éducation et l'accompagnement de chacun à l'IA, en s'appuyant sur les secteurs éducatifs, de l'inclusion et de la médiation numérique, afin de capitaliser sur les structures et initiatives existantes, tout en portant une attention particulière aux plus vulnérables. Ces dispositifs doivent être pensés pour s'adapter aux besoins particuliers, tels que la perte d'autonomie ou le handicap.

2. **Maintenir et développer la médiation numérique par un accompagnement humain,** notamment *via* des ateliers gratuits dans les espaces collectifs territoriaux (Espaces France Services, médiathèques, associations, structures sociales, *etc.*), avec un ciblage prioritaire des publics éloignés du numérique. Accompagner systématiquement les citoyens dans

leurs démarches administratives lorsqu'ils sont confrontés à des outils automatisés et garantir une voie de recours vers un contrôle humain pour l'ensemble des décisions. L'omnicanalité doit rester la norme et non devenir l'exception. **Ces mesures nécessitent l'engagement d'un financement pérenne pour la médiation numérique de terrain**.

3. **Diffuser** dans les établissements d'enseignement secondaire, les établissements supérieurs, sur les réseaux sociaux ou encore dans les lieux publics, **des contenus de sensibilisation et des recommandations sur la manière d'utiliser les systèmes d'IA conversationnelle à des fins de vérification des faits, y compris en période de crise.** Ces recommandations, émises par le gouvernement devraient détailler les avantages et les limites des *chatbots* dans ce contexte, le risque accru pour les utilisateurs d'être exposés aux fausses informations en période de crise, ainsi que les bonnes pratiques numériques pour vérifier l'information.

4. **Pour les professionnels et acteurs clés** (enseignants, agents publics, journalistes, travailleurs sociaux, professionnels de l'inclusion et de la médiation numérique, entreprises privées, *etc.)* : **renforcer leur formation initiale et continue par la création de modules dédiés aux risques de l'IA générative**, à la manière d'accompagner l'ensemble des publics (citoyens, salariés, usagers du service public, élèves, étudiants, *etc.*) et de guides sectoriels pratiques, régulièrement mis à jour.

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS

Recommandation n° 10 : Renforcer la répression de la création de contenus à caractère sexuel non consentis et pédocriminels truqués (hypertrucages) générés par intelligence artificielle.

Message clé : Il est aujourd'hui nécessaire de renforcer le cadre pénal pour poursuivre pénalement les auteurs de *deepfakes* à caractère sexuel non consentis et pédocriminels générés par IA, en complément du cadre légal existant, pour lutter contre la banalisation de la création et de la circulation de ces contenus.

Exposé des enjeux

Sur le cadre réglementaire existant

- **L'Omnibus numérique en matière d'IA** crée deux interdictions supplémentaires, au sein de l'article 5 du RIA, de la mise sur le marché, de la mise en service et de l'utilisation d'un système d'IA qui génère ou diffuse des *deepfakes* pédocriminels et à caractère sexuel non consentis.
- **L'article 99 du RIA** prévoit que le non-respect de l'interdiction visée à l'article 5 peut faire l'objet d'une sanction administrative pouvant aller jusqu'à 35 millions d'euros ou 7 % du chiffre d'affaires annuel mondial.

Cette approche est nécessaire mais ne répond que partiellement à l'enjeu : il semble nécessaire de compléter les dispositions européennes, assorties de sanctions administratives, de dispositions nationales assorties de sanctions pénales, dans le respect du principe *non bis in idem*.

Sur les évolutions législatives à venir (UE)

- **La refonte de la directive relative à la lutte contre les abus sexuels et l'exploitation sexuelle des enfants et la pédopornographie[386] érige en infraction pénale le fait de concevoir ou d'adapter des systèmes d'IA destinés à produire des contenus pédocriminels, ainsi que de les diffuser**. Une fois officiellement adoptée, la directive révisée devra être transposée en droit français.

Sur les évolutions législatives à venir (France)

- **Une lacune juridique persistante :** L'article 227-23 du Code pénal sanctionne déjà la détention, la diffusion et la consultation de contenus pédopornographiques, y compris

[386] Directive 2011/93/UE du Parlement européen et du Conseil du 13 décembre 2011 relative à la lutte contre les abus sexuels et l'exploitation sexuelle des enfants, ainsi que la pédopornographie et remplaçant la décision-cadre 2004/68/JAI du Conseil.

ceux générés par IA. Cependant, il ne couvre pas explicitement la création de ces contenus. De même, la loi SREN du 21 mai 2024 (article 21) a introduit un délit spécifique pour la diffusion de contenus truqués à caractère sexuel (montages ou *deepfakes* générés par IA), sans aborder leur création. Les peines sont alourdies en cas de diffusion *via* un service de communication en ligne (article 15), reflétant les enjeux de viralité, mais le vide juridique persiste pour l'acte de production initial.

- **Pour combler cette lacune, la proposition de loi apportant une réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants**[387] propose de :
  - Créer une infraction visant la mise à disposition et la diffusion de modèles algorithmiques spécifiquement conçus pour produire des contenus sexuels non consentis, y compris des contenus pédocriminels (article 27).
  - Créer une infraction spécifique réprimant la création, la modification, la diffusion ou le partage de contenus pédocriminels générés par traitement algorithmique (article 28).

Mise en œuvre

Les différents textes législatifs adoptés ou en cours d'examen, au niveau européen et au niveau national, sont bien complémentaires et ne font pas courir de risques de double sanction d'un même acte illicite, car les sanctions portent d'une part sur de la matière administrative (RIA) et d'autre part sur de la matière pénale (directive 2011/93/UE et proposition de loi intégrale).

La présente Commission soutient la proposition de loi dite intégrale qui vise notamment à poursuivre individuellement les auteurs de contenus générés par IA à caractère sexuel non consentis ou pédocriminels. Cette proposition de loi a été adoptée à l'Assemblée nationale le 13 octobre 2026, et sera prochainement examinée par le Sénat. Cet ajout est complémentaire à l'interdiction posée à l'article 5 par l'Omnibus numérique en matière d'IA.

[387] Proposition de loi apportant une réponse intégrale au phénomène des violences sexuelles et sexistes contre les femmes et les enfants, n° 3106, déposée à l'Assemblée nationale le mardi 11 août 2026.

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS

Mesures additionnelles pour les mineurs

Recommandation n° 11 : Instaurer une autorisation préalable de mise sur le marché pour les jouets intégrant un système d'IA générative et définir des seuils d'âge restrictifs d'accès.

Message clé : Le marché des jouets dotés d'IA générative est en forte croissance et pose des risques spécifiques pour les mineurs. Il est essentiel de garantir un niveau élevé de sécurité lorsqu'ils interagissent avec des jouets intégrant des systèmes d'IA (en adéquation avec les recommandations n° 2 et n° 3).

Exposé des enjeux

1. Risques psychologiques et cognitifs

- **Attachement et dépendance affective** : risque d'altération de la distinction entre un humain et un jouet doté d'IA générative, avec développement d'une confiance excessive ou d'un attachement émotionnel envers l'outil. Ce phénomène est amplifié par l'illusion de réciprocité (mémoire de l'IA, langage affectif).
- **Substitution des repères éducatifs** : les jouets IA peuvent être perçus comme une autorité morale ou cognitive, notamment en raison de la nature immédiate et assurée de leurs réponses. Cela peut :
  - Entrer en conflit avec le développement du discernement ;
  - Remplacer les interactions humaines essentielles ;
  - Normaliser des comportements à risque.
- **Non-adaptation des réponses générées** au niveau de maturité ou de compréhension du mineur.

2. Risques liés à la vie privée et à la sécurité des données

- **Collecte intrusive d'une quantité massive de données personnelles et cognitives** (voix, conversations, préférences, données biométriques). Ces données peuvent être : exploitées à des fins commerciales, accessibles à des tiers en cas de fuite ou de piratage, ou encore réutilisées sans consentement éclairé, *etc.*
- **Absence de transparence** sur la finalité de la collecte des données, les destinataires des données, les durées de conservation ou encore les mécanismes de suppression.
- **Absence de garde-fous robustes** limitant les interactions.

Mise en œuvre

La conception, la commercialisation et l'usage de ces jouets doivent être strictement régulés pour limiter les risques identifiés. Bien que des cadres existent déjà (RIA, annexe I, articles 5 et 6, et règlement (UE) 2025/2509 relatif à la sécurité des jouets), des lacunes persistent et nécessitent des actions complémentaires :

1. **Établir des balises claires par tranche d'âge** à l'instar des repères 3-6-9-12[388] qui guident les usages numériques, il est crucial de définir des seuils d'âge pour l'accès aux jouets intégrant un modèle d'IA.

Pour ce faire, il est proposé d'**organiser une conférence de consensus sur les barrières d'âge.** Cette conférence doit associer acteurs institutionnels et décideurs politiques, organisations internationales de protection de l'enfance, experts scientifiques et chercheurs interdisciplinaires, représentants des familles et des mineurs, acteurs du numérique et de l'industrie du jouet et enfin, société civile. Elle sera l'occasion de définir notamment l'âge minimal pour une première exposition à un jouet IA et les modalités d'encadrement (durée, supervision) à mettre en place.

2. **Prévoir une autorisation préalable de mise sur le marché pour les jouets à destination des mineurs intégrant des modèles d'IA.**

Compte tenu de la vulnérabilité particulière des mineurs et des risques identifiés, la commercialisation des jouets dotés d'IA pourrait être **subordonnée à une autorisation préalable**, similaire à celle applicable aux médicaments par exemple. Ce mécanisme viserait à prévenir l'exploitation des vulnérabilités des utilisateurs et garantir un niveau élevé de protection.

Conditions requises :

- **Tests d'adversité (*red-teaming*)** : basés sur une **taxonomie publiée des risques spécifiques aux mineurs**, ces tests doivent être validés avant toute commercialisation.
- **Surveillance continue** : une fois sur le marché, des **évaluations récurrentes** (sur le modèle des évaluations *ex post* présentées dans la Recommandation n° 4) doivent être menées pour s'assurer que le jouet reste conforme aux exigences de sécurité, notamment en cas d'évolution de son système d'IA.

[388] 3-6-9-12. Pour un développement numérique durable et solidaire, avec et sans IA.

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS

Mesures additionnelles pour les mineurs

Recommandation n° 12 : Consacrer un programme d'éducation à l'intelligence artificielle dès le primaire.

Message clé : Renforcer l'éducation au numérique, aux médias et à l'information (EMI) pour en faire un réel pilier des parcours scolaires en l'élargissant aux technologies numériques et à l'IA. La Commission réitère cette recommandation fondamentale et complémentaire à la politique nationale ambitieuse d'éducation à l'IA pour tous les citoyens.

Exposé des enjeux

- L'éducation au numérique doit être instaurée dès le plus jeune âge, du primaire aux études secondaires, en adaptant les contenus en fonction de la maturité et des usages des élèves.
- Il est nécessaire de s'appuyer sur les dispositifs existants, à l'instar du parcours PixIA obligatoire pour les élèves de quatrième et de seconde, complété par des enseignements au lycée, ou encore la Charte pour l'éducation à la culture et à la citoyenneté numériques.
- Malgré les dispositifs existants, l'EMI reste insuffisante en classe par manque de temps et de moyens et doit faire l'objet d'un programme scolaire d'éducation au numérique, à l'IA, aux médias et à l'information structuré, en adaptant les apprentissages à l'âge des élèves et à leurs usages, sur le modèle du programme dédié à l'éducation à la vie affective, relationnelle et sexuelle paru en 2025[389].

Mise en œuvre

1. **Établir un socle commun fixant des objectifs minimaux en termes de nombre d'heures et de compétences à acquérir**.
    - Donner à tous les élèves une culture commune des outils numériques et de l'IA, en abordant leurs fonctionnements techniques, leurs enjeux éthiques et sociétaux (biais

[389] Ministère de l'Education nationale. « Un programme ambitieux : éduquer à la vie affective et relationnelle et à la sexualité ». 2025.

algorithmiques, désinformation, impact environnemental, *etc.*), leurs usages responsables, ainsi que leurs modèles économiques.

- Ce socle commun gagnerait à être généralisé à l'échelle européenne *via* des objectifs minimaux en termes de nombre d'heures et de compétences à acquérir, que chaque État membre intégrerait dans ses programmes scolaires. Ce programme doit être accompagné par un financement dédié pour garantir l'effectivité de la mesure.

2. **Confier l'élaboration de ce programme à des enseignants spécifiquement formés et volontaires.**

- Cela permettrait de cibler la formation sur un corps enseignant restreint, plutôt que de l'imposer à l'ensemble des professeurs, certains pouvant ne pas se sentir légitimes ou compétents pour aborder ces sujets.
- La consécration d'un tel programme n'exclut cependant pas une approche transversale de l'IA et du numérique dans les autres disciplines, où ces outils peuvent être mobilisés de manière contextuelle. En combinant un enseignement dédié et des applications disciplinaires, cette recommandation vise à garantir la diffusion d'une culture numérique partagée et critique, tout en préservant la liberté pédagogique des enseignants et des établissements.

RECOMMANDATIONS POUR ENCAPACITER ET PROTÉGER LES UTILISATEURS, NOTAMMENT LES MINEURS

Recommandation n° 13 : Garantir dans les salles de classe des modèles d'IA générative spécifiquement adaptés pour les mineurs.

Message clé : L'usage de l'IA doit s'opérer dans un cadre pédagogique précisément défini, en recourant à des outils labellisés ayant démontré leur efficacité, en adéquation avec l'âge et le niveau de développement des élèves.

Exposé des enjeux

- **Un risque spécifique d'affaiblissement cognitif propre au contexte scolaire :** contrairement à un usage récréatif, l'usage de l'IA en classe pose la question de la substitution de l'effort d'apprentissage. Une IA générative « classique », conçue pour maximiser la satisfaction immédiate de l'utilisateur (réponse directe, complète, sans détour), est opposée à l'objectif pédagogique, qui suppose au contraire la recherche, l'erreur, la mémorisation active et la construction progressive du savoir. L'usage de l'IA générative impacte la capacité à résoudre des problèmes, la mémoire à long terme, ainsi que la créativité.
- **Un risque d'homogénéisation des usages sans différenciation par âge** : les modèles d'IA générative grand public ne distinguent pas, dans leur *design*, un élève de CP d'un lycéen, alors que les besoins cognitifs, les capacités d'autorégulation et les vulnérabilités diffèrent radicalement selon les tranches d'âge (cf. chapitre 1).
- **Un risque de renforcement des inégalités scolaires** : en l'absence de cadre, les usages d'IA en classe dépendent des choix individuels des enseignants, des établissements ou des familles, ce qui peut créer des écarts importants entre élèves selon leur environnement pédagogique, social et économique.

Face à ces risques, il est préconisé de développer des applications et des usages différenciés en fonction de l'âge des élèves (telles que les solutions socratiques) permettant à la fois de prévenir ces risques et de garantir l'acquisition effective des savoirs.

Mise en œuvre

Si cette liste est non exhaustive, les recommandations renvoient au rapport « *Sortir de la clandestinité : mettre l'IA au service d'une nouvelle ambition pour le système éducatif* » du Conseil de l'IA et du numérique de juin 2026.

1. **Développer des modèles d'IA générative éducatifs spécifiques pour les mineurs**

   - Concevoir des modèles éducatifs différenciés par tranche d'âge, *« spécifiquement adaptés, capables de stimuler la curiosité et la réflexion plutôt que de fournir des réponses directes »*, adaptés aux besoins et développement des mineurs (ex : cycle 2, cycle 3, collège lycée), avec des paramètres de langage (ex. : simplifié pour les élèves du primaire ; approche plus analytique pour les lycéens), de complexité et de contenu adaptés au développement cognitif de chaque tranche d'âge.
   - Privilégier les outils dits « socratiques » qui guident l'élève par le questionnement plutôt que d'apporter une réponse clé en main.
   - Interdire par défaut, pour ces modèles destinés aux mineurs, les fonctionnalités identifiées comme problématiques.

2. **Garantir une labellisation et une évaluation systématique des solutions d'IA auxquelles sont exposés les élèves**

   - Mettre en place une procédure de labellisation pour certifier les modèles d'IA destinés aux élèves par tranches d'âge, en particulier les outils socratiques, garantissant leur conformité aux objectifs pédagogiques et éthiques.
   - Définir un cahier des charges précis que doivent respecter les éditeurs pour obtenir ce label, conditionnant l'entrée en classe de toute solution.
   - Rendre le processus de candidature accessible aux acteurs publics comme privés (start-ups EdTech, associations, éditeurs scolaires), afin de ne pas réserver le marché aux seuls grands fournisseurs et de favoriser une offre souveraine et diversifiée.
   - Établir un processus d'évaluation rigoureux, en confiant au Conseil scientifique de l'Éducation nationale la responsabilité d'évaluer les modèles d'IA proposés, en collaboration avec des experts en pédagogie, psychologie du mineur et éthique du numérique.
   - Des critères d'évaluation doivent être instaurés : (i) le respect des principes pédagogiques (adaptation à l'âge, outil socratique) ; (ii) l'absence de biais discriminatoires ou de mécanismes addictifs ; (iii) la conformité à la règlementation européenne et nationale en vigueur.

3. **Sanctuariser des espaces et temps d'apprentissage sans IA**

   - Instaurer une éducation duale : apprendre et enseigner « sans » IA et « avec » IA : avec « *des espaces et temps sans IA pour développer les compétences cognitives fondamentales, tout en intégrant de manière encadrée et réfléchie les outils d'IA pour préparer les élèves aux risques et opportunités qu'elle permet. Cette approche duale doit être coconstruite avec les enseignants, qui sont les mieux placés pour déterminer quand et comment l'IA apporte une réelle plus-value pédagogique* ».
   - Définir des plages horaires ou des disciplines où l'usage de l'IA est interdit (ex. : cours de rédaction, débats, activités manuelles) pour préserver le développement de compétences fondamentales.

- Confier aux enseignants, en concertation avec les équipes pédagogiques, la détermination des moments où le recours à l'IA apporte une réelle plus-value.
- Réviser les programmes scolaires pour y intégrer progressivement les compétences liées à l'IA, et garantir une formation adaptée, initiale et continue des enseignants.

RECOMMANDATIONS POUR RENFORCER LA COHÉSION SOCIALE ET DÉMOCRATIQUE À L'ÈRE DE L'IA GÉNÉRATIVE

Recommandation n° 14 : Détecter de manière précoce les manipulations de l'information par les systèmes d'IA générative.

Message clé : Face à l'industrialisation de la désinformation générée par IA, l'État doit renforcer sa capacité de détection précoce et systémique, articulée avec le cadre européen (RSN) et accessible à la société civile, afin de repérer les campagnes de désinformation générées par IA avant qu'elles ne deviennent virales.

Exposé des enjeux

La désinformation générée par IA (*deepfakes*, faux sites, contenus synthétiques, comptes coordonnés) représente une menace croissante pour la démocratie, en particulier en période électorale ou de crise (sanitaire, sécuritaire, géopolitique). L'État doit renforcer sa capacité à détecter les signaux faibles (premières apparitions de récits trompeurs, coordination inauthentique de comptes, anomalies dans les flux de diffusion) et soutenir la recherche sur des méthodes innovantes (détection non supervisée, analyse en temps réel, analyse multimodale texte/image/vidéo/audio) pour anticiper les campagnes de désinformation avant qu'elles ne deviennent virales.

Mise en œuvre

1. **Construire une capacité d'audit « *system-wide* » des plateformes**, à même de détecter seule ces manipulations là où elles se logent, en s'appuyant sur l'article 40 du RSN (accès aux données pour les chercheurs agrémentés) et l'acte délégué de la Commission européenne pour obtenir un accès structuré et régulier aux données de recommandation et de diffusion des grandes plateformes.

2. **Créer un dispositif de veille partagée** associant l'Arcom, VIGINUM, l'INESIA et les organisations de *fact-checking*, avec un canal de signalement accéléré vers les plateformes en période électorale.

3. **Financer la recherche** (*via* l'ANR, les Programmes prioritaires de recherche IA type PEPR) sur la détection automatique de désinformation et des signaux faibles en temps réel.

4. **Expérimenter un bac à sable réglementaire (*sandbox*)** permettant à des chercheurs et des *start-ups* de tester des outils de détection sur des jeux de données réels, sous supervision de la CNIL.

5. **Imposer aux entreprises spécialisées en IA générative proposant des services de *chatbot* en France de modifier leurs interfaces utilisateur conversationnelles afin d'y inclure une mise en garde plus visible et plus précise à l'intention des utilisateurs concernant les limites de la vérification des faits lors d'événements de crise en temps réel.** Ces avertissements pourraient prendre la forme d'un message d'alerte contextuel lorsqu'un *chatbot* est interrogé sur une crise en cours et en lien avec des mots-clés associés à celle-ci, signalant aux utilisateurs qu'ils ne doivent pas considérer les réponses comme des faits avérés tant qu'un incident se déroule en temps réel.

RECOMMANDATIONS POUR RENFORCER LA COHÉSION SOCIALE ET DÉMOCRATIQUE À L'ÈRE DE L'IA GÉNÉRATIVE

Recommandation n° 15 : Instaurer un code de bonne conduite sur l'usage de l'IA générative en période électorale.

Message clé : Le changement du rapport à la vérité des citoyens, renforcé par la généralisation des systèmes d'IA générative, exige la mise en place d'un code de bonne conduite pour les candidats et les partis politiques sur l'usage des systèmes d'IA générative en période électorale.

Exposé des enjeux

Pour préserver la confiance dans les institutions démocratiques, et, en période électorale, garantir des élections libres et éclairées, il est impératif que les candidats et les partis politiques s'engagent publiquement à ne pas utiliser l'IA générative pour tromper, manipuler ou désinformer l'électorat. À l'approche de l'élection présidentielle française de 2027, un code de bonne conduite permettrait d'encadrer l'usage de ces technologies et de sensibiliser tant les acteurs politiques que les citoyens.

Mise en œuvre

1. **Rédiger et adopter un code de bonne conduite national**, en s'inspirant du modèle québécois (IVADO/CEIMIA, 2025-2026), qui interdit l'usage de l'IA pour générer ou diffuser des contenus visant à tromper, manipuler ou désinformer l'électorat, y compris par négligence dans la vérification des faits. Le code reposerait sur quatre principes fondamentaux :
    - Transparence : toute utilisation d'IA pour produire du contenu électoral (visuel, sonore ou textuel) doit être clairement signalée, par un étiquetage standardisé et lisible (pictogramme, mention systématique).
    - Usage responsable : interdiction de recourir à l'IA pour créer ou amplifier de fausses informations, des deepfakes ou des messages ciblés trompeurs (microciblage manipulatoire).
    - Sécurité et gouvernance : mise en place, au sein de chaque équipe de campagne, d'un point de contact responsable de la vérification des contenus générés par IA avant diffusion.
    - Sensibilisation et prévention : formation obligatoire des équipes de campagne et information du public sur les risques liés à l'IA.

Ce code s'inscrit dans le prolongement des textes existants sur la transparence électorale et la lutte contre la manipulation de l'information (loi du 22 décembre 2018, RGPD) ; son caractère volontaire limite toutefois la portée contraignante des sanctions.

2. **Engager les partis politiques et les candidats à signer le code** lors d'une cérémonie publique avant le début officiel de la campagne électorale, avec médiatisation pour renforcer l'adhésion citoyenne.

3. **Prévoir un mécanisme de suivi et de signalement des manquements**, porté par une autorité indépendante (Haute Autorité pour la transparence de la vie publique, Arcom ou une instance *ad hoc*), avec publication régulière d'un bilan des signalements et, le cas échéant, des suites données.

RECOMMANDATIONS POUR RENFORCER LA COHÉSION SOCIALE ET DÉMOCRATIQUE À L'ÈRE DE L'IA GÉNÉRATIVE

Recommandation n° 16 : Garantir la représentativité des langues et des cultures minoritaires dans l'IA générative, en mobilisant les corpus et les catalogues de référence.

Message clé : Garantir que les modèles d'IA générative intègrent et valorisent la diversité linguistique, culturelle est un enjeu de souveraineté numérique et de préservation du pluralisme démocratique. Les langues et les cultures peu dotées doivent être activement soutenues pour ne pas être marginalisées dans l'écosystème numérique mondial.

Exposé des enjeux

- **Biais culturels et linguistiques :** entraînés surtout sur des corpus anglophones et occidentaux, les grands modèles de langage véhiculent principalement les valeurs des pays anglophones, quelle que soit la langue dans laquelle on les interroge. Il en résulte un risque d'uniformisation culturelle.
- **Marginalisation des langues peu dotées** : mal représentées dans les corpus, ces langues sont moins bien servies par les outils numériques, ce qui limite l'accès et les usages de leurs locuteurs et menace la vitalité des langues elles-mêmes.
- **Uniformisation culturelle** : faute de transparence sur les corpus d'entraînement et les méthodes d'alignement (apprentissage par renforcement à partir de rétroaction humaine, garde-fous, *etc.*), les modèles tendent à reproduire les références culturelles dominantes issues des données d'entraînement et de leurs concepteurs.

Mise en œuvre

1. **Valoriser les corpus et catalogues existants** plutôt que de multiplier les nouvelles annotations. Il s'agit de :
    - soutenir l'ouverture des catalogues des bibliothèques, des institutions culturelles et des administrations (par exemple Gallica, de la BnF) ;
    - cibler de nouveaux financements sur les langues peu dotées et les données orales, en lien avec les communautés linguistiques ;
    - s'appuyer sur les infrastructures européennes (European Language Grid, Projet High Performance Language Technologies (HPLT), Alliance for Language Technologies) ;

- mutualiser les efforts avec l'Organisation internationale de la Francophonie (OIF) et dans le cadre de la stratégie France-Québec 2025-2030.

2. **Soutenir des modèles ouverts adaptés aux langues et cultures minoritaires**, en s'inspirant d'initiatives comme ALIA en Espagne ou Mila au Québec qui développent des outils améliorant l'utilisation de certaines langues dans le développement et le déploiement de l'IA dans le monde.

3. **Inciter les fournisseurs à utiliser ces corpus** avec une contribution financière réduite qui serait accordée à ceux qui concluent des partenariats avec les organismes détenteurs de ces catalogues.

4. **Mesurer la représentativité linguistique et culturelle des modèles.** Il s'agit de développer des bancs d'essai publics qui évaluent la qualité des principaux modèles en français et dans les langues peu dotées, ainsi que leur prise en compte des références culturelles, puis d'en publier régulièrement les résultats.

5. **Renforcer la découvrabilité des contenus culturels et journalistiques francophones** dans les réponses des systèmes d'IA conversationnelle, en dialogue avec les éditeurs, sur le modèle des obligations imposées aux plateformes de *streaming*[390].

[390] Les obligations françaises imposées aux services de vidéo à la demande (SMAD) combinent des quotas d'œuvres, leur mise en avant et, pour certains services, une contribution au financement de la production.

RECOMMANDATIONS POUR ACCROÎTRE LA TRANSPARENCE ET LA RESPONSABILITÉ DES ACTEURS DU MARCHÉ DE L'IA GÉNÉRATIVE

Recommandation n° 17 : Encadrer les inférences de données personnelles réalisées à partir des *prompts* de l'utilisateur.

Message clé : Les systèmes d'IA conversationnelle déduisent de simples échanges des informations que l'utilisateur ne leur a jamais confiées : état de santé, opinions politiques, difficultés financières, fragilité psychologique. **Ces inférences doivent être rendues visibles, puis strictement encadrées**, notamment **lorsqu'elles servent à des fins commerciales** (ciblage publicitaire, personnalisation tarifaire, revente à des tiers).

Il est nécessaire d'obliger les fournisseurs d'IA à **déclarer, dans leur politique de confidentialité, les catégories d'attributs personnels qu'ils sont susceptibles d'inférer, y compris lorsque l'utilisateur ne les a jamais divulgués**. Ensuite, nous recommandons de **soumettre toute exploitation commerciale de ces profils inférés aux garanties prévues pour les données sensibles à l'article 9 du RGPD**, dès lors que ces inférences portent sur la santé, les opinions ou tout autre attribut sensible.

Exposé des enjeux

Le principal enjeu est l'asymétrie d'information. L'utilisateur ignore ce que l'agent déduit de lui et donc ce qui peut être exploité commercialement. Un système d'IA conversationnelle reçoit des échanges longs et personnels, bien plus riches qu'une requête de moteur de recherche. On y confie des symptômes, des inquiétudes d'argent, des doutes sur sa vie de couple ou son travail. De ces confidences, un modèle peut inférer des attributs que personne n'a déclarés : une pathologie probable, une orientation politique, un niveau de revenu, un état de détresse.

Le cadre actuel ne répond que partiellement à ce risque ; trois limites ressortent :

- Les articles 13 et 14 du RGPD imposent d'informer sur les catégories de données traitées, mais les politiques de confidentialité décrivent en général les données fournies, rarement celles qui sont déduites.
- L'article 9 du RGPD protège la santé, les opinions politiques, les convictions religieuses, l'orientation sexuelle ou l'appartenance syndicale, entre autres. La situation financière, l'état mental ou l'état émotionnel n'y figurent pas, alors que leur exploitation est tout aussi sensible.
- La CJUE a jugé que des données révélant indirectement une information sensible relèvent de l'article 9. Ce principe reste rarement appliqué aux profils construits à partir de conversations avec un système d'IA conversationnelle.

L'exploitation commerciale de ces profils (ciblage publicitaire, partage avec des tiers, personnalisation tarifaire) amène un risque supplémentaire. Elle peut viser précisément les

personnes les plus fragiles, ce qui rejoint les pratiques d'exploitation des vulnérabilités que le RIA (article 5) cherche déjà à interdire.

Mise en œuvre

1. **Transparence des inférences**. Les fournisseurs de systèmes d'IA conversationnelle doivent déclarer explicitement, dans leur politique de confidentialité, les catégories d'attributs personnels qu'ils sont susceptibles d'inférer à partir des échanges conversationnels, y compris lorsque l'utilisateur ne les a jamais divulgués. Un modèle de déclaration harmonisé, publié par l'autorité de contrôle, en facilitera la comparaison entre opérateurs.

2. **Garanties renforcées pour l'exploitation commerciale**. Toute exploitation commerciale de profils inférés (ciblage publicitaire, partage ou revente à des tiers, personnalisation tarifaire) est soumise aux garanties de l'article 9 du RGPD, notamment le consentement explicite, lorsque les inférences portent sur la santé, les opinions ou un autre attribut sensible. Cette protection est étendue aux inférences sur la situation financière et sur l'état mental et émotionnel.

3. **Droits des utilisateurs**. Chaque utilisateur doit pouvoir accéder aux inférences le concernant, les contester et s'opposer à leur usage commercial, en prolongeant les droits d'accès et d'opposition des articles 15 et 21 du RGPD.

4. **Contrôle et sanctions**. Les autorités de protection des données vérifient la conformité des déclarations, notamment par des audits. Le non-respect entraîne les sanctions du RGPD.

RECOMMANDATIONS POUR ACCROÎTRE LA TRANSPARENCE ET LA RESPONSABILITÉ DES ACTEURS DU MARCHÉ DE L'IA GÉNÉRATIVE

Recommandation n° 18 : Renforcer la responsabilité des fournisseurs de systèmes d'IA générative.

Message clé : Parmi les instruments relatifs à la responsabilité civile applicables aux fournisseurs de systèmes d'IA générative, la directive de 2024 relative à la responsabilité du fait des produits défectueux, devant être transposée en droit français en décembre 2026, consacre la mise en œuvre de la responsabilité civile desdits fournisseurs en cas de mise sur le marché d'un système ou d'un modèle d'IA défectueux ayant causé un dommage. Ce cadre juridique peut être renforcé en écartant, ainsi que l'article 18 de cette directive le permet, l'exonération pour risque de développement.

Exposé des enjeux

Les règles relatives à la responsabilité civile des fournisseurs d'IA générative doivent être précisées et renforcées. La directive relative à la responsabilité du fait des produits défectueux prévoit que « À l'ère numérique, les produits peuvent être corporels ou incorporels. Les logiciels, tels que les systèmes d'exploitation, les micrologiciels, les programmes informatiques, les applications ou les systèmes d'IA, sont de plus en plus répandus sur le marché et jouent un rôle de plus en plus important pour la sécurité des produits ».

Les systèmes et les modèles d'IA sont susceptibles de causer des dommages, en ce compris des décès ou des atteintes à la santé mentale des usagers.

Exonération pour risque de développement : certains fournisseurs d'IA générative pourraient invoquer l'exonération pour risque de développement (impossibilité de déceler le défaut au regard de l'état objectif des connaissances scientifiques et techniques) prévue par l'article 11 de la directive pour échapper à leur responsabilité, si celle-ci n'est pas écartée par le législateur français, ainsi que le permet l'article 18 de la directive relative à la responsabilité du fait des produits défectueux.

Mise en œuvre

La directive (UE) 2024/2853 relative à la responsabilité du fait des produits défectueux :

- définit dans son article 4.1, la notion de « produit » comme étant *« tout meuble, même s'il est incorporé dans un autre meuble ou dans un immeuble ou interconnecté avec celui-ci ; le terme comprend […] les logiciels » ;*
- définit dans son article 4.3 la notion de « composant » comme *étant « tout élément, corporel ou incorporel, matière première ou service connexe, intégré dans un produit ou interconnecté avec celui-ci » ;*

- précise dans son considérant 13 : « *À l'ère numérique, les produits peuvent être corporels ou incorporels. Les logiciels, tels que les systèmes d'exploitation, les micrologiciels, les programmes informatiques, les applications ou les systèmes d'IA, sont de plus en plus répandus sur le marché et jouent un rôle de plus en plus important pour la sécurité des produits. Les logiciels peuvent être mis sur le marché en tant que produits autonomes ou être intégrés par la suite dans d'autres produits en tant que composants, et ils sont susceptibles de causer des dommages lors de leur exécution* ».

Le RIA précise dans son considérant 97 que les modèles d'IA sont des composants des systèmes d'IA.

À l'occasion de la transposition de la directive relative à la responsabilité du fait des produits défectueux, transposer explicitement le contenu du considérant 13, de l'article 4.1, de l'article 4.3 dudit texte, établissant une correspondance entre la notion de « système d'IA » et celle de « produit », ainsi qu'entre la notion de « modèle d'IA » et celle de « composant » du système d'IA, lorsqu'on les articule avec le considérant 97 du RIA. Cela permettrait de **renforcer juridiquement la mise en œuvre de la responsabilité civile en matière d'IA générative et de rappeler le principe de la responsabilité solidaire entre fabricants de modèles et fabricants de systèmes d'IA**, lorsqu'ils sont des personnes juridiquement distinctes.

À l'occasion de la transposition de la directive relative à la responsabilité du fait des produits défectueux, prévue pour le mois de décembre 2026, utiliser le mécanisme dérogatoire à l'harmonisation maximale présent à l'article 18 pour **exclure de l'exonération pour risque de développement (sous certaines conditions)** les fournisseurs de systèmes d'IA générative de type IA conversationnelle à destination du grand public, et en particulier de type IA conversationnelle à destination des mineurs, tels que visés dans la proposition d'EU KIDS Act.

Annexe 1 : Lettre de mission

MINISTÈRE
CHARGÉ DE L'INTELLIGENCE
ARTIFICIELLE
ET DU NUMÉRIQUE
Liberté
Égalité
Fraternité

Paris, le 20 FEV. 2026

LA MINISTRE DÉLÉGUÉE

Nos références : MEFI-D26-01853

Madame la co-présidente,
Monsieur le co-président,

Depuis plus de trois ans, les agents conversationnels fondés sur l'intelligence artificielle (IA) générative se sont imposés de façon croissante dans les usages numériques des citoyens français et européens : quand le réseau social Instagram mettait 30 mois pour atteindre 100 millions d'utilisateurs, il n'a fallu que 60 jours à ChatGPT pour atteindre ce seuil symbolique. Désormais, les modèles se sont multipliés, sont devenus plus puissants, et les applications qui en découlent se sont diversifiées ; IA conversationnelles ou *chatbots*, génération de texte, d'image et de vidéo, « compagnons IA », jouets aux systèmes d'IA intégrés, etc.

Si ces développements technologiques ouvrent des perspectives d'innovation et de transformation, tant de l'action publique que des métiers au sein des entreprises, l'IA générative est d'ores et déjà à l'origine de risques graves et de dérives d'ordres divers, notamment en matière de cybercriminalité, de production et de diffusion massive de contenus illicites ou pouvant porter atteinte à la dignité humaine.

Les applications grand public fondées sur l'IA générative soulèvent ainsi des interrogations sur leurs possibles atteintes à la santé mentale et au bien-être social, ainsi que sur les troubles du développement cognitif et éducatif de ses utilisateurs. Ces questions se posent en particulier pour les usagers les plus jeunes, par ailleurs confrontés à l'anthropomorphisation croissante des systèmes d'IA. Pour autant, d'autres risques, à l'image de la violation des données personnelles, de l'inégalité d'accès aux services numériques, de la manipulation de l'information ou de biais discriminants, prouvent que les vulnérabilités induites par l'IA générative n'épargnent personne. Nous y sommes toutes et tous confrontés.

Dans ce cadre, dans la continuité des réflexions initiées par le Président de la République sur le rôle du numérique dans la démocratie et les effets de ces technologies sur les mineurs, il apparaît aujourd'hui essentiel de renforcer notre compréhension des vulnérabilités induites par ces technologies sur l'ensemble de la population, afin de mieux prévenir et encadrer leurs conséquences.

1.

Madame Anne BOUVEROT
Monsieur Guillaume POUPARD
Co-présidents du Conseil de l'intelligence artificielle et du numérique
6 rue Louise Weiss
75013 PARIS

C'est pour ces raisons et afin de faire toute la lumière sur ces enjeux que j'ai l'honneur de vous confier pour mission de constituer une Commission dite « IA générative et vulnérabilités », dont la co-présidence est confiée à Amine Benyamina, professeur d'addictologie à l'APHP, Raphaël Gaillard, professeur de psychiatrie à l'Université Paris Cité et Serena Villata, directrice de recherche CNRS en Intelligence artificielle au centre INRIA d'Université Côte d'Azur.

Les travaux seront pilotés et coordonnés par le secrétariat général du Conseil de l'intelligence artificielle et du numérique (CIANum), sous votre supervision directe. Je vous demande de bien vouloir constituer un groupe de travail multidisciplinaire qui sera composé d'une dizaine d'experts indépendants, retenus pour leurs compétences et leurs mérites dans leurs domaines respectifs.

Cette Commission sera chargée de rédiger un rapport qui devra :

- considérer l'IA générative dans ses différentes applications en dressant une cartographie des usages (support pour l'enseignement, pour le travail, pour la recherche, *chatbots*, « compagnons IA », jouets, *etc.*) ;
- identifier l'ensemble des risques ainsi que les vulnérabilités les plus critiques, par grandes catégories identifiées dans le cadre de la cartographie, en les hiérarchisant selon leur impact réel et potentiel sur les utilisateurs, en traitant avec une attention spécifique les effets sur les mineurs et en veillant à les objectiver au vu des travaux scientifiques à date ;
- analyser le cadre réglementaire et technique existant au niveau national et européen, les enjeux éthiques sous-jacents et les réflexions analogues conduites en dehors de nos frontières ;
- formuler des propositions étayées et concrètes à l'échelle française et européenne.

Le rapport de la Commission « IA générative et vulnérabilités » sera restitué au cours du mois de mai 2026 et comprendra des recommandations opérationnelles et consensuelles, susceptibles d'éclairer les décisions publiques et d'engager des actions pour renforcer le cadre de protection. Des observations ou des recommandations complémentaires, de nature plus prospective, pourraient être rendues à la rentrée 2026.

Dans cette perspective, la commission pourra s'appuyer sur l'expertise des services concernés de l'État, en particulier au sein du ministère de l'Économie, des Finances et de la Souveraineté industrielle, énergétique et numérique. En outre, vous travaillerez en étroite collaboration avec le Comité consultatif national d'éthique du numérique (CCNEN) qui apportera tout son concours à ces réflexions.

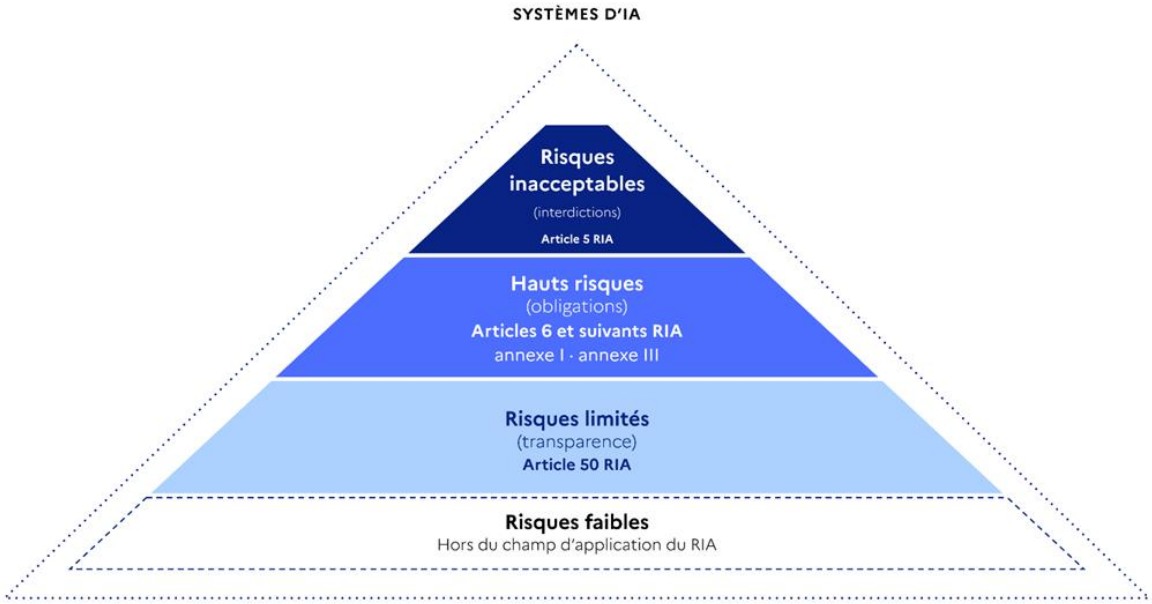
Je vous prie d'agréer, Madame la co-présidente, Monsieur le co-président, l'expression de ma considération distinguée.

Anne LE HÉNANFF

2.

Annexe 2 : L'appréhension de l'IA générative par le Règlement européen sur l'IA (RIA)[391]

L'approche par les risques du RIA



Le RIA est souvent présenté comme une *« réglementation fondée sur une approche par les risques »*, comportant quatre niveaux de risques concernant les systèmes d'IA[392] :

- **Le niveau du risque inacceptable[393]** :

Tous les systèmes d'IA considérés comme une menace manifeste pour la sécurité, la santé et les droits fondamentaux des personnes sont interdits. L'article 5 interdit huit pratiques[394]. L'Omnibus numérique en matière d'IA ajoute deux nouvelles pratiques interdites.

[391] Schémas conçus par une juriste de la Commission ; conception graphique adaptée à l'aide d'un outil d'intelligence artificielle.
[392] Commission européenne. Législation sur l'IA.
[393] Article 5 du RIA.
[394] Par exemple : un système d'IA qui a recours à des techniques subliminales, au-dessous du seuil de conscience d'une personne, ou à des techniques délibérément manipulatrices ou trompeuses, avec pour objectif ou effet d'altérer substantiellement le comportement d'une personne ou d'un groupe de personnes en portant considérablement atteinte à leur capacité à prendre une décision éclairée, amenant ainsi la personne à prendre une décision qu'elle n'aurait pas prise autrement, d'une manière qui cause ou est raisonnablement susceptible de causer un préjudice important à cette personne, à une autre personne ou à un groupe de personnes ; ou un système qui exploite des vulnérabilités dues à l'âge, au handicap ou à la situation sociale ou économique spécifique d'une personne physique ou d'un groupe de personnes donné avec pour

- **Le niveau du haut risque**[395] :

Les cas d'utilisation de l'IA susceptibles de présenter des risques graves pour la santé, la sécurité ou les droits fondamentaux sont classés comme présentant un risque élevé. Les annexes I et III du RIA répertorient les cas d'utilisation à haut risque.

Les systèmes d'IA à haut risque sont soumis à des obligations strictes à la charge des opérateurs avant de pouvoir être mis sur le marché.

- **Le niveau de risque limité** :

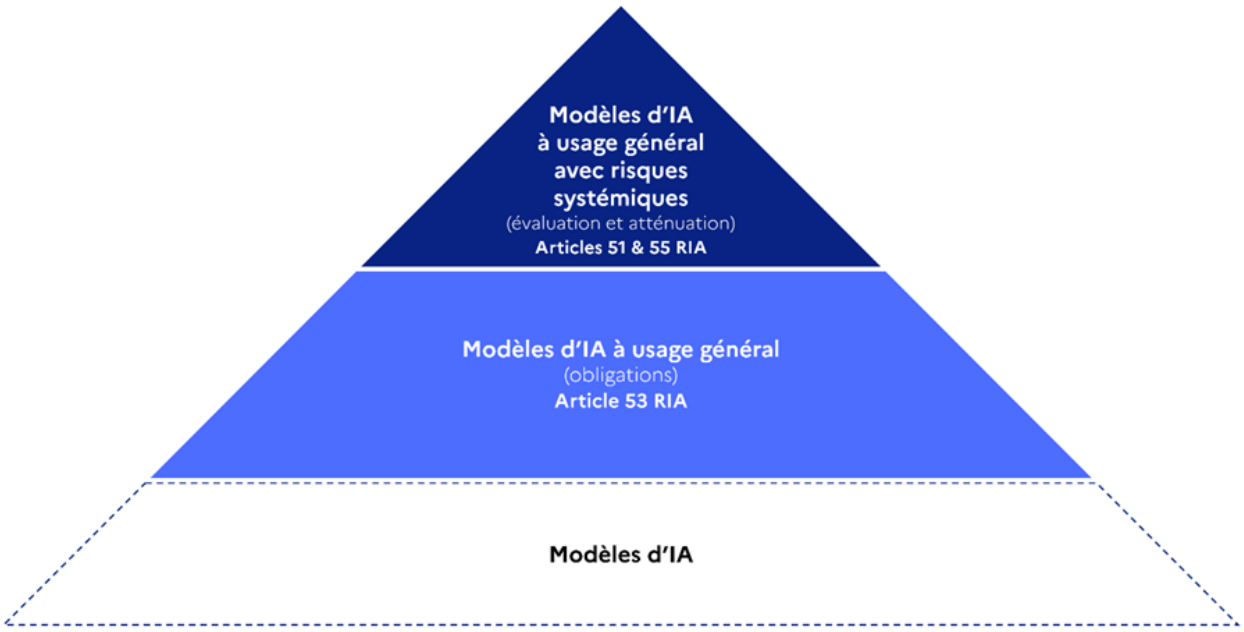
Ce niveau fait référence aux risques pour lesquels il est nécessaire d'assurer la transparence concernant l'utilisation de l'IA. L'article 50 introduit des obligations spécifiques de divulgation afin de garantir la protection des personnes.

- **Le niveau de risque minimal ou nul :**

Le RIA n'introduit pas de règles pour les IA considérées comme présentant un risque minimal ou nul.

*

Le RIA établit, en outre, des règles harmonisées pour la **mise sur le marché des modèles d'IA à usage général**. À cette fin, le règlement comprend une taxonomie des modèles d'IA, sous la forme d'une pyramide conceptuelle de modèles : la notion la plus générale est celle de modèle d'IA, qui n'est pas réglementée dans le RIA. Au-dessus de cette première notion se trouve un concept plus précis et plus spécifique : le **modèle d'IA à usage général** qui impose un premier lot d'obligations aux fournisseurs desdits modèles. Au sommet de la pyramide se trouve un concept encore plus spécifique : le **modèle d'IA à usage général et à risque systémique**, qui entraîne, en particulier, une obligation d'atténuation de ce type de risque aux fournisseurs desdits modèles.



objectif ou effet d'altérer substantiellement le comportement de cette personne ou d'un membre de ce groupe d'une manière qui cause ou est raisonnablement susceptible de causer un préjudice important à cette personne ou à un tiers (b).

[395] Article 6 et suivants du RIA.

À cet égard, si beaucoup d'exacerbations de vulnérabilités par des usages des IA génératives peuvent être saisies par le biais des pratiques d'IA interdites de l'article 5 ou des règles de transparence de l'article 50 du RIA, les règles attachées aux systèmes d'IA à haut risque des articles 6 et suivants, de même que les règles attachées aux modèles d'IA à usage général et à risque systémique des articles 51 et suivants ne permettront pas systématiquement de saisir la multiplicité des usages des IA génératives engendrant des vulnérabilités, tant et si bien qu'au-delà des solutions de *lege lata* (selon la loi en vigueur) qui ont été évoquées tout au long de ce rapport, ont également été formulées des propositions de *lege ferenda* (selon la loi à proposer).

*

Les risques systémiques : une notion protéiforme

La notion de risque systémique est une notion protéiforme qui se retrouve, assortie de définitions diverses, dans plusieurs instruments européens, dont le RIA et le RSN. Aux termes de l'article 3, paragraphe 65, du RIA, **le risque systémique** est défini comme *« un risque spécifique aux capacités à fort impact des modèles d'IA à usage général, ayant une incidence significative sur le marché de l'Union en raison de leur portée ou d'effets négatifs réels ou raisonnablement prévisibles sur la santé publique, la sûreté, la sécurité publique, les droits fondamentaux ou la société dans son ensemble, pouvant être propagé à grande échelle tout au long de la chaîne de valeur »*.

Dans le prolongement de l'article 56 et du considérant 116 du RIA, **le code de bonnes pratiques relatif aux modèles d'IA à usage général** a, à cet égard, dressé une taxonomie des risques systémiques.

Même si cette taxonomie des risques systémiques présente dans l'annexe du code de bonnes pratiques des modèles d'IA à usage général[396] peut être considérée comme de « l'infra-droit », elle n'en demeure pas moins le fruit d'un processus normatif synthétisant les avis divergents de nombreuses parties prenantes et d'experts[397].

Or, il ressort de cette taxonomie que **certains risques systémiques se logent principalement dans le modèle d'IA**, pendant que d'autres risques systémiques ne peuvent faire l'économie, au titre de leur étude, d'autres éléments que le modèle, comme le système dans lequel le modèle s'insère, mais également l'environnement sociotechnique dans lequel se déploient le modèle et le système, pour finalement étudier leur impact sur la totalité de la société.

En particulier, la taxonomie de l'annexe 1.4 du chapitre « Sûreté et sécurité » du code de bonnes pratiques de l'IA à usage général comprend, à côté des risques qui sont intrinsèquement inhérents aux modèles, des **risques qui puisent leur origine, non seulement dans les capacités du modèle mais également dans le *design* du système l'accueillant, ainsi que dans les caractéristiques et vulnérabilités évolutives de l'environnement sociotechnique et de la société**, dans lesquels l'un et l'autre sont déployés.

Cette observation est tout particulièrement adaptée au second et au quatrième risque systémique identifiés dans l'annexe 1.4 du chapitre « Sûreté et sécurité » du code de bonnes pratiques de l'IA à usage général : *« (2) Perte de contrôle : risques liés à la perte par l'homme de la capacité à diriger, modifier ou arrêter de manière fiable un modèle. Ces risques peuvent résulter d'un décalage par rapport*

[396] Commission européenne « Code de bonnes pratiques de l'IA à usage général ». 2025.
[397] *Ibid.*

à l'intention ou aux valeurs humaines, du raisonnement autonome, de l'autoreproduction, de l'autoamélioration, de la tromperie, de la résistance à la modification des objectifs, d'un comportement visant à acquérir du pouvoir, ou de la création ou de l'amélioration autonomes de modèles ou de systèmes d'IA » et *« (4) Manipulation nuisible (harmful manipulation) : les risques liés au fait de permettre une distorsion stratégique du comportement ou des convictions humaines en ciblant de larges populations ou des décideurs jouant un rôle crucial, par le biais de la persuasion, de la tromperie ou du ciblage personnalisé »*. **Cela inclut le renforcement significatif des capacités de persuasion, de tromperie et de ciblage personnalisé**, en particulier par le biais d'interactions à plusieurs étapes et lorsque les individus n'ont pas conscience de cette influence ou ne sont pas en mesure de la détecter de manière raisonnable. **De telles capacités pourraient porter atteinte aux processus démocratiques et aux droits fondamentaux, y compris l'exploitation fondée sur des caractéristiques protégées**.

Dans le cas de la manipulation nuisible (*harmful manipulation*), l'objet de l'évaluation est encore plus vaste, car c'est bien le comportement du modèle, mais également du système interagissant avec les usagers, et plus encore les vulnérabilités individuelles et collectives ainsi que les réactions évolutives de la société et des individus la composant, dans leurs interactions, dans la durée, avec l'IA, qu'il conviendra d'étudier.

*

Des distinctions importantes : modèles *versus* systèmes, usage général *versus* usage spécifique

Le droit du numérique ne saisit les systèmes d'IA générative, que s'il considère que le modèle, le système ou la pratique permise par le modèle ou le système, présentent un risque suffisant, au regard d'une pyramide des risques allant du risque inacceptable au faible risque, pour entraîner le déclenchement d'une interdiction ou d'une obligation sanctionnée. Finalement, les systèmes d'IA générative ne sont pas systématiquement régis par le droit du numérique en général, et par le RIA en particulier.

Les systèmes d'intelligence artificielle constituent l'une des notions centrales du RIA[398] qui reprend une définition de l'Organisation de coopération et de développement économiques (OCDE)[399]. Il apparaît ainsi que toutes les IA sont des IA génératives, dès lors que toute sortie de l'IA est considérée par ces textes comme une *« génération »* de prédictions, de contenus, de recommandations, de décisions.

La notion de modèle d'IA constitue l'autre notion centrale du RIA[400] : le modèle d'IA est le composant principal du système d'IA, un système d'IA pouvant, à cet égard, être composé d'un ou de plusieurs modèles, auxquels viendront s'ajouter d'autres composants techniques.

Ainsi, si les IA génératives peuvent s'incarner dans de nombreuses applications, plusieurs précisions importantes, à l'interface de l'informatique et du droit, doivent être formulées.

---

[398] RIA. Article 3, paragraphe 1 : « système d'IA », un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels ».

[399] OCDE. « Explanatory Memorandum on the Updated OECD Definition of an AI System ». *OECD Artificial Intelligence Papers*, Mars 2024.

[400] Considérant 97 du RIA : « Bien que les modèles d'IA soient des composants essentiels des systèmes d'IA, ils ne constituent pas à eux seuls des systèmes d'IA. Les modèles d'IA nécessitent l'ajout d'autres composants, tels qu'une interface utilisateur, pour devenir des systèmes d'IA. Les modèles d'IA sont généralement intégrés dans des systèmes d'IA et en font partie ».

Les IA génératives sont tantôt des systèmes d'IA à usage général, tantôt des systèmes d'IA à usage spécifique.

Les « *modèles de base* » (*Foundation Models*) selon la définition de l'Université de Stanford, sont des modèles d'IA à grande échelle, souvent de type « transformers »[401], entraînés sur des quantités massives de données variées et diversifiées, qui peuvent être adaptés pour effectuer une grande variété de tâches, servant ainsi de « base » à de nombreuses applications différentes. Ce terme met l'accent à la fois sur leur échelle (généralement des milliards de paramètres) et sur leur rôle de base unique et polyvalente capable d'alimenter de nombreuses applications d'IA.

- Lorsque l'IA générative est un système d'IA à usage général servant à une multitude d'usages distincts (ex : ChatGPT, Claude, Gemini), cela signifie que son composant principal est également un modèle d'IA à usage général.

À cet égard, le modèle d'IA à usage général est défini par le RIA[402] comme étant « *un modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur mise sur le marché* ». **Le système d'IA à usage général** correspond à *« un système d'IA qui est fondé sur un modèle d'IA à usage général et qui a la capacité de répondre à diverses finalités, tant pour une utilisation directe que pour une intégration dans d'autres systèmes d'IA* »[403].

La distinction entre systèmes d'intelligence artificielle à usage général et systèmes d'intelligence artificielle à usage spécifique repose sur l'étendue des cas d'usage envisagés lors de leur conception et tient pour une large part à la nature, au volume et à la diversité des données mobilisées lors de leur entraînement. Un système d'IA à usage général se caractérise par sa capacité à exécuter une large gamme de tâches distinctes, souvent non anticipées explicitement par ses concepteurs, et à être intégré ou adapté par des tiers dans des contextes d'application variés, qu'il s'agisse de la génération de texte, d'image, de code ou de tout autre contenu, sans finalité prédéterminée unique. Un système d'IA à usage général est donc entraîné sur des corpus de données massifs et hétérogènes, couvrant un large spectre de domaines, de formats et de sources. Cette diversité des données lui confère une capacité à généraliser et à être adapté ou réutilisé par des tiers dans des contextes d'application multiples.

Un système d'IA à usage spécifique est conçu et entraîné pour répondre à un ensemble de requêtes circonscrites, tel que le diagnostic médical. Un système d'IA à usage spécifique repose donc sur des jeux de données ciblés, sélectionnés et souvent annotés en fonction d'une tâche précise et d'un domaine circonscrit, tel qu'un corpus de dossiers médicaux pour un outil de diagnostic. Cette spécialisation des données d'entraînement restreint corrélativement le champ fonctionnel du système, mais permet généralement d'en optimiser la performance et la fiabilité sur la tâche visée.

---

[401] VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, *et al.* Attention is all you need. *Advances in neural information processing systems*, 2017, vol. 30.
[402] Article 3, paragraphe 63 RIA.
[403] Article 3, paragraphe 66 du RIA.

- Lorsque l'IA générative est un système d'IA à usage spécifique, ne servant qu'à un usage donné (par exemple, un compagnon IA de type Character.AI, une IA à visée médicale telle que Symptom.AI), cela ne signifie pas nécessairement que son composant principal soit un modèle d'IA à usage spécifique. Le composant principal du système spécifique, peut, en effet, tantôt être un modèle d'IA à usage général, tantôt un modèle d'IA à usage spécifique.

Dans le premier cas, la spécialisation est apportée par le système lui-même (données, outils, règles métier, orchestration, guardrails, *etc.*). Ces systèmes sont parfois qualifiés de systèmes d'IA hybrides, car le modèle d'IA à usage général qui sous-tend le système est complété par des composants spécialisés, dès lors que cela apporte un gain en performance, en fiabilité ou en efficacité.

Dans le second cas, la spécialisation est apportée par le modèle. En miroir de ce qui vient d'être énoncé, les systèmes d'IA générative, intégrant comme composant principal un ou plusieurs modèles d'IA à usage général, seront donc, tantôt des systèmes d'IA à usage général (de type ChatGPT, Claude ou Gemini), tantôt des systèmes d'IA à usage spécifique (IA compagnon ou IA médicale par exemple).

*

Les usages particuliers d'une IA générative (comme la sollicitation par un usager d'un compagnonnage affectif, la sollicitation de conseils médicaux, financiers, politiques ou encore la production de contenus pédopornographiques) peuvent tantôt être le fait d'un système d'IA à usage général - qui par essence permet tous les usages spécifiques - tantôt celui d'un système d'IA à usage spécifique - qui par essence ne permet que certains usages spécifiques.

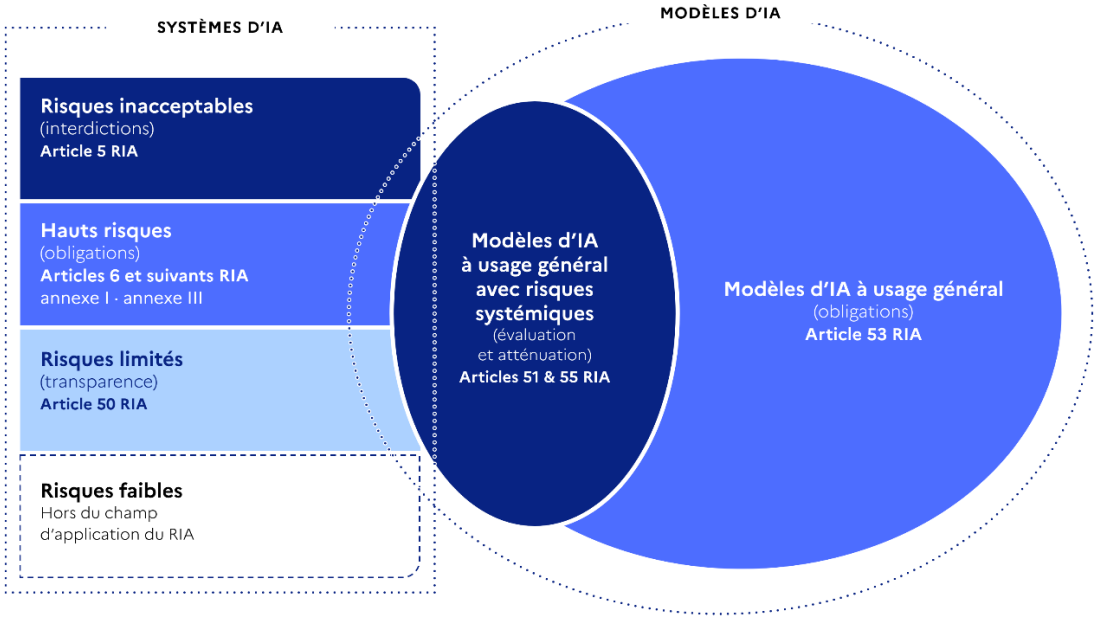
Une difficulté de qualification réside dans le fait qu'un utilisateur peut utiliser un système d'IA générative à des fins de compagnonnage avec un système à usage général (ex. : ChatGPT détourné) ou des services spécialisés (comme Replika ou Character.AI). **Il est ce faisant très difficile de saisir l'IA générative en tant que telle ou au travers du prisme des seuls systèmes d'IA à usage spécifique problématiques** (compagnons IA, IA à visée médicale ou financière *etc.*), **alors que des systèmes d'IA à usage général permettront les mêmes usages spécifiques problématiques. Ce premier constat oblige à traiter la question des vulnérabilités au regard des usages spécifiques problématiques des IA génératives**, qu'elles soient à usage général ou spécifique, et non au regard les différents types de systèmes d'IA.

En outre, les IA génératives, qu'elles soient des systèmes d'IA à usage général ou à usage particulier, ont nécessairement comme composant principal un modèle d'IA, qui sera dans de nombreuses hypothèses, à usage général. **Toute la difficulté, dans ce cas, sera de déterminer si l'exacerbation de vulnérabilités provient du système ou du modèle, en sachant que des garde-fous pour éviter certains risques peuvent être, soit placés directement dans le modèle** (IA constitutionnelle ou raisonnement socratique type *chain-of-thought* intégrées dans le modèle), **soit ajoutés comme une brique complémentaire du système**. Cette détermination sera d'autant plus importante, que le système et le modèle ne sont pas nécessairement fournis par le même fournisseur. **Ce second constat oblige à saisir la question des vulnérabilités engendrées par les IA génératives, non seulement sous l'angle des systèmes d'IA, mais également sous l'angle des modèles qui composent les systèmes.**

*

À ces complexités techniques, s'ajoute un deuxième niveau de complexité, d'ordre juridique, découlant du fait que **toutes les exacerbations de vulnérabilités engendrées par les usages des IA génératives n'ont pas encore, été identifiées à ce stade, par les dispositions du RIA, comme étant des risques devant déclencher des obligations et des interdictions**.

Seules les IA génératives présentant un risque d'ores et déjà identifié par le législateur font actuellement l'objet d'obligations ou d'interdictions au sein du RIA, tel que modifié par l'Omnibus numérique en matière d'IA.



Annexe 3 : Les défis économiques de l'IA générative

Le modèle économique de l'IA générative risque de reproduire les dérives des plateformes numériques.

S'interroger sur les modèles économiques de l'IA générative est crucial pour comprendre les risques qui en découlent

L'IA générative se démocratise à une vitesse inédite : son adoption est plus rapide que celle de l'ordinateur personnel ou d'Internet[404] et près de la moitié des Français l'utilisent déjà trois ans seulement après son apparition[405]. Cette facilité d'adoption de l'IA générative rend la comparaison avec les plateformes numériques (réseaux sociaux, moteurs de recherche) encore plus pertinente que celle avec l'équipement en ordinateurs ou en *smartphones* qui est faite fréquemment. C'est en effet la gratuité de certaines versions et l'absence de friction à l'entrée (autre que l'équipement premier permettant d'y accéder) qui favorisent cette rapidité de diffusion, plus rapide encore que lors du lancement des réseaux sociaux. Or, ce sont ces mêmes plateformes numériques qui ont permis l'émergence de l'économie de l'attention et de ses dérives[406].

Derrière l'apparente gratuité des outils d'IA générative – notamment des IA conversationnelles proposés en version freemium – se cache une réalité économique implacable : ces technologies coûtent extrêmement cher et leur viabilité dépend désormais d'une recherche urgente de rentabilité.

Le paradoxe est ainsi posé : sous une gratuité apparente se loge un coût réel, dont le financement reste une question ouverte. Il convient donc de rappeler les leçons tirées des réseaux sociaux, avant de présenter la réalité économique qui sous-tend aujourd'hui l'IA générative ainsi que l'émergence des modèles d'affaires[407] qui pourrait permettre de rendre la diffusion de l'IA générative viable afin d'anticiper l'émergence de ces dérives et promouvoir des modèles d'affaires vertueux[408].

[404] « Generative AI reached 53% adoption in three years, faster than the personal computer or the internet ». Artificial Intelligence Index Report. Stanford. 2026.
[405] Arcep, Arcom, CGE, ANCT. Baromètre du numérique – édition 2026. Février 2026. *« Près de la moitié des Français utilisent l'IA générative 3 ans seulement après son apparition. Cette vitesse d'adoption est inédite : il a fallu 7 ans aux réseaux sociaux et à l'équipement en smartphones pour atteindre ce seuil. »*
[406] Direction Générale du Trésor. « L'économie de l'attention à l'ère du numérique ». 4 septembre 2025.
[407] Le modèle d'affaires (ou modèle économique) décrit la manière dont une entreprise crée de la valeur, génère des revenus et des profits ; il se distingue des sources de revenus qui le composent (cf. *infra*).
[408] À la différence d'un modèle d'affaire *viable*, c'est-à-dire seulement capable de couvrir durablement ses coûts, un modèle d'affaire *vertueux*, ne vise pas à exploiter les vulnérabilités des utilisateurs. C'est précisément l'écart entre ces deux exigences, mis en évidence par les réseaux sociaux, que cette annexe cherche à anticiper pour l'IA générative

- **Les réseaux sociaux et l'économie de l'attention offrent un contre-exemple à ne pas reproduire**

Moteurs de recherche, réseaux sociaux, *streaming* vidéo, messageries, cartographie et navigation GPS : ces services financent leur gratuité en vendant aux annonceurs l'attention de leurs utilisateurs, acquise pendant qu'ils consomment le service ; plus un utilisateur y reste engagé, plus il génère de données exploitables pour un ciblage publicitaire précis[409]. **C'est la logique de l'économie de l'attention. Deux registres de dérives en découlent, aujourd'hui bien documentés : celles liées au contenant**, c'est-à-dire aux mécanismes de conception de la plateforme eux-mêmes, et **celles liées au contenu**, c'est-à-dire à la nature de ce qui y circule et à sa diffusion.

(i) Le contenant : les *dark patterns* au service de la captation de l'attention

Pour maximiser l'engagement et donc les revenus publicitaires, les plateformes recourent à des interfaces conçues pour capter l'attention au-delà de l'intention de l'utilisateur : algorithmes personnalisant des contenus destinés à générer des émotions fortes, interfaces trompeuses, notifications fréquentes. Ces techniques, connues sous le nom de *dark patterns*, sont devenues un instrument central du modèle économique des plateformes.

Cette logique s'accompagne d'une irresponsabilité structurelle : en subordonnant la modération des contenus à ses effets sur l'engagement et donc sur les revenus, les plateformes n'ont pas intérêt à un internet plus sûr[410], alors que les risques juridiques et réputationnels qu'elles encourent restent limités[411].

(ii) Le contenu : de l'*enshittification* à la prolifération des contenus choquants

Une fois les effets de réseau installés, les utilisateurs ont beaucoup de difficulté à quitter une plateforme même quand sa qualité se dégrade : c'est le processus qu'on appelle l'*enshittification*[412], qui s'est notamment illustré avec l'augmentation de la publicité sur Instagram. Comme le montrent théoriquement Choi et Jeon[413], c'est le pouvoir de marché de ces acteurs qui rend possible la maximisation de l'attention des utilisateurs dans l'intérêt des annonceurs et donc de leurs propres revenus, au détriment de la qualité de service.

Cette même logique d'engagement maximal favorise la visibilité des contenus les plus clivants ou choquants, seuls capables de générer les émotions fortes recherchées par les algorithmes de recommandation : incitation à des comportements dangereux pour soi-même, comme l'anorexie, ou pour autrui, *via* la mise en avant de pratiques violentes ou interdites ; polarisation des opinions ; agressivité, *etc.*

[409] D'où le raccourci conduisant à dire que c'est l'utilisateur lui-même qui est vendu : *« si c'est gratuit, c'est toi le produit »*.
[410] Cf. encore récemment les limites de l'action publique en cas de présence d'objets interdits sur les sites d'e-commerce.
[411] Certes, le RSN permet de s'attaquer aux « risques systémiques » mais les mesures récentes d'interdiction d'accès des enfants témoignent des limites de la méthode.
[412] Le Monde. « La « merdification », ce processus inexorable de dégradation de la qualité des services sur les plateformes numériques ». 24 avril 2025.
[413] CHOI Jay Pil, JEON Don-Shin. « Platform design biases in ad-funded two-sided markets ». *The Rand Journal of Economics*. 2023. vol. 54.

Quant à la publicité comme moteur économique en contrepartie de la gratuité, elle est susceptible de financer la désinformation[414], des sites illicites[415] et les pratiques dangereuses. De plus, comme l'ont montré toute une série de scandales, les risques réputationnels et économiques étant minimes, les acteurs sont peu incités à être regardants sur la publicité illicite, voire sur des arnaques[416].

Le modèle économique de l'IA générative reste incertain

Les grands services d'IA générative sont fournis à partir d'infrastructures lourdes (les centres de données ou *datacenters*) et d'énormes quantités de données pour développer et mettre au point des modèles très coûteux, notamment en raison des coûts d'infrastructure (qui représentent aujourd'hui l'essentiel du coût de production d'un *token*), d'entraînement et d'inférence[417] pour répondre aux requêtes des utilisateurs. Les services obtenus à partir de ces modèles peuvent à leur tour avoir besoin d'une très importante puissance de calcul.

Contrairement aux plateformes numériques du début du XXIe siècle qui pouvaient se développer avec des moyens limités « dans un garage » et passer à l'échelle avec des coûts marginaux très faibles voire quasi nuls[418], nous sommes face à des services très intensifs en capital dont les dynamiques de prix, de concurrence, de financement et de demande font douter que la rentabilité soit au rendez-vous pour tous les intervenants[419].

- **L'économie de l'IA générative se décompose en plusieurs strates**

La chaîne de valeur

Sur le plan strictement fonctionnel, cette chaîne de valeur peut se résumer, en simplifiant, comme une transformation de l'énergie en contenu : l'association des puces (*chips*) et de leur hébergement (*hosting*) donne accès à la capacité de calcul et à l'électricité nécessaires pour entraîner et faire fonctionner un modèle d'IA générative ; ce dernier convertit cette puissance de calcul en *tokens*[420] ; enfin, la couche applications (*apps*) convertit ces *tokens* en contenu susceptible d'avoir une valeur pour l'utilisateur final.

En pratique, les acteurs se positionnent différemment le long de cette chaîne. En amont, les fabricants de puces (Nvidia, AMD, Intel) et les fournisseurs de *cloud* (*hyperscalers* comme AWS, Google Cloud, Microsoft Azure, ou acteurs spécialisés dits « neoclouds », qui ne fournissent que de

[414] AHMAD Wajeeha., SEN Ananya, EESLEY Charles et al. « Companies inadvertently fund online misinformation despite consumer backlash ». *Nature*. 2024. vol 630, 123–131.
[415] The Athletic. « World Cup illegal streaming sites have ad revenue cut off in latest crackdown ». 6 octobre 2026.
[416] Jeff HORWITZ. « Meta is earning a fortune on a deluge of fraudulent ads, documents show ». *Reuters*. 6 novembre 2025.
[417] L'inférence désigne la phase d'utilisation d'un modèle déjà entraîné, au cours de laquelle il traite une requête pour produire une réponse (cf. infra, « Structure des coûts »).
[418] Quand Facebook acquiert WhatsApp en 2014 pour 22 milliards de dollars, il y avait 450 millions d'utilisateurs actifs par mois dont 72 % quotidiennement et 32 salariés ! Quant à Instagram il y avait 13 salariés et 30 millions d'utilisateurs. L'acquisition a coûté 1 milliard de dollars.
[419] Cf. parmi d'autres : The New York Times. « Big Tech's A.I. Spending Keeps Rising. So Do the Jitters ». 30 Juillet 2026. ; Frédéric Lordon « IA krach ? ». 1er juillet 2026. ; Joachim KLEMENT « The impossible maths of the AI boom ». *Financial Times*. 19 Mai 2026.
[420] Les *tokens* sont les unités élémentaires (fragments de mots) que traite un modèle de langage, en entrée comme en sortie ; ils servent d'unité de mesure de l'usage et de facturation (cf. infra, « Structure des coûts »).

la puissance de calcul GPU, comme CoreWeave ou, en Europe, Mistral AI) conditionnent l'accès de la plupart des acteurs, à l'exception de la puissance de calcul dont disposent les fournisseurs de modèles eux-mêmes ou les entreprises qui hébergent elles-mêmes des modèles, notamment ouverts (*self-hosting*)[421].

Sur les couches modèle et application, les grands acteurs du numérique sont souvent intégrés verticalement sur l'ensemble de la chaîne (Alphabet, Microsoft) ou plus ponctuellement (Meta), tandis que les développeurs de modèles (Anthropic, Mistral AI, OpenAI) leur sont fréquemment liés par des partenariats capitalistiques qui brouillent la frontière entre nouveaux entrants et acteurs installés : historiquement OpenAI avec Microsoft, Anthropic avec Amazon et Google, et plus récemment des financements qualifiés par la BRI de « circulaires » entre les acteurs de la chaîne de valeur[422] avec Nvidia au cœur du système de financement. Les grands fournisseurs de modèles disposent désormais de leur infrastructure d'inférence, en propre ou *via* des capacités dédiées chez leurs partenaires.

Par « IA générative grand public », on désigne les applications tournées vers un utilisateur non professionnel et qui s'appuient sur un ou plusieurs modèles de langage (ChatGPT, Claude, Gemini, Meta AI, Vibe, *etc.*).

L'intégration verticale des entreprises qui proposent des systèmes d'IA générative grand public les expose à la pression concurrentielle sur les deux niveaux simultanément : sur le modèle, *via* la concurrence des modèles *open source* et sur l'application, *via* la multiplication des interfaces concurrentes. À la différence des réseaux sociaux, dont les effets de réseau verrouillent les utilisateurs, les services d'IA générative restent à ce stade largement substituables, ce qui renforce cette pression.

Face à ces pressions, les développeurs de modèles répondent par deux mouvements concomitants. D'un côté, l'intégration vers l'aval (*downstream*) : développement d'applications verticales propriétaires dans des secteurs à haute valeur (droit, santé, finance). De l'autre, l'intégration vers l'amont (*upstream*) *via* des investissements massifs dans l'infrastructure de calcul[423] visant à réduire la dépendance aux fournisseurs d'hébergement tiers et à sécuriser le substrat physique de leur compétitivité[424].

Cette chaîne de valeur permet de situer où se logent les coûts détaillés dans la section suivante (« Structure des coûts ») : l'entraînement et l'inférence se rattachent tous deux à la couche modèle, tandis que la couche puces et hébergement en constitue le substrat physique commun.

Les acteurs majeurs

En septembre 2026, quatre entreprises américaines concentrent l'essentiel du marché de l'IA générative grand public : OpenAI, Anthropic, Google, Microsoft[425].

---

[421] Autorité de la concurrence. Avis 24-A-05 du 28 juin 2024 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative.
[422] Phurichai Rungcharoenkitkul, 2026. « The AI investment race » BIS Working Papers 1367, Bank for International Settlements.
[423] À titre d'exemple pharaonique, le projet Stargate à 500 milliards de dollars. *L'Usine digitale*. « Stargate : le projet IA à 500 milliards de dollars signé Donald Trump ». 22 janvier 2025.
[424] ZeroHedge. « Nadella's Hedge: Microsoft Wants To Make AI Models Cheap - Then Own The Rails They Run On ». 22 juin 2026.
[425] First Page Sage. « Top Generative AI Chatbots by Market Share ». Octobre 2026.

Cette concentration est le produit d'une stratégie délibérée : la course à l'utilisateur *via* la gratuité. Depuis novembre 2022 (lancement de ChatGPT), les acteurs ont absorbé des coûts d'infrastructure colossaux afin d'acquérir des bases d'utilisateurs massives avant toute monétisation. En atteignant 100 millions d'utilisateurs en 60 jours (contre 28 mois pour Instagram), ChatGPT a posé les bases d'un « effet de réseau »[426] qui semblait, il y a quelques mois encore, difficile à inverser. Aujourd'hui, OpenAI semble pris en étau entre les stratégies généralistes de Google et Microsoft d'une part, et d'Anthropic d'autre part. Pour autant, aucun de ces acteurs n'a jusqu'à présent fait la preuve de sa capacité à rentabiliser des investissements gigantesques.

Ces acteurs investissent des sommes colossales pour tenir ce rythme : le CAPEX (ou dépenses d'investissement en capital) cumulé des cinq principaux *hyperscalers* américains[427] dépasse déjà 1 600 milliards de dollars depuis 2022, et devrait encore atteindre 660 à 690 milliards pour la seule année 2026[428]. Ces montants sont sans commune mesure avec les revenus dégagés à ce stade par les développeurs de modèles. À titre d'exemple, OpenAI aurait ainsi perdu près de neuf milliards de dollars en 2025 pour 13,5 milliards de chiffre d'affaires[429] ; quant aux chiffres de 2026 ils sont largement discutés. *The Economist* insistait récemment sur la fragilité des montants annoncés par la plupart des acteurs[430]. De son côté, Anthropic frôlerait l'équilibre opérationnel ces deux derniers trimestres avec un chiffre d'affaires en très forte croissance de 33 milliards de dollars depuis le début de l'année 2026. Le document d'introduction en Bourse d'Anthropic (prévue d'ici à la fin 2026) totalise par ailleurs des engagements d'achat et de location d'environ 900 milliards de dollars, dont la quasi-totalité concerne les capacités de calcul et les infrastructures informatiques dont près de la moitié d'ici 2030, ce qui donne la mesure du gap de chiffre d'affaires à réaliser.

Ces capitaux alimentent aussi une intense dynamique de rachats et de prises de participations croisées entre acteurs, par laquelle les plus solides cherchent à renforcer leur position sur la chaîne de valeur (cf. *supra*) plutôt qu'à se contenter de la seule concurrence par les produits.

Ce tableau d'ensemble révèle une asymétrie fondamentale : du côté de la demande, une adoption de masse en voie de consolidation accélérée dans des conditions de rémunération fragiles ; du côté de l'offre, des modèles d'affaires structurellement déficitaires dont la pérennité reste à construire.

Les modèles (à poids) ouverts

Le modèle dit *open weight*, c'est-à-dire dont les poids ou les paramètres internes – les coefficients numériques spécifiques obtenus lors de la phase d'entraînement du modèle de fondation – sont rendus publiquement accessibles, permet à quiconque de les télécharger, de les exécuter localement ou de les réentraîner sur des données spécifiques pour des tâches ciblées[431]. Il exerce à ce titre une pression directe sur la chaîne de valeur (cf. *supra*) et sur les prix pratiqués par les acteurs propriétaires. Les éditeurs de modèles ouverts ne monétisent donc pas l'accès au modèle lui-même, mais les services qui l'entourent (hébergement et API, versions commerciales plus performantes,

[426] Il n'y a pas à proprement parler, à ce stade, d'effet de réseau (indirects) comme pour les plateformes ; pour autant, l'augmentation du nombre d'utilisateurs permet d'augmentation de l'usage et des données permettant l'amélioration de la qualité des réponses.
[427] Google, Amazon, Meta, Microsoft et Oracle.
[428] Visual Capitalist. « Charted: The $448B AI Spending Surge by Big Tech ». 20 avril 2026.
[429] Fortune. « OpenAI says it plans to report stunning annual losses through 2028—and then turn wildly profitable just two years later ». 12 novembre 2025.
[430] The Economist. « The AI boom is threatened by an epidemic of unreliable numbers ». 17 octobre 2026.
[431] On parle alors de *fine-tuning* ou réglage fin.

intégration et support), ou poursuivent des objectifs stratégiques (diffusion d'un standard, vente de matériel). Leurs revenus sont très inférieurs à ceux des entreprises américaines.

Le centre de gravité de ce type de modèle est plutôt chinois. Toutefois, d'autres acteurs, comme Nvidia, Microsoft, Meta aux États-Unis ou Mistral en Europe, chacun ayant des intérêts et des logiques stratégiques propres, défendent l'ouverture – certes partielle – des modèles. Au-delà du choc DeepSeek de janvier 2025 (performances comparables aux modèles frontière américains pour un coût d'entraînement déclaré d'environ six millions de dollars, contre plusieurs centaines de millions pour GPT-4)[432], Alibaba domine aujourd'hui l'écosystème mondial des modèles ouverts[433] avec sa famille Qwen et en août 2026 ses plus de deux milliards de téléchargements cumulés, selon Hugging Face[434]. Cette stratégie répond à une logique autant géopolitique qu'économique[435] : favoriser le développement de l'usage des modèles chinois tout en faisant pression sur les marges des acteurs américains, dont l'usage par API est facturé jusqu'à 50 fois plus cher avec des modèles propriétaires[436].

Le mouvement a suscité deux réponses opposées : Meta, pionnier de l'ouverture avec Llama, a au contraire durci ses licences par crainte de voir des concurrents s'appuyer sur ses poids ouverts avant de revenir plus fréquemment à des modèles plus ouverts, tandis que Mistral AI et OpenAI ont choisi d'ouvrir davantage certains modèles (Mistral Large 3[437], gpt-oss[438]).

Cette dynamique exerce une pression à la baisse sur les prix des modèles propriétaires, la valeur captée se déplaçant vers les services associés (conseil, intégration, hébergement managé), et pose un défi réglementaire : une fois les poids publiés, il devient difficile d'encadrer l'usage qui en est fait[439], y compris pour contourner les garde-fous mis en place par les développeurs[440].

- **Une structure des coûts à forte intensité capitalistique : l'économie de l'IA générative est marquée par des coûts fixes et des coûts marginaux importants**

Avant d'aborder les modèles d'affaires de l'IA générative, il convient d'en comprendre la structure de coûts, qui en conditionne la viabilité. Celle-ci comporte des coûts fixes : l'entraînement du modèle, préalable à tout usage, et, sur l'ensemble du cycle de vie, les infrastructures de calcul (GPU), dont la durée d'amortissement constitue un enjeu clé ; mais elle se distingue surtout, structurellement, de celle des plateformes numériques qui l'ont précédée. Alors que celles-ci reposaient sur un coût marginal par utilisateur quasi nul, chaque réponse générée par un modèle d'IA a un coût réel, l'inférence, détaillé *infra*.

[432] ITSocial. « DeepSeek-V3 redéfinit le marché de l'IA et bouscule son modèle économique en construction ». 29 janvier 2025.
[433] Particulièrement s'agissant des grands modèles vision-langage.
[434] Hugging Face. « State of Open Models: Summer 2026 ».
[435] Cette stratégie s'appuie semble-t-il sur la maîtrise de la distillation à grande échelle, qui permet aux laboratoires chinois de reproduire à moindre coût les performances des modèles fermés américains.
[436] AgentLand. « DeepSeek-V4 fait sauter les prix avec 1M tokens open source ». 24 avril 2026.
[437] Mistral. Introducing Mistral 3. 2 décembre 2025.
[438] OpenAI. Lancement de gpt-oss. 5 août 2025.
[439] « Demis Hassabis has a plan to harness AI safely ». 14 juillet 2026
[440] IBM. « Risques liés à l'open source : les dangers croissants de l'IA générative non réglementée ».

C'est cette différence qui structure l'ensemble des modèles d'affaires présentés *infra*. Tant que ce coût marginal reste incertain ou significatif, la question de savoir si, et comment, la gratuité de l'IA générative pour le grand public peut être financée demeure ouverte.

L'entraînement, un coût fixe croissant

L'entraînement représente un coût initial, massif et ponctuel, de construction du modèle à partir des données. Le coût amorti (matériel et énergie) de l'entraînement d'un modèle frontière croît d'un facteur 2,4 par an depuis 2016, une dynamique qui s'est encore accélérée depuis 2020 (facteur 3,5 par an, soit un doublement tous les sept mois)[441] : à ce rythme, les plus gros entraînements devraient dépasser le milliard de dollars dès 2027[442]. Ce coût est fixe : il ne croît pas avec l'usage une fois le modèle déployé, mais doit être réengagé à chaque nouvelle version, et mobilise, lors de la phase d'entraînement, une consommation électrique intense et concentrée dans le temps[443]. Les éditeurs restent toutefois très discrets sur l'évolution de ces coûts, dont la part relative tendrait à diminuer à mesure que les GPU sont mutualisés entre entraînement et inférence.

L'inférence à l'origine d'un coût marginal

L'inférence est le processus par lequel un modèle d'IA utilise ses connaissances d'abord pour traiter une requête utilisateur (*tokens* entrants), ensuite pour générer une réponse (*tokens* sortants), ces deux composantes pesant toutes deux dans le coût. Le processus est d'autant plus coûteux que les calculs réalisés par le modèle pour répondre au prompt sont complexes. C'est le coût variable par excellence, le coût marginal au sens économique du terme, proportionnel au volume d'utilisation et mesuré en *tokens* (en entrée comme en sortie). C'est cette composante qui conditionne directement la marge brute (contribution marginale)[444] de chaque requête traitée, et donc si chaque requête supplémentaire rapproche ou éloigne l'opérateur de l'équilibre. Pour donner un ordre de grandeur, une estimation largement reprise depuis 2023 indiquait qu'un échange simple (sans *chain-of-thought*) avec un agent conversationnel coûtait environ dix fois plus en énergie de calcul qu'une requête Google Search classique[445]. Aujourd'hui l'écart de coût énergétique est presque nul, mais il demeure pour les requêtes longues ou à raisonnement explicite, plus consommatrices en *tokens*[446]. L'énergie ne représente toutefois qu'une faible part du coût total, l'essentiel relevant de l'amortissement du matériel de calcul[447]. Le coût que représente l'inférence d'un modèle étant important, il ne peut être négligé dans l'adoption d'un modèle d'affaires viable. Ce point met en exergue une différence notable avec la structure des coûts des plateformes numériques dont le coût marginal est quasi nul[448].

[441] Epoch AI. Trends in AI. 5 février 2026.
[442] COTTIER Ben, RAHMAN Robi, FATTORINI Loredana et al. The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*. 2024.
[443] Epoch AI. *Op.cit.*
[444] La marge brute (*contribution margin*) isole le rapport entre le prix facturé et le seul coût variable d'inférence.
[445] Reuters. « Focus: For tech giants, AI like Bing and Bard poses billion-dollar search problem ». 22 février 2023.
[446] Epoch AI.*Op.cit.*
[447] Selon Epoch AI, l'énergie représente 2 à 6 % du coût amorti de développement des modèles frontière (GPT-3, OPT-175B, GPT-4, Gemini Ultra), l'essentiel relevant du matériel et du personnel : COTTIER Ben et al., « How Much Does It Cost to Train Frontier AI Models? ». Epoch AI. 2024.
[448] ROCHET Jean-Charles et TIROLE Jean. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 2003, vol. 1, Issue 4, pp. 990-1029.

La baisse du coût marginal du *token* et l'effet rebond

La trajectoire du coût d'inférence par *token* est en forte baisse depuis 2023[449], sous l'effet combiné des optimisations logicielles (quantisation[450], distillation des modèles[451]), des améliorations matérielles et de l'intensification de la concurrence, notamment par des modèles chinois, qui permettent de réaliser l'inférence à un prix bien inférieur à celui pratiqué par les modèles américains[452]. Le coût de traitement d'un million de *tokens* a été divisé par un facteur supérieur à dix en moins de deux ans pour les modèles équivalents[453].

Cette tendance ne se traduit pas pour autant par une réduction équivalente des coûts totaux, en raison d'un effet rebond. À mesure que le *token* devient moins cher à produire, les modèles en génèrent davantage. Les architectures dites « de raisonnement » produisent des dizaines ou centaines de milliers de *tokens* internes avant de formuler une réponse, au prix d'un coût d'inférence par requête sans commune mesure avec les modèles précédents. Parallèlement, des utilisateurs plus nombreux, engagés dans des sessions plus longues et formulant des requêtes plus complexes, augmentent le volume total de calcul[454].

Dès lors, malgré la baisse du coût marginal par *token*, l'augmentation concomitante de la production de *tokens*, qui s'apparente à un effet rebond, empêche le coût marginal par requête, et le coût total, de diminuer dans les mêmes proportions, voire les conduit à augmenter. Une simple baisse mécanique du coût du *token* ne saurait donc suffire à résoudre la question du financement de la gratuité. Ainsi, le niveau auquel se stabilisera le coût marginal réel de l'inférence, et pour quels usages, demeure à ce stade largement incertain.

- **Les modèles d'affaires ne sont pas encore stabilisés, en particulier sur le versant B2C**

Aujourd'hui, les acteurs de l'IA générative explorent des sources de revenus variées, mais celles-ci restent sans commune mesure avec les coûts engendrés par l'entraînement et l'usage de ces modèles comme présenté *supra*.

Pour analyser cette dynamique, il est essentiel de distinguer ce qu'on appelle un modèle économique (ou modèle d'affaires), qui décrit comment une entreprise crée de la valeur, génère des revenus et des profits, des sources potentielles de revenus susceptibles d'être utilisées pour parvenir à construire un modèle économique. Les premiers relèvent de la stratégie globale des acteurs (pouvant combiner différents usages et différentes sources de revenus), tandis que les secondes concernent les mécanismes concrets de monétisation (abonnements, commissions, exploitation des données, *etc.*). **Il n'existe pas aujourd'hui d'entreprise d'IA générative ayant réussi à mettre en place un modèle d'affaires pérenne incontestable.** Même pour Anthropic, qui se concentre dans le B2B et apparait

[449] Par exemple, le coût (mesuré en $ par million de tokens) de GPT-4 (en 2023) était supérieur au coût du dernier modèle d'Open AI GPT-6.1 Sol (Septembre 2026) alors que les capacités de ce dernier sont massivement supérieures.
[450] Selon IBM, la quantisation est le processus qui consiste à réduire la précision d'un signal numérique, généralement d'un format de haute précision à un format de plus faible précision.
[451] Selon CaféTech, la distillation est une technique qui permet de transférer des capacités d'un *modèle* "enseignant" vers un *modèle* "élève", afin d'améliorer son efficacité sans sacrifier significativement sa précision.
[452] The Economist. « China is having another AI moment ». 21 juin 2026.
[453] BERGEAUD Antonin et GRASS Etienne. « L'augmentation des factures d'IA et la montée de l'*open weight* ». Laboratoire de l'innovation. Octobre 2026.
[454] *Ibid*.

actuellement comme le plus prometteur, les débats sur les risques et la fiabilité des prévisions sont nombreux. C'est la raison pour laquelle nous examinons, en particulier pour ses services grand public, au cœur de cette analyse, chaque source éventuelle de revenus, ses potentialités, mais aussi éventuellement les risques qu'elle est susceptible d'engendrer.

Pour lire ces différentes sources de revenus, il est utile de les distinguer par les usages, professionnels ou grand public. **D'un côté le B2B, où l'IA est vendue comme un outil professionnel qui est essentiellement facturé à l'usage (API, licences, intégration)**, auquel s'apparente le B2P (*business-to-public*), qui désigne spécifiquement sa vente aux administrations et organismes publics ; **de l'autre le B2C, où elle est proposée directement au grand public (abonnement, publicité, commissions)**. Du côté des acteurs eux-mêmes, la frontière est en réalité moins nette (en tout cas pour l'instant), presque tous combinant les deux approches. Ainsi, ils se distinguent davantage selon qu'ils disposent déjà d'un écosystème numérique établi (les *hyperscalers*) ou qu'ils sont de nouveaux entrants nés avec l'IA générative.

Pour autant la distinction B2B/B2P et B2C permet d'analyser plus simplement les risques qu'ils engendrent en termes de vulnérabilité. L'hypothèse implicite est qu'il est peu envisageable que les modèles d'IA conversationnelle grand public soient totalement financés par les ressources de l'IA générative à destination des professionnels. En tout état de cause, la rudesse de la concurrence dans les services aux entreprises et l'intensité capitalistique observée pour les services aux entreprises ne devraient pas permettre de financer les importants coûts, fixes comme variables, des services grand public.

A contrario, il est utile d'évoquer la **capacité, qu'auraient des acteurs numériques déjà en place, à subventionner leur offre d'IA générative grand public par les revenus de leurs activités historiques modernisées par l'introduction de l'IA** (*search*, systèmes d'exploitation, logiciels généralistes, réseaux sociaux). Cela représente un avantage structurel sur les nouveaux entrants exclusivement adossés à l'IA générative.

Les acteurs chinois (DeepSeek, Alibaba, Baidu, ByteDance) occupent une position à part dans cette dynamique. Comme présenté *supra,* leur stratégie repose surtout sur la diffusion de modèles ouverts utilisés en Europe ou aux États-Unis essentiellement par les entreprises, tout en déployant, pour certains d'entre eux (Qwen Chat d'Alibaba, Doubao de ByteDance, Ernie Bot de Baidu)[455], leurs propres applications grand public sur le marché domestique[456].

[455] South China Morning Post. « Alibaba's Qwen AI app hits 100 million users as upgrade broadens role in daily life ». 15 janvier 2026.
[456] South China Morning Post. « Baidu launches Ernie 5.0 as the firm's AI assistant users reach 200 million a month ». 22 janvier 2026.

Le segment B2B et B2P

Le marché professionnel, B2B au sens large, incluant le B2P (*business-to-public*)[457] constitue le principal terrain de monétisation des acteurs dont l'activité se concentre exclusivement sur l'IA générative. Leur modèle repose sur deux piliers complémentaires, l'usage et l'installation (celle-ci incluant les contrats gouvernementaux, où s'exprime le B2P) auxquels s'ajoute un troisième pilier encore expérimental, le partage de la valeur.

- *L'usage*

L'accès à l'API à la consommation (facturation par *token*) et les licences entreprise constituent la principale source de revenus[458], portée par une forte croissance des dépenses des entreprises en IA générative ces dernières années[459].

Cette facturation à l'usage rencontre toutefois ses limites côté clients : plusieurs entreprises, à l'image d'Uber, ont dû plafonner leurs dépenses en outils d'IA facturés au *token*, faute de pouvoir établir un lien clair entre consommation et livrée aux utilisateurs[460], révélateur d'une tension plus large entre facturation à l'usage et retour sur investissement perçu[461].

- *L'installation*

Au-delà de l'accès à l'API, le déploiement de modèles au sein des systèmes d'information des entreprises (intégration aux outils métier, déploiements sur site ou en cloud privé) représente une ligne de revenus distincte et croissante. Cette composante est particulièrement stratégique pour Mistral AI[462], dont l'offre souveraine et la compatibilité avec les contraintes réglementaires européennes constituent un argument commercial différenciant sur le marché français et européen.

Les contrats gouvernementaux représentent également un levier financier croissant, poussant les géants technologiques à adapter leurs offres pour répondre aux exigences des institutions. Google, Meta, OpenAI, Anthropic et xAI proposent désormais des versions « gouvernementales » de leurs solutions, conçues pour se conformer aux normes strictes de confidentialité, de sécurité et de performance requises par les administrations[463]. Ces contrats, souvent conclus avec des agences fédérales ou des organismes publics, offrent une source de revenus stable, tout en renforçant la légitimité des entreprises dans des secteurs sensibles.

[457] Il s'agit de la vente aux administrations et organismes publics.
[458] WANG Sarah, KAHL Justin et XU Shangda. « Leaders, gainers and unexpected winners in the Enterprise AI arms race ». *Andreesen Horowitz*, 30 janvier 2026.
[459] TULLY Tim, REDFERN Joff, DAS Deedy et XIAO Derek. « *2025: The State of Generative AI in the Enterprise* ». MENLO. 9 décembre 2025.
[460] Fortune. « Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it ». 26 mai 2026.
[461] BELLAN, Rebecca. « The token bill comes due: Inside the industry scramble to manage AI's runaway costs ». *TechCrunch*, 5 juin 2026.
[462] SERAMOUR, Célia. « Développer son "premier cluster européen d'IA" : Pourquoi Mistral AI lève 830 millions de dollars de dette ». *L'Usine digitale*. 30 mars 2026.
[463] Center for Democracy and Technology. « Risky Business: Advanced AI Companies' Race for Revenue ». 7 janvier 2026.

En France, Mistral AI illustre cette dynamique de partenariat : le gouvernement français a déployé un *chatbot* basé sur ses modèles pour équiper un million de fonctionnaires[464]. Comme l'indique Audrey Herblin-Stoop, senior vice-president, Global Affairs chez Mistral AI : *« l'essentiel de l'activité de Mistral AI reste le B2B. Nous laissons à nos clients, notamment les institutions, la liberté de choisir où et comment ils organisent le post-training et, par conséquent, les limites d'utilisation de nos modèles »* (audition du 11 mai 2026).

- *Le partage de la valeur*

Un troisième modèle, encore marginal, consiste à ne plus facturer l'usage du modèle mais à percevoir une part de la valeur qu'il contribue à créer chez le client (*pay per success*). OpenAI a ainsi évoqué publiquement cette hypothèse par la voix de sa directrice financière Sarah Friar, à Davos en janvier 2026 : l'entreprise pourrait percevoir une part des revenus générés par une découverte à laquelle ses modèles auraient contribué, par exemple un nouveau médicament, plutôt qu'une facturation classique à la consommation[465]. Il s'agit toutefois à ce stade d'une piste explorée publiquement, qui suscite des interrogations sur sa cohérence. En effet, si la valeur créée par le modèle était aussi importante que le suggère ce type d'accord, pourquoi les acteurs de l'IA continuent-ils, pour l'essentiel de leur activité, à facturer au *token* plutôt qu'au pourcentage de la valeur produite[466].

Le segment B2C

Les grandes entreprises technologiques disposent de bases d'utilisateurs grand public considérables et déploient plusieurs logiques de monétisation, souvent combinées. Plusieurs modèles sont testés à ce stade.

- *L'abonnement*

Le principe est bien connu : proposer un abonnement afin d'accéder au service. Cette formule, largement répandue dans de nombreux secteurs, vise à offrir une utilisation continue en échange d'un paiement régulier.

Aujourd'hui, une convergence tarifaire s'observe autour de 20 dollars par mois (ChatGPT Plus, Claude Pro, Copilot Pro, Gemini AI Pro)[467], avec une segmentation par intensité d'usage.

Les taux de conversion vers l'abonnement payant restent faibles au regard de la base d'utilisateurs gratuits[468], et varient sensiblement d'un acteur à l'autre. En effet, l'acteur qui gagne la bataille de l'audience n'est pas nécessairement celui qui gagne celle de la monétisation par utilisateur[469].

L'usage gratuit, majoritaire, générant des coûts sans revenus directs, l'abonnement reste un modèle complémentaire, aujourd'hui insuffisant à lui seul pour assurer la viabilité économique. Cette source

[464] Gouvernement français, IA dans l'État. « Lancement de l'expérimentation Mistral AI dans l'Assistant IA interministériel ». 17 novembre 2025.
[465] CNBC. « CNBC Transcript: OpenAI CFO Sarah Friar Speaks with CNBC's "Squawk Box" in Davos, Switzerland Today ». 21 janvier 2026.
[466] CNBC. « Palantir's Karp bashes OpenAI, Anthropic token model: "Something has gone completely wrong" ». 1er juillet 2026.
[467] Il existe également des abonnements « pro » tels que ChatGPT Pro ou Perplexity Max à plus de 100 dollars par mois qui ne sont pas à proprement parler des offres purement B2C.
[468] Roic.ai. « OpenAI's ChatGPT Surpasses 900 Million Weekly Users, 50 Million Paying Subscribers ». 27 février 2026.
[469] The Next Web. « Claude earns more revenue per user than ChatGPT, Sensor Tower finds ». 16 juin 2026.

de revenu soulève également la question d'un accès par tous et à moindre coût à un service de qualité.

- *Les objets connectés (ou « devices »)*

Les objets connectés intégrant un système ou un modèle d'IA générative émergent comme un levier pour les entreprises cherchant à diversifier leurs revenus et à ancrer leurs technologies dans le quotidien des utilisateurs. Cette approche se matérialise par la création de nouveaux marchés ou l'extension de services d'IA existants vers des supports physiques innovants. Parmi les exemples les plus emblématiques, les lunettes connectées de Meta[470], permettant un accès direct à son *chatbot*, illustrent bien cette intégration. Ces dispositifs posent des questions pressantes en matière de vie privée, notamment en raison de leur capacité à enregistrer en continu des données, qui pourraient ensuite être exploitées à des fins publicitaires. Parallèlement, les jouets IA se profilent comme un nouveau terrain financier pour les entreprises[471].

- *La valorisation des données cognitives au sein de l'écosystème*

Les fournisseurs d'IA conversationnelles ont accès à une source d'information bien plus riche que les moteurs de recherche classiques : là où une requête révèle une intention ponctuelle, une conversation dévoile des préoccupations, des hésitations, un contexte émotionnel. Cela constitue une matière dont ils sont en mesure d'extraire une valeur commerciale (intentions d'achat notamment), y compris au-delà de ce que l'utilisateur choisit délibérément de partager[472]. Pour les acteurs intégrés dans de vastes écosystèmes numériques (Google, Meta, Microsoft), cette valeur peut ensuite être exploitée au sein même de cet écosystème, par exemple sous forme de publicité ciblée sur leurs autres services : Meta a ainsi annoncé l'utilisation des conversations avec son assistant IA à cette fin (hors Union européenne)[473].

Pour les IA dits « compagnons » (Character.AI, Replika, *etc.*), dont les conversations longues et personnalisées atteignent une profondeur psychologique particulière[474], cette valorisation s'opère davantage auprès d'acteurs tiers : une étude récente a ainsi révélé que 17 des 20 principaux *chatbots* analysés transmettent, aux États-Unis, des données de conversation à des entreprises tierces[475].

- *La publicité conversationnelle*

L'entrée de la publicité dans les IA conversationnelles ne procède pas d'un modèle stabilisé. Elle avance par tâtonnements, expérimentations et ajustements successifs, avec des formats et des logiques qui varient selon les acteurs et les stades de déploiement. D'un côté, une réalité commerciale qui commence à s'installer ; de l'autre, une série d'hypothèses encore incertaines sur ce que sera la publicité conversationnelle dans les prochaines années.

---

[470] BBC. « Facebook owner unveils new AI-powered smart glasses ». 18 septembre 2025.
[471] Research and Markets. « Smart Toys Market Report 2026 ». Janvier 2026.
[472] STAAB Robin, VERO Mark, BALUNOVIC Mislav et al. Beyond memorization: Violating privacy via inference with large language models. *International Conference on Learning Representations*. 2024. pp. 33832-33878.
[473] VITARD, Alice. « Meta va utiliser les conversations avec son assistant IA pour le ciblage publicitaire ». *L'Usine digitale*. 2 octobre 2025.
[474] Fortune Business Insights. « AI Companion Market Size, Share, and Industry Analysis By Type (Text-Based, Voice-Based, and Multi-Model), By Application (Mental Health Support, Social Interaction, Education and Learning, Personal Assistance), By Industry (Consumer, Business, Healthcare, and Education), and Regional Forecast, 2026-2034 ». 21 septembre 2026.
[475] JAZLAN Muhammad, WANG Ethan, VEKARIA Yash, et SHAFIQ Zubair. Tracking Conversations: Measuring Content and Identity Exposure on AI Chatbots. *arXiv:2604.27438*. 2026.

- *Les formats intégrés à la réponse*

Le principe consiste à accoler à la réponse une annonce distincte et clairement étiquetée (« *Sponsored* »), sans la mêler au contenu généré : sous forme de bandeau en dessous du texte, ou générée dynamiquement selon le contexte de la conversation (Showroom Ads de Microsoft Copilot)[476]. Perplexity AI a été le premier acteur à l'expérimenter, dès 2024, avant de l'abandonner face aux résistances des utilisateurs[477]. OpenAI l'a depuis déployé à plus grande échelle, d'abord aux États-Unis en janvier 2026[478], puis en Europe à partir d'août 2026[479].

- *Les formats en cours de conversation ou l'hypothèse de la publicité déguisée*[480]

Bien que proches des formats intégrés par leur nature publicitaire, les formats en cours de conversation s'en distingueraient par leur capacité à s'insérer de manière fluide et pertinente dans le flux naturel de l'échange. Ils seraient susceptibles d'exploiter le contexte continu de l'échange pour insérer des messages commerciaux à des moments jugés pertinents : lors d'une expression de besoin, d'une comparaison de produits ou d'une demande de recommandation. Leur efficacité dépendrait alors de leur fluidité d'intégration dans le fil du dialogue[481], soulevant *de facto* des questions d'identification pour les utilisateurs.

Sur l'hypothèse de la publicité déguisée, des pratiques masquées d'influence commerciale et de toute forme de monétisation, il convient de maintenir dans la réflexion la possibilité d'orientations commerciales qui s'exerceraient sur les réponses sans étiquetage formel, comme conséquence structurelle d'accords commerciaux influençant l'entraînement ou l'alignement des modèles.

Selon une expérience académique récente où des participants ont interagi avec un assistant d'achat basé sur une IA conversationnelle, la faisabilité technique et l'efficacité de ces outils pour orienter discrètement les choix des consommateurs sont désormais démontrées[482]. Toutefois, à ce stade, aucune preuve ne permet d'affirmer que les entreprises exploitent ces capacités, ni dans quelles conditions elles pourraient le faire.

Cette hypothèse restera difficile à confirmer empiriquement, mais difficile à exclure. C'est précisément cette ambiguïté qui justifie une vigilance importante, en particulier pour les IA conversationnelles utilisées par des mineurs, qui ont un esprit critique moins développé pour remarquer ce type de publicité implicite.

[476] Formations Analytics. « Les nouvelles fonctionnalités de Microsoft Copilot : Showroom ads et filtres dynamiques ». 6 mars 2025.
[477] Perplexity. « Why we're experimenting with advertising ». 12 novembre 2024.
[478] CNBC. « OpenAI to begin testing ads on ChatGPT in the U.S. ». 16 janvier 2026.
[479] OpenAI. « ChatGPT Ads se déploie en Europe ». 18 août 2026.
[480] Nous retenons le terme « déguisée » pour désigner cette forme d'influence dans la mesure où il s'agit d'une notion établie en droit français et européen qui renvoie aux pratiques commerciales déloyales ou trompeuses.
[481] WebFX. « Inside Microsoft's Copilot Search Ads (And What Makes Them 25% More Effective Than Traditional Search Ads) ». 4 juin 2025.
[482] WERNER Tobias, SORAPERRA Ivan, CALVANO Emilio et al. Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection. *arXiv preprint arXiv:2409.12143*. 2024.

Annexe 4 : Liste des auditionnés

Auditions menées entre le 10 mars et le 10 juillet 2026

Organisme	Auditionnés et fonctions
AI Collaborative	• **M. TISNÉ Martin**, Président-directeur général
AI & Society Institute, École normale supérieure (ENS)	• **Mme POPA-FABRE Murielle**, Senior Tech & Policy Advisor
Anthropic	• **Mme DILÉ Claire**, Responsable des affaires extérieures pour la France
Artpak	• **M. DOSHI Jigar**, Director of machine learning • **M. SHAHED Rehan**, Project Manager GEN AI -LLMs Public Health Initiatives
Assistance publique des Hôpitaux de Paris AP-HP	• **M. MOUHOUD Théo**, Pédopsychiatre, Responsable de la maison des adolescents de l'Hôpital Avicenne • **Mme VERGERON Laure**, Psychiatre addictologue, Hôpital de Jour Adolescents et Jeunes Adultes de l'hôpital Paul Brousse
Autorité de contrôle prudentiel et de résolution (ACPR)	• **M. FLICHE Henri-Olivier**, Directeur de l'Innovation, des Données et des Risques technologiques
Autorité de régulation de la communication audiovisuelle et numérique (Arcom)	• **M. BONAN Jeremy**, Directeur adjoint des plateformes en ligne • **M. LOUTREL Benoît**, Membre du collège • **Mme PÉCAUT–RIVOLIER Laurence**, Membre du collège, Conseiller à la Cour de cassation
Banque de France	• **M. LASALLE Julien**, Adjoint au Directeur des études et de la surveillance des paiements
CAMELEON Association France	• **Mme CAMI Églantine**, Chargée de Sensibilisation et de Plaidoyer
Carnegie Endowment for International Peace	• **M. BATEMAN Jon**, Senior Fellow and Co Director Technology and International Affairs Program

CEA Saclay	• **M. GRINBAUM Alexei**, Directeur de recherche, Président du Comité opérationnel pilote d'éthique du numérique du CEA
Commission européenne	• **Mme GORNET Mélanie**, Technology Specialist in AI, CNECT.A.2
Commission nationale consultative des droits de l'homme (CNCDH)	• **M. CASTEX Lucien**, Membre • **M. DUMORTIER Thomas**, Conseiller juridique • **Mme ZOLYNSKI Célia**, Membre
Commission nationale de l'informatique et des libertés (CNIL)	• **Mme ELBAZ Jennifer**, Cheffe du service de sensibilisation du public
Comité consultatif national d'éthique du numérique (CCNEN)	• **M. CHATILA Raja**, Membre • **M. KIRCHNER Claude**, Président
Croix-Rouge française	• **Mme MARTINEZ Charlotte**, Responsable résilience numérique
Défenseur des droits	• **Mme BÉNICHOU Sarah**, Directrice de la direction Promotion de l'égalité et de l'accès aux droits
Délégation interministérielle à la lutte contre le racisme, l'antisémitisme et la haine anti-LGBT (DILCRAH)	• **Mme BENOUALID Shani**, Conseillère en charge du numérique et de la lutte contre la haine en ligne
École Polytechnique de Milan, département de *Design*	• **Mme BROADBENT Stefana**, Anthropologue des pratiques numériques
E-Enfance / 3018	• **Mme BÉCHU Véronique**, Directrice de l'Observatoire contre le harcèlement et les violences numériques faites aux mineurs e-Enfance 3018 • **M. COMBLEZ Samuel**, Psychologue de l'enfance et de l'adolescence, Directeur général adjoint de l'Association e-Enfance et Directeur du 3018
Everyone.AI / iRAISE Coalition	• **Dr. CERIOLI Mathilde**, Directrice Scientifique et cofondatrice • **Mme SERET Anne-Sophie**, Directrice Exécutive et cofondatrice
FairPatterns	• **Mme POTEL-SAVILLE Marie**, Co-fondatrice
FDJ UNITED	• **Mme KLEIN Sandrine**, Cheffe de département Data & IA Loterie • **M. PERROTIN Sylvain**, Chief Sustainability Officer

Fédération des Centres sociaux et socioculturels de France	• **Mme ROUSSET Mélanie**, Déléguée nationale • **Mme ZELLER Isabelle**, Vice-présidente
Fondation pour l'Enfance	• **Mme CABANETTES Justine**, Assistante plaidoyer • **Mme LEFRANC Angèle**, Responsable du plaidoyer • **Mme SICAMOIS Joëlle**, Directrice
Fordham University	• **M. SYLVAIN Olivier**, Professor of Law
Google	• **Mme BOITEUX Sarah**, Public Policy and government affairs manager • **M. ESPER Olivie**r, Public Policy, Senior Manager • **M. VERGNES Arnaud**, Public Policy, Senior Manager
Grand Orient de France	• **M. BERTINOTTI Pierre**, Grand maître • **M. SEDOGBO Célestin**, Conseiller de l'Ordre en charge du numérique et de l'impact sociétal
Haut Conseil à l'égalité entre les femmes et les hommes	• **Mme COUILLARD Bérangère**, Présidente
Human Technology Foundation	• **M. SALOBIR Éric**, Président
Internet Sans Crainte – TRALALERE	• **Mme DESAINT Axelle**, Directrice d'Internet Sans Crainte • **Mme ELALOUF-LEWINER Deborah**, Présidente de TRALALERE / Internet Sans Crainte
Institut Jean Nicod (ENS-PSL, EHESS, CNRS)	• **M. DARCY Grégoire**, Doctorant en Sciences Cognitives • **M. MERCIER Hugo,** Directeur de recherche
Institut national de recherche en sciences et technologies du numérique (Inria)	• **M. FERGUSON Yann**, Sociologue et directeur scientifique du LaborIA
Institut national de la santé et de la recherche médicale (Inserm)	• **Mme MELCHIOR Maria**, Directeur de Recherche (DR1)
La Quadrature du Net	• **M. LE QUERREC Bastien**, Juriste
Microsoft	• **Mme CRAIG Amanda**, General Manager, Responsible AI Public Policy • **Mme FRANÇOIS Elvire,** Senior Government Affairs Manager

Mistral AI	• **Mme HERBLIN STOOP Audrey**, Senior Vice President, Global Affairs
NewsGuard	• **Mme LABBÉ Chine**, Rédactrice en chef et Vice-présidente senior chargée des partenariats, Europe et Canada
Next	• **Mme SALIOU Mathilde**, Journaliste et autrice
Observatoire des Algorithmes Publics (ODAP)	• **Mme PENICAUD Soizic**, Co-fondatrice, chercheuse et consultante indépendante
OpenAI	• **M. SIGNOUX Martin**, AI Policy Lead, EMEA
Point de Contact	• **Mme MARISCAL-LOPEZ Alejandra**, Directrice générale
Projet Résilience Inspir	• **Mme DURAND-GIRARD Cécile**, Responsable relations institutionnelles • **M. GONTHARET David**, Président • **M. MARICHEZ Raphaël**, Membre fondateur et Trésorier • **M. SEZNEC Jean-Christophe**, Médecin psychiatre
Que Choisir Ensemble	• **M. MICHAELSEN Frithjof**, Chargé de mission relations institutionnelles Europe & secteur numérique
Sciences Po	• **M. BOULLIER Dominique**, Professeur de sociologie
Université Caen Normandie	• **Mme GALLARDO Anna**, Doctorante en droit du numérique
Université Columbia	• **Mme SCHIFFRIN Anya**, Ph.D, Senior lecturer at the school of international public affairs • **Mme WILLIAMS Kaylee**, Ph.D student
Université de Montréal	• **Mme REGIS Catherine**, Professeure de droit
Université Paris-Panthéon-Assas	• **M. EL HILALI Kamel**, Docteur en droit public

Université Paris 1 Panthéon-Sorbonne	• **M. ALIX Pascal**, Chercheur et avocat
Université d'Oslo	• **M. ESS Charles**, Philosopher and professor specializing in the ethics of AI, media, and digital technologies
Université de Toronto	• **M. PAPERNOT Nicolas**, Associate Professor
Wikimedia Foundation	• **M. GERLACH Jan**, Director of Public Policy • **Mme PUTZ Franziska**, Senior Movement Advocacy Manager • **Mme SCIUBBA CANIGLIA Costanza**, Knowledge Integrity Strategy Lead
Wikimédia France	• **Mme FRANÇOISE Camille**, Membre du conseil d'administration • **M. GERBET Rémy**, Directeur exécutif

Annexe 5 : Liste des contributions internationales sur les effets des usages de l'IA générative sur les mineurs

Organisme	Auditionnés et fonctions
ALAFIA LAB (Brésil)	• **M. CARREIRO Rodrigo**, Directeur et co-fondateur • **Mme ALMADA Maria Paula**, Co-directrice exécutive
ALANA (Brésil)	• **Mme CARVALHO SILVIA Letícia**, Consultante juridique internationale • **Mme CIFALI Ana Claudia**, Coordinatrice juridique • **Mme FERNANDES DE MENDONÇA Julia**, Avocate • **Mme GÓES DE MELLO Maria**, Coordinatrice des programmes • **Mme HENRIQUES Isabella**, Directrice générale • **M. NEJM Rodrigo,** Docteur spécialisé en éducation numérique • **Mme OTSUKA YAMAZOE Júlia**, Consultante juridique internationale
Balázs-Diák (Hongrie)	• **M. NAGY Róbert**, Maître de conférences, Chercheur en pédagogie numérique
Børns Vilkår (Danemark)	• **Mme VESTH RASMUSSEN Sørine**, Responsable senior des politiques numériques
CAMELEON Association (France)	• **Mme CAMI Églantine**, Chargée de sensibilisation et de plaidoyer • **Mme SINIC Anne**, Directrice générale adjointe
Center for Digital Ethics (Irlande)	• **Mme LENIHAN Niamh**, Fondatrice
Copenhagen Business School (Danemark)	• **M. FLYVERBOM Mikkel**, Professeur à la Copenhagen Business School, membre du Conseil danois d'éthique des données
Ctrl+alt+reclaim (France)	• Non précisé

DATA PRIVACY Brasil (Brésil)	• **M. BARCELLOS Victor**, Doctorant en communication et culture à l'Université fédérale de Rio de Janeiro • **M. BIONI Bruno**, Directeur fondateur • **M. GOMES MENDONÇA Eduardo**, Chercheur junior • **Mme NOVOA Natasha**, Spécialiste en sécurité numérique, gestion des données et gouvernance de l'Université pontificale catholique de Rio Grande do Sul (PUCRS) • **Mme RIELLI Mariana**, Co-directrice • **Mme RODRIGUES Carla**, Candidate à la maîtrise et au doctorat en communication et culture contemporaine de l'Université fédérale de Bahia (UFBA) • **M. ZANATTA Rafael**, Co-directrice
Dublin City University (Irlande)	• **M. CELESTE Edoardo**, Professeur associé, directeur adjoint du Dublin European Law Institute et membre du Law and tech research cluster • **Mme RIPPICH Karolin**, Chercheuse doctorante à Dublin City University Anti-Bullying Centre
E-Enfance / 3018 (France)	• **Mme ATLAN Justine**, Directrice générale • **Mme BECHU Véronique**, Directrice de l'Observatoire contre le harcèlement et les violences numériques faites aux mineurs e-Enfance 3018 • **M. COMBLEZ Samuel**, Psychologue de l'enfance et de l'adolescence, Directeur général adjoint de l'Association e-Enfance et Directeur du 3018
Everyone.AI / iRAISE Coalition (États-Unis)	• **Dr. CERIOLI Mathilde**, Directrice Scientifique et cofondatrice • **Mme SERET Anne-Sophie**, Directrice exécutive et cofondatrice
FARI – AI for the Common Good Institute (Irlande)	• **M. MÖRCH Carl**, Co-créateur
Fondation pour l'Enfance (France)	• **Mme LEFRANC Angèle**, Responsable du plaidoyer
Génération Numérique (France)	• **M. DI PALMA Cyril**, Délégué général
King's College London (Royaume-Uni)	• **M. SLOVAK Petr**, Professeur en interaction homme-machine
KORA (France)	• **Mme HERLIN Stéphie**, Co-fondatrice

Observatoire de l'IA à Rome Saint-Siège (Italie)	• **Séminaire du 29 avril 2026**, PFG7 – « IA et jeunesse : quels enjeux pour le développement et l'autonomie ? »
Pioneer Centre for AI (Danemark)	• **Mme LOUIZI GHONIEM Nermeen**, Responsable du programme technologique du groupe LEGO, membre du Pioneer Centre for AI, fondatrice de Hello Ada
Point de Contact (France)	• **Mme MARISCAL-LOPEZ Alejandra**, Directrice générale
Pontifical Catholic University of São Paulo (Brésil)	• **Mme FORTIM Ivelise**, Professeure et psychologue clinicienne
Stanford Accelerator for Learning (États-Unis)	• **Mme HAU Isabelle**, Directrice exécutive
Thorn (États-Unis)	• **Mme SLIFER Emily**, Directrice
Trinity College Dublin (Irlande)	• **Mme RAUH Maribeth**, Chercheuse • **Mme SHARYGINA Angelika**, Chercheuse doctorante
TMW Center for Early Learning + Public Health (États-Unis)	• **Mme SUSKIND Dana**, Docteure spécialisée en développement cérébral fondamental
Université catholique de Louvain (Belgique)	• **M. DEVILLE Yves**, Professeur
Université Charles (République tchèque)	• **Mme SLUSSAREFF Michaela**, Chercheuse et fondatrice de Slow Tech Institute
University of Eastern Finland (Finlande)	• **M. KAHILA Juho**, Chercheur postdoctoral • **M. TEDRE Matti**, Professeur • Mme VARTIAINEN Henriikka, Professeure
University of Oxford (Royaume-Uni)	• **M. PRZYBYLSKI Andrew**, Psychologue chercheur
Université de São Paulo (Brésil)	• **M. NABUCO DE ABREU Cristiano**, Psychologue clinicien
Université Sapienza di Roma (Italie)	• **M. MONTI Andrea**, Professeur en droit du numérique

VUB - Vrije Universiteit Brussel (Belgique)	• **M^me^ DUSHI Desara**, Chercheuse
WeProtect (Royaume-Uni)	• **M^me^ BURKE Katie**, Responsable des politiques publiques

Le Conseil de l'intelligence artificielle et du numérique (CIANum) est une instance indépendante placée auprès de la ministre déléguée chargée de l'Intelligence artificielle et du Numérique. Il a pour mission d'étudier toute question relative au développement du numérique et de l'intelligence artificielle ainsi que leur impact sur la société, l'économie et les territoires. Il est composé d'un collège interdisciplinaire de membres nommés par le Premier ministre et de parlementaires nommés par les présidents du Sénat et de l'Assemblée nationale. Il est institué par le décret du 8 décembre 2017 modifié par le décret du 4 septembre 2025.

Contact presse

presse@conseil-ia-numerique.fr

+33 6 24 33 29 08

Contact

info@conseil-ia-numerique.fr

Conseil de l'intelligence artificielle et du numérique

6 rue Louise Weiss,

75013 Paris

Impression : BPS – Pôle IRNF – Reprographie Bercy