MINISTÈRE
DE LA CULTURE

Liberté
Égalité
Fraternité

Rapport d'étape relatif à la mission sur les enjeux pour les secteurs culturels et créatifs des hypertrucages générés ou manipulés par l'intelligence artificielle

PRÉSENTÉ AU CONSEIL SUPÉRIEUR DE LA PROPRIÉTÉ LITTÉRAIRE ET ARTISTIQUE

Co-présidentes de la mission : Joëlle Farchy et Célia Zolynski

Co-rapporteurs : Bastien Blain et Alexandre Trémolière

CO-PRÉSIDENTES DE LA MISSION

Joëlle Farchy
Professeure à l'Université Paris 1 Panthéon-Sorbonne
Membre d'honneur du CSPLA

Célia Zolynski
Professeure à l'Université Paris 1 Panthéon-Sorbonne
Personnalité qualifiée au CSPLA

CO-RAPPORTEURS

Bastien Blain
Professeur junior à Paris 1 Panthéon - Sorbonne

Alexandre Trémolière
Administrateur de l'Etat

Rapport d'étape présenté à la réunion plénière du CSPLA du 9 juillet 2026.

Son contenu n'engage que ses auteurs.

TABLE DES MATIERES

La créativité,
c'est l'intelligence qui s'amuse !

Albert Einstein

INTRODUCTION

Le président du Conseil supérieur de la propriété littéraire et artistique a souhaité, dans un contexte où les conséquences de l'émergence et du développement de l'intelligence artificielle (IA), et plus encore de l'intelligence artificielle générative (IAG) demeurent, à plusieurs égards, incertaines, qu'une mission s'intéresse aux enjeux pour les secteurs culturels et créatifs des hypertrucages générés ou manipulés par l'intelligence artificielle. L'actualité ne cesse en effet de mettre en lumière les potentialités de ces technologies, qu'il s'agisse de l'apparition de Val Kilmer dans le récent film *As Deep as the Grave*, sorti un an après son décès, ou de la reconstitution sous la forme d'une vidéo du discours prononcé par le général de Gaulle à Londres le 18 juin 1940, alors qu'il n'en existe aucun enregistrement, sans oublier les nombreuses vidéos qui circulent sur les réseaux sociaux, devenant des outils de propagande dans de récents conflits.

L'intelligence artificielle générative permet en effet de générer ou manipuler des contenus qui présentent une ressemblance troublante avec la réalité. Ces contenus sont qualifiés d'hypertrucages. Le terme anglais, largement usité, « *deepfake* », né de la contraction de « *deep learning* » – apprentissage profond – et de « *fake* » – faux – apparaît en 2017 sur le forum Reddit, lorsqu'un utilisateur publie des vidéos pornographiques dans lesquelles les visages des actrices sont remplacés par ceux de célébrités hollywoodiennes.

La possibilité de tels montages n'est, en elle-même, pas nouvelle. Depuis Méliès, les effets spéciaux et tous les procédés de manipulation de l'image et du son ont toujours contribué à la « magie » du résultat et à repousser les limites de la créativité humaine. Dès les années 1930, Walter Benjamin, dans son classique *L'œuvre d'art à l'époque de sa reproductibilité technique*, soulignait d'ailleurs qu'au cinéma, en raison même du montage, la « *nature illusoire est une nature au second degré* ». L'auteur relevait en outre que, par sa reproductibilité, l'œuvre avait perdu son aura, sa qualité de l'*ici et maintenant*, de sorte que, comme l'écrivait Jean Baudrillard à sa suite dans les années 1970, « *la forme la plus avancée, la plus moderne de ce déroulement et que lui décrivait dans le cinéma, la photo, et les mass - média contemporains, est celle où l'original n'a plus même jamais lieu, puisque les choses sont d'emblée conçues en fonction de leur reproduction illimitée* »[1].

Toutefois, l'émergence de l'intelligence artificielle générative introduit une rupture ; ce qui change, c'est la rapidité et la simplicité avec lesquelles chacun peut désormais produire des illusions, grâce à des outils techniquement sophistiqués mais accessibles au « grand public », pour un coût limité, voire nul, sans qu'il soit toujours aisé de distinguer les contenus authentiques ou manipulés par IA. S'agissant par exemple de la photographie, il n'est plus possible d'affirmer de manière absolue que « l'*effet qu'elle [la photographie] produit sur moi n'est pas de restituer ce qui est aboli (par le temps, la distance), mais d'attester que cela que je vois, a bien été* »[2].

Combiné à la rapidité de diffusion des contenus en ligne, en particulier sur les réseaux sociaux, le risque est celui d'une perte irrémédiable de lien à la réalité, ainsi transformée : il n'est plus besoin de l'imagination pour que Don Quichotte voit « *ce qu'il ne peut voir et qui n'existe pas* » et Sancho n'est alors plus présent pour rappeler la réalité à son maître, qui, désormais sans contradiction, pourra croire encore plus fort « *qu'il en est exactement comme je le dis, et je me représente en mon imagination telle que je la désire* »[3]. Le lien au réel est requestionné donnant une nouvelle actualité aux inquiétudes de Jean Baudrillard : « *Cette course au réel et à l'hallucination réaliste est sans issue car, quand un objet est exactement semblable à un autre, il ne l'est pas exactement, il l'est un peu plus. Il n'y a jamais de similitude, pas plus qu'il n'y a d'exactitude. Ce qui est exact est déjà trop exact, seul est exact ce qui s'approche de la vérité sans y prétendre* »[4].

Les hypertrucages constituent désormais un défi aux multiples facettes compte tenu de la multiplicité des usages associés. Certains de ces usages font consensus tandis que d'autres portent atteinte aux intérêts de la société et des personnes visées : escroquerie, cyberharcèlement, désinformation, cybersécurité, diffamation à grande échelle, diffusion aisée de contenus à caractère raciste ou sexuel, nouvelles pratiques pédocriminelles… Le rapport Europol sur les défis posés aux forces de l'ordre par ces

[1] Jean Baudrillard, *Simulacres et simulation*, Gallimard, 2025 (éd. Originale 1981), p. 149.
[2] Roland Barthes, *La chambre claire. Note sur la photographie, Œuvres complètes*, t. 3, Paris, Seuil, 1993-1995, p. 1166.
[3] Cervantès, *Don Quichotte de la Manche*, I, 18 et I, 25.
[4] *Op. cit.*, p. 160-161.

outils d'IAG confirme par ailleurs que les hypertrucages saturent désormais les modes opératoires en matière de fraude documentaire, d'usurpation d'identité et de désinformation politique, justifiant un changement d'échelle dans les outils d'investigation[5].

Le droit français s'est saisi des enjeux du rapport au réel il y a plusieurs décennies : depuis la loi n° 70-643 du 17 juillet 1970 tendant à renforcer la garantie des droits individuels des citoyens, le droit français punit en effet celui qui a sciemment publié, par quelque voie que ce soit, le montage réalisé avec les paroles ou l'image d'une personne, sans le consentement de celle-ci, s'il n'apparait à l'évidence qu'il s'agit d'un montage ou s'il n'en est fait expressément mention.

Face aux problématiques plus récentes[6], l'adoption par l'Union européenne du règlement sur l'intelligence artificielle[7] (RIA) a constitué une étape importante à l'échelle mondiale[8]. Ce règlement s'intéresse essentiellement aux enjeux démocratiques et pour les droits fondamentaux liés au développement de l'intelligence artificielle, manifestant le caractère systémique pour nos sociétés des questions soulevées. Cette approche est déclinée à l'égard des hypertrucages, qui font l'objet d'une disposition spécifique. Le législateur européen a en effet constaté que : « l*a large disponibilité et les capacités croissantes de ces systèmes ont des conséquences importantes sur l'intégrité de l'écosystème informationnel et la confiance en celui-ci, ce qui pose de nouveaux risques de désinformation et de manipulation à grande échelle, de fraude, d'usurpation d'identité et de tromperie des consommateurs* »[9].

La réponse apportée repose sur la transparence, seule obligation substantielle que l'article 50 du RIA impose aux hypertrucages ; cette transparence est conçue comme un Sancho rappelant constamment à son maître d'interroger son rapport au réel puisque cette notion « *renvoie au fait que les systèmes d'IA sont développés et utilisés de manière à permettre une traçabilité et une explicabilité appropriées, faisant en sorte que les personnes réalisent qu'elles communiquent ou interagissent avec un système d'IA, que les déployeurs soient dûment informés des capacités et des limites de ce système d'IA et que les personnes concernées soient informées de leurs droits* »[10].

Dans le RIA, les enjeux, en ce qui concerne les secteurs de la culture et de la création, ne sont guère mentionnés. Or, dans ces secteurs, les hypertrucages ne se limitent pas à une technologie d'automatisation ou de fraude : ils interviennent au cœur même de la production de valeur, c'est-à-dire dans la création, l'interprétation, la mise en scène et la circulation des œuvres. Ils soulèvent donc des enjeux distincts de ceux observés dans la finance, la cybersécurité ou l'administration, où le problème principal est souvent l'usurpation ou la falsification d'identité.

La présente mission a précisément pour objet de mieux appréhender les enjeux liés aux hypertrucages pour ces secteurs en combinant une approche juridique et une approche économique. **Le rapport**, nourri des réponses adressées par les parties prenantes au questionnaire diffusé par la mission (annexes II et III) et des auditions menées au cours des derniers mois (annexe IV), **constitue un travail d'étape, destiné à constituer une base partagée sur laquelle des discussions pourront s'engager sur les réponses restant à apporter aux défis soulevés.**

Compte tenu de l'ampleur du sujet, de son caractère largement nouveau et évolutif, la mission a cherché, en premier lieu, à définir les contours précis de la notion (chapitre 1). Une nouvelle économie se met en place que nous présentons (chapitre 2) en réalisant une cartographie de la fabrique des hypertrucages (qui les produit, pourquoi et avec quelles conséquences pour les acteurs des filières

---

[5] Europol Innovation Lab, *Facing Reality? Law Enforcement and the Challenge of Deepfakes* (Europol, 2022), https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

[6] Voy. par ex. Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1 ou Robert Chesney, Danielle K. Citron, « Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security », *California Law Review*, 2019, vol. 107, p. 1753 – Adde, Rapport International sur la sécurité de l'IA, 2026, pt. 2.1.

[7] Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle.

[8] Voy. par ex. le colloque *Deepfakes: In Search of a Global Solution*, organisé le 24 octobre 2025 par l'Université de Columbia, l'Ecole Polytehnique, Sciences Po et l'Université Paris 1 Panthéon-Sorbonne : https://kernochan.law.columbia.edu/content/symposium-2025-deepfakes

[9] Considérant 133 du RIA.

[10] Considérant 27 du RIA.

culturelles). Le chapitre 3 présente l'encadrement juridique des hypertrucages dans le cadre du droit positif actuel et le chapitre 4 les conséquences pour le public de la perte de confiance associée à l'existence des hypertrucages. Les trois derniers chapitres examinent plus précisément le contenu juridique de l'obligation de transparence imposée par le RIA (chapitre 5), les modalités techniques d'identification et de détection des hypertrucages (chapitre 6) et enfin les responsabilités des différents acteurs de la chaîne de valeur (chapitre 7).

1. UN VASTE CHAMP PARTIELLEMENT PRIS EN COMPTE PAR LE RIA

La notion d'hypertrucage fait l'objet d'un traitement spécifique dans le RIA, permettant de définir la portée de l'obligation spécifique à des contenus, telle qu'elle figure à l'article 50, paragraphe 4. Cet article, en énonçant les obligations de transparence propres à différentes catégories de systèmes d'IA générative, au sein desquelles figurent les hypertrucages, contribue en outre, en creux, à compléter la définition donnée par l'article 3, point 60 :

> **« 60) «hypertrucage», une image ou un contenu audio ou vidéo généré ou manipulé par l'IA, présentant une ressemblance avec des personnes, des objets, des lieux, des entités ou événements existants et pouvant être perçu à tort par une personne comme authentiques ou véridiques ».**

Cette définition doit être replacée dans son contexte (1.1) avant d'en définir les contours précis, tels que la mission estime qu'ils doivent être retenus (1.2).

1.1. Une sous-catégorie des contenus générés ou manipulés par IA

L'analyse des différents termes de la définition des hypertrucages retenue par le RIA révèle les frontières qu'elle trace au cœur de l'ensemble des contenus pouvant être manipulés ou générés par IA. La lecture de l'article 50 du RIA montre en effet que l'hypertrucage est une catégorie de contenus générés par IA, les deux notions ne se recoupant pas : si l'alinéa 2 du paragraphe 4 porte spécifiquement sur les hypertrucages, le paragraphe vise les « *contenus de synthèse de type audio, image, vidéo ou texte* » générés par les systèmes d'IA (et exclusivement ceux-ci, qui sont par ailleurs soumis aux obligations du chapitre V lorsqu'ils constituent des modèles d'IA à usage général).

La confrontation de ces différents paragraphes et de la définition de l'article 3, point 60, fait apparaître d'emblée qu'**aux côtés des hypertrucages tels que le RIA les définit et les appréhende, deux autres catégories tangentent le champ de cette définition** et interrogent les mécanismes de régulation mis en place. **Il s'agit d'une part des textes et d'autre part des autres contenus générés par IA qui peuvent satisfaire une partie des critères de la définition, sans y répondre entièrement** : on conçoit ainsi qu'un personnage fictif entièrement généré par IA, sous forme simplifiée, ne constitue pas, de manière évidente, un hypertrucage lorsqu'il est nouveau et ne comporte aucune ressemblance avec un être humain ; on approche les limites de la définition dès lors que d'autres hypothèses sont étudiées, comme la création de personnages virtuels dont le réalisme rend difficile la distinction avec une personne existante ou un objet qui s'avèrerait être un pastiche d'un objet existant.

Les hypertrucages, au-delà de prouesses techniques, forment en effet un ensemble d'usages structurés par la relation à un existant ou par la finalité qui leur est assignée, plutôt qu'un bloc homogène. La figure 1 tente de fournir une typologie de ce foisonnement.

Figure 1 : Contenus synthétiques intégralement ou partiellement générés ou manipulés par IA.

Source : Butraud, Farchy, 2026

Certains contenus images, audios ou vidéos **(bulle n°1)** présentent des « ressemblances » évidentes selon les catégories proposées par le RIA : avec des personnes (leur apparence complète ou

simplement leur voix ou leur visage) ou avec des objets (archives...), des lieux (actuels, historiques, détruits ou inaccessibles), des entités (univers d'artistes, identité de marque, personnage fictif célèbre...) ou événements (interview, entretien, témoignage, discours historique...) existants. Ce sont donc les contenus qui satisfont à l'ensemble des critères de définition d'un hypertrucage.

Bien que n'étant pas considérés comme des hypertrucages par le RIA, de nombreux textes générés par IA présentent une forte ressemblance avec des éléments connus (bulle n° 2) : imitation du style littéraire donnant lieu à de faux inédits, de fausses traductions ou des variantes manipulées ; faux entretiens ou discours d'un auteur, préfaces ou manuscrits attribués à un écrivain… Ces textes générés par IA n'entrent pas dans le champ de la définition des hypertrucages, mais des seuls contenus générés par IA soumis à l'article 50, paragraphe 2, bien que l'article 50, paragraphe 4, en tienne compte d'une manière très limitée, se bornant à étendre l'obligation de marquage aux textes publiés dans le but d'informer le public sur des questions d'intérêt public. Cela ne couvre pas, sauf de manière marginale, les œuvres textuelles protégées alors que, de manière générale, des problématiques sont communes avec les contenus audio et vidéo, notamment s'agissant des risques de désinformation et de manipulation à grande échelle, de fraude, d'usurpation d'identité et de tromperie des consommateurs.

En outre, **le RIA soulève des interrogations quant au sort à accorder aux « créations fantômes » qui, sans avoir une ressemblance directe avec une personne, un lieu, une entité, un événement ou un objet existant *stricto sensu*, n'en présentent pas moins une illusion de réalité**. Pour les motifs mentionnés ci-après, la mission considère qu'il y a lieu de retenir une lecture étendue du terme *existant*. Pour autant, le sort à donner aux créations purement fantômes, tels que les influenceurs virtuels, n'est pas réglé par ce seul choix. En effet, certaines créations ne cherchent pas à s'appuyer sur une ressemblance avérée avec l'existant mais simplement à être perçues comme authentiques par le public. Des voix, visages, lieux, écrits, peuvent être créés en donnant une forte impression d'authenticité, sans pour autant ressembler à des éléments connus **(bulle n°3)**. Une scène imaginaire ou un avatar inventé de toutes pièces parce qu'il n'imite pas forcément le réel, bien qu'il ne relève pas des hypertrucages au sens du RIA, n'en est pas moins susceptible de soulever des questions. Ces créations « fantômes », si elles ne présentent pas de ressemblance avec des personnes ou éléments existants, répliquent en effet une apparence humaine (artistes, groupes musicaux, influenceurs, toutes « personnalités fantômes »…), des éléments d'apparence humaine (voix, visage…) ou des références à des créateurs ou auteurs pouvant être perçus comme des humains et tromper le public.

La définition de l'hypertrucage n'est, en tout état de cause, pas évidente car elle doit allier une réalité technique encore nouvelle et évolutive avec une catégorisation qui ne saurait être trop rigide afin de permettre la prise en compte d'évolutions futures. Sa définition a, au demeurant, été confortée au cours des travaux législatifs sur le RIA. Il s'avère à cet égard que, **si le législateur européen a fait le choix d'une définition qui s'avère relativement ouverte[11], elle revêt surtout un caractère finaliste**, lié à la définition du champ d'application de l'obligation d'étiquetage prévue à l'article 50, paragraphe 4. Il s'agit d'identifier un type de contenu synthétique, présentant des enjeux spécifiques pour la démocratie : le considérant 133 retient l'enjeu de « *l'intégrité de l'écosystème informationnel et la confiance en celui-ci* », la génération de grandes quantités de contenu de synthèse qu'il est de plus en plus difficile de distinguer du contenu authentique pour l'œil humain, créant de « *nouveaux risques de désinformation et de manipulation à grande échelle, de fraude, d'usurpation d'identité et de tromperie des consommateurs* ». **C'est la logique globale du règlement tenant à la minimisation des risques pour les droits fondamentaux qui s'y retrouve[12], sans que les enjeux des secteurs créatifs et culturels aient été directement pris en compte[13]**.

Les hypertrucages ont toutefois une particularité, qui justifie leur traitement spécifique dans le cadre du RIA : **c'est le risque de confusion avec le réel** ou, plus largement, selon le terme retenu, l'existant.

---

[11] Voy. l'intervention de Jennifer Rothman lors du colloque *Deepfakes: In Search of a Global Solution*, organisé le 24 octobre 2025 par le Kernochan Center (Columbia Université) et l'IRJS (Université Paris 1) : https://kernochan.law.columbia.edu/content/symposium-2025-deepfakes

[12] Arnaud Latil, « Le Règlement relatif à l'intelligence artificielle et l'approche par les risques : application d'une méthode législative structurante », *RTD eur.*, 2024, p. 563 ; Isabel Kusche, « Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk », *Journal of Risk Research*, 11 mai 2024.

[13] Andres Guadamuz, « The EU's Artificial Intelligence Act and copyright », *The Journal of World Intellectual Property*, 2025, p. 213.

Il en découle en effet une double interrogation qu'il faut garder à l'esprit lorsqu'on chercher à définir les contours de la notion :

- La première porte sur **le risque de tromperie** pour la personne confrontée à l'hypertrucage : la notion de manipulation revêt alors cumulativement sa double signification technique et péjorative. Cette tromperie s'entend au-delà de la notion usuellement retenue, notamment en droit de la consommation[14] ou, dans une perspective plus ouverte, dans certaines infractions du code pénal comme l'escroquerie ou l'abus de confiance ;
- La seconde interrogation concerne **le consentement de la personne dont les attributs sont manipulés par IA** (pour autant que le contenu hypertruqué concerne une personne) ou dont l'œuvre est utilisée par l'IA : on serait tenté de considérer que l'accord donné à cette utilisation correspond à une adhésion de la personne au contenu ainsi généré, mais cela repose, d'une part, sur une vision de la relation contractuelle fondant ce consentement qui, bien qu'idéale en ce qu'elle permet de ne consentir qu'à une utilisation déterminée, n'est pas toujours confirmée par la réalité des contrats et accords – mais il ne s'agit pas là d'une question spécifique aux hypertrucages. D'autre part, on ne peut exclure que la personne ait donné son consentement dans la perspective d'une simple manipulation.

La prise en compte de ces deux interrogations souligne qu'**il y a lieu pour appréhender les hypertrucages d'envisager ceux-ci sous leurs différentes formes et finalités : il peut y avoir des hypertrucages préjudiciables, mais d'autres qui ne le sont pas**[15]. Créer un personnage virtuel avec un univers destiné à créer une identité dans une logique artistique, voire commerciale, ne soulève pas de difficulté de principe (si l'ensemble des droits de propriété intellectuelle sont préservés) ; en faire usage pour manipuler les destinataires ou propager de fausses informations s'inscrit dans une autre perspective.

Or, le législateur européen a fait un choix structurant : définir l'hypertrucage par des éléments aussi objectifs que possible, sans prise en compte d'une dimension positive ou négative tirée de sa finalité, qui aurait eu pour effet d'en relativiser la portée. Il se focalise ainsi sur la technologie utilisée, dans la perspective de réduire les effets négatifs d'usages problématiques.

A cet égard, la mission souhaite **lever une ambiguïté terminologique**, que la consultation écrite a permis de faire ressortir. Le RIA traduit le terme « *deepfake* », qui figure dans la version anglaise (ainsi que dans plusieurs versions, comme l'allemande et l'italienne), par « hypertrucage » dans la version française : si on perçoit une nuance entre les deux termes, tenant à la connotation de la dénomination anglaise, qui contient le mot « *fake* » (qu'on retrouve par exemple dans le terme utilisé la version espagnole, « *ultrafalsificación* », ou celui de la portugaise, avec « *Falsificações profundas* »), **les deux termes doivent être considérés sur le plan juridique comme strictement équivalents**. La commission d'enrichissement de la langue française avait, en 2020, proposé d'utiliser les termes *vidéotox* ou *infox vidéo*[16], qui se sont avérés doublement restrictifs (par la référence à la seule vidéo et le présupposé d'une volonté d'infox, oubliant le caractère, entre autres, parodique qu'un tel contenu peut avoir) et donc insatisfaisants, ce qui a conduit à préférer le terme proposé en 2019 par l'Office québécois de la langue française[17]. Les principaux dictionnaires édités récemment en France (notamment le Robert et le Larousse) reprennent désormais ce terme. **La mission s'en tient à cette stricte équivalence, qui est conforme à la neutralité du terme retenu et qui, quel qu'il soit, ne préjuge pas de la licéité du contenu hypertruqué**.

1.2. L'hypertrucage au sens du RIA : une notion circonscrite et finaliste

La notion d'hypertrucage, telle que définie à l'article 3, point 60, suppose **la réunion de trois conditions, qui soulèvent des questions distinctes** : le type de contenu ; une ressemblance avec des personnes, objets, lieux, entités ou événements existants ; un risque de perception erronée. Ces questions sont déterminantes pour connaître la portée de l'obligation résultant de l'article 50, paragraphe 4, et sont apparues suffisamment ouvertes pour conduire certains auteurs à souligner la nécessité de clarifications par les autorités compétentes et les juridictions[18].

---

[14] Pratiques commerciales trompeuses définies par les articles L. 121-2 à L. 121-5 du code de la consommation

[15] Voy. à ce propos les exemples donnés par J. Rothman dans son intervention précitée.

[16] Ce qui figure toujours sur le site officiel : https://www.culture.fr/franceterme/terme/CULT753

[17] https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26552557/hypertrucage

[18] Thomas Gils, « A Detailed Analysis of Article 50 of the EU's Artificial Intelligence Act », *in* C. N. Pehlivan, N. Forgó and P. Valcke (eds.), *The EU Artificial Intelligence (AI) Act: A Commentary*, Kluwer Law International, 2025, p. 776-823.

Encadré – La notion de « digitial replicas » dans le cadre du droit américain

La mission ayant, compte tenu de l'importance de la littérature universitaire sur le sujet, étudié les différentes problématiques du présent rapport à la lumière comparatiste du droit américain – sans qu'il y soit fait nécessairement référence – il est intéressant de mettre en rapport la définition retenue par le RIA avec celle des « *digital replicas* » (réplique numérique) retenue par le rapport du *US Copyright Office*[19] :

« *Le rapport utilise le terme de « réplique numérique » pour se référer à une vidéo, une image ou un enregistrement audio qui a été créé ou manipulé numériquement en vue de représenter de manière réaliste mais fausse un individu.* »[20]

Les différences avec le RIA sont notables en ce qu'elle se limite aux personnes et se réfère au caractère réaliste mais faux de la représentation, et ne s'appuie pas sur l'appréciation d'un risque de perception erronée. Par ailleurs, la définition proposée par le *US Copyright Office* est technologiquement neutre puisqu'elle ne se réfère pas à l'utilisation d'un système d'IA générative.

La définition proposée dans le *No Fakes Act*, dans sa version présentée au Sénat le 4 septembre 2025[21], est plus détaillée (Section 2, a, 2) :

« *Le terme « réplique numérique »*

(A) désigne une représentation électronique nouvelle créée, générée par ordinateur et hautement réaliste qui est facilement identifiable comme la voix ou l'image d'une personne qui

(i) est intégrée à un enregistrement sonore, une image, une œuvre audiovisuelle, y compris une œuvre audiovisuelle ne comportant aucun son, ou une transmission

(I) dans laquelle la personne réelle n'a pas joué ou n'est pas apparue réellement ; ou

(II) qui est une version d'un enregistrement sonore, d'une image ou d'une œuvre audiovisuelle dans lequel la personne réelle a joué ou est apparue, et dont le caractère fondamental de la performance ou de l'apparence a été substantiellement manipulé ; et

(B) ne comprend pas la reproduction électronique, l'utilisation d'un échantillon d'un enregistrement sonore ou d'une œuvre audiovisuelle dans une autre œuvre, le remixage, la masterisation ou la remasterisation numérique d'un enregistrement sonore ou d'une œuvre audiovisuelle autorisée par le titulaire des droits d'auteur. »[22]

On trouve par ailleurs, dans le droit positif américain, une notion distincte, celle de contrefaçon numérique (« *digital forgery* »), qui figure par exemple dans le *Take It Down Act* du 19 mai 2025[23] :

« *Le terme « contrefaçon numérique » désigne toute représentation visuelle intime d'une personne identifiable créée au moyen d'un logiciel, d'une machine apprenante, d'une intelligence artificielle, ou tout autre moyen de génération par ordinateur ou technologique, y compris par adaptation, modification, manipulation, ou altération d'une représentation visuelle authentique, qui, lorsqu'elle est considérée dans son ensemble par une personne raisonnable, n'est pas distinguable d'une représentation visuelle authentique de la personne.* »[24]

1.2.1. Les contenus concernés : l'exclusion de l'essentiel des textes

Le législateur européen a fait le choix de limiter la notion d'hypertrucage aux images et contenus audio ou vidéo générés ou manipulé par l'IA. La portée de cette référence à la manipulation par une IA sera évoquée ultérieurement (5.1.2), mais il y a lieu ici de relever le caractère relativement large du champ qui correspond à l'ensemble des contenus visuels et auditifs générés par une IA, même si, contrairement à l'article 5 de la directive (UE) 2024/1385 du 14 mai 2024 sur la lutte contre la violence à l'égard des femmes et la violence domestique au matériels similaires, le législateur n'a pas étendu la référence aux « *matériels similaires* », ce qui ne semble pas devoir faire l'objet d'une interprétation *a contrario*.

---

[19] Sur les autres définitions débattues aux Etats-Unis : Jennifer Rothman, « Reframing Deepfakes », *Columbia Journal of Law & the Arts*, 2026, vol. 49, n° 4, p. 101.

[20] *Copyright and Artificial Intelligence*, Part 1: Digital Replicas, juillet 2024, p. 2.

[21] 119e Congrès, 1re session, S. 1367 *To protect intellectual property rights in the voice and visual likeness of individuals, and for other purposes*.

[22] Traduction des auteurs.

[23] 119e Congrès, *Public Law* 119-12 : *An Act To require covered platforms to remove nonconsensual intimate visual depictions, and for other purposes*.

[24] Traduction des auteurs.

La conséquence principale de cette délimitation du type de contenu est l'exclusion des textes du champ des hypertrucages. Il en résulte qu'**alors que l'obligation mise à la charge des fournisseurs de systèmes d'IA générative à caractère général par l'article 50, paragraphe 2, s'applique aussi à la génération de textes** (le RIA le prévoit expressément), **tel n'est pas le cas de l'obligation d'étiquetage imposée à tous les déployeurs par l'article 50, paragraphe 4**, alinéa 1er (le second alinéa ne couvrant en tout état de cause qu'une partie très limitée des textes, qui ne sont pas considérés comme des hypertrucages).

Ce choix, qui a été discuté au cours des travaux préparatoires[25] et contesté[26], peut s'expliquer à la fois par la difficulté d'identifier des textes hypertruqués (et les marquer)[27] ainsi que par un risque de tromperie potentiellement moindre compte tenu de la différence de média concerné. Il n'en constitue pas moins **un sujet d'inquiétude légitime**, que les réponses à la consultation écrite ont souligné à plusieurs reprises, en particulier auprès des acteurs de l'information. Il est d'ailleurs paradoxal que l'obligation de marquage de l'article 50, paragraphe 2, soit applicable, tandis que celle d'étiquetage du paragraphe 4 ne le soit que dans des cas très limités. On peut certes relever qu'outre les obligations pesant sur les éditeurs de presse au titre de la loi du 29 juillet 1881, l'article 1-1 de la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (LCEN) impose l'identification des éditeurs d'un service de communication au public en ligne, conformément à l'article 5 de la directive 2000/31/CE du Parlement européen et du Conseil du 8 juin 2000 relative à certains aspects juridiques des services de la société de l'information, et notamment du commerce électronique, dans le marché intérieur (directive sur le commerce électronique), tandis que le principe de neutralité ne dispense les fournisseurs de services intermédiaires de responsabilité à l'égard des contenus que dans les conditions prévues par le règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques (règlement sur les services numériques - DSA).

Néanmoins, alors qu'en l'état, on peut s'interroger sur le point de savoir si l'absence de respect des obligations de l'article 50 implique l'illicéité de l'hypertrucage (voy. *infra*, partie 7), **l'enjeu est d'abord celui de l'information du lecteur et d'éviter des manipulations**. Par cette exclusion, le RIA conduit à ce que ces enjeux soient exclusivement pris en compte dans une démarche répressive à l'égard des contenus textuels diffusés. Il est vrai que, de ce point de vue, **le droit français offre plusieurs possibilités, qui devraient permettre de couvrir la plupart des cas d'atteintes à des droits**. On retrouve d'ailleurs la même exclusion des textes en droit français s'agissant de l'infraction de l'article 226-8 du code pénal : celle-ci réprime le montage réalisé avec les paroles ou l'image d'une personne. Mais de nombreuses infractions sont invocables à l'encontre d'écrits litigieux, ainsi que l'a relevé le Conseil constitutionnel pour censurer la création d'un délit de diffusion en ligne d'un contenu qui soit porte atteinte à la dignité d'une personne ou présente à son égard un caractère injurieux, dégradant ou humiliant, soit crée à son encontre une situation intimidante, hostile ou offensante[28]. Le code pénal, la loi du 29 juillet 1881 ou le code de la consommation en sont les principaux supports. **Mais, face à de fausses informations, ce sont essentiellement des dispositions aux conditions d'invocation restreintes qui sont applicables** : le délit de diffusion de fausses nouvelles réprimé par l'article 27 de la loi de 1881 suppose un risque de trouble à la paix publique[29] et les outils destinés à lutter contre la diffusion d'allégations inexactes ou trompeuses de nature à altérer la sincérité d'un scrutin ne peuvent être utilisés que trois mois avant un scrutin[30].

Ces deux cas illustrent d'ailleurs l'hypothèse prévue par le RIA d'un **étiquetage « *des textes publiés dans le but d'informer le public de questions d'intérêt public* »**, dont les contours sont, pour le reste, incertains. S'agissant des textes intervenant en période électorale, le Conseil constitutionnel avait validé ces dispositions en relevant que « *les contenus d'information se rattachant à un débat d'intérêt général visés par*

---

[25] Voy. la proposition d'inclusion de la commission pour la culture et l'éducation du Parlement européen dans son avis du 16 juin 2022, doc. PE 719.637, amendements 18, 23 et 53.

[26] Voy. les principaux éléments de débat retracés par Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1.

[27] Jiameng Pu *et al.*, « Deepfake Text Detection: Limitations and Opportunities », IEEE Symposium on Security and Privacy (SP), 2023.

[28] Décision n° 2024-866 DC du 17 mai 2024.

[29] Voy. à ce sujet David Dechenaud, Philippe Conte, V° Presse et communication – Outrage envers les agents diplomatiques étrangers – Fausses nouvelles, *JurisClasseur Lois pénales spéciales*, fasc. 70, 2025.

[30] Art. L. 163-2 du code électoral, issu de la loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information.

les dispositions contestées sont ceux qui présentent un lien avec la campagne électorale »[31]. S'il ressort, par *a contrario*, de l'exception que comporte l'alinéa 2 du paragraphe 4 de l'article 50 du RIA que les articles de presse relèvent également des contenus d'information se rattachant à un débat d'intérêt général, **deux incertitudes** sont à relever :

- Si la référence au but informatif permet d'exclure les communications commerciales[32] et les textes à caractère artistique, on peut néanmoins s'interroger sur la portée réelle de ce but (par ex. s'agissant d'une tribune ou d'un autre texte exprimant une opinion, qui peut aussi prendre la forme d'un texte à caractère scientifique dont l'intérêt public peut être retenu). On pourrait être tenté de faire une analogie avec la notion de *nouvelle* figurant à l'article 27 de la loi du 29 juillet 1881, mais celle-ci paraît trop restrictive, alors qu'il est manifestement conforme à l'objectif poursuivi que des publications scientifiques générées par un système d'IA générative entrent dans le champ de cet alinéa ;
- Les « *débats d'intérêt public* » sont potentiellement très larges. Le considérant 96 énonce d'ailleurs, de manière non exhaustive, parmi les « *missions d'intérêt public* » l'éducation, les soins de santé, les services sociaux, le logement et l'administration de la justice. Ce considérant apparaît comme un élément de référence essentiel pour déterminer le champ de cette notion et la Commission, dans ses lignes directrices sur l'article 50 du RIA en reprend largement les éléments énoncés, tout en relevant que la liste n'est pas exhaustive, ni figée (paragr. 131).

Face à ces incertitudes, le cas des informations à caractère scientifique, et notamment en matière de santé publique, paraît particulièrement révélateur : dans un contexte où la distinction entre information et opinion devient floue[33], la publication d'un article présentant des informations prétendument scientifiques doit pouvoir être appréhendée comme un hypertrucage dès lors qu'il entend revêtir un caractère informatif – quand bien même il conserve la forme d'un travail de recherche, dont on pourrait estimer qu'il a le caractère d'une opinion. A cet égard, **on ne peut méconnaître les exigences liées à la liberté d'expression** : celle-ci implique que ne puissent être qualifiées d'allégations inexactes ou trompeuses des opinions, des parodies, des inexactitudes partielles ou de simples exagérations[34]. Pour autant, **la qualification d'hypertrucage n'implique pas celle de fausseté ; il s'agit seulement de faire connaître, par l'obligation de transparence, les conditions de génération d'un texte et, par suite, inciter au regard critique**. S'il ne peut ainsi être donné une portée trop large à l'obligation de transparence applicable aux textes, alors même qu'elle est en grande partie vidée de sa substance par la légitime réserve dont bénéficient les services soumis à un contrôle éditorial, l'obligation en cause ne paraît pas, par principe, contraindre la liberté d'expression, d'autant qu'elle s'applique au déployeur qui, par définition, est conscient des conditions de génération du texte concerné. L'exemple récent de la « *bixonimanie* »[35] souligne l'importance qui s'y attache en outre dans un contexte où les systèmes d'IA générative se nourrissent des informations disponibles en ligne.

S'agissant plus directement du champ de la présente mission, à ces incertitudes s'ajoute le fait qu'**en excluant la plupart des textes répondant, pour le reste, à la définition de l'hypertrucage, le législateur européen restreint la possibilité de protéger les droits de propriété intellectuelle** : l'utilisation d'un système d'IA générative pour prêter à une personne des écrits qu'elle n'a pas produits peut lui porter préjudice, sans qu'il s'agisse nécessairement d'une diffamation ou d'un outrage. Il y aurait certes une pratique commerciale trompeuse si était mis en vente un ouvrage attribué à un auteur qui ne l'a pas écrit, mais le préjudice résultant de cette infraction serait invocable par les lecteurs potentiels et non par l'auteur, qui ne peut invoquer qu'un préjudice réputationnel dans les conditions de droit commun. L'enjeu est potentiellement massif comme le souligne le cas de Grammarly et sa fonction « *Expert Review* », qui offre des avis inspirés du style d'auteurs célèbres[36] Quant au droit de l'auteur sur les écrits, alors qu'on peut raisonnablement considérer que le système d'IAG a pu bénéficier de données issues d'œuvres protégées au sens de l'article L. 112-2 du code de la propriété intellectuelle, il se heurtera aux difficultés que la création

---

[31] Décision n° 2018-773 DC du 20 décembre 2018, paragr. 8.

[32] Celles-ci relèveront le cas échant des pratiques commerciales déloyales sanctionnées par la directive 2005/29/CE du 11 mai 2005 et les articles L. 121-1 et suivants du code de la consommation.

[33] Vincent de Coorebyter, « L'internet : démocratie ou démagogie », *in* Yves Poullet (éd.), *Vie privée, liberté d'expression et démocratie dans la société numérique, Cahier du CRIDS*, n° 47, Bruxelles, Larcier, 2020, p. 252.

[34] Cons. Const., déc. n° 2018-773 DC du 20 décembre 2018, paragr. 21.

[35] Chris Stokel-Walker, « Scientists invented a fake disease. AI told people it was real », *Nature*, 7 avril 2026.

[36] Julia Angwin, « Why I'm Suing Grammarly », *the New York Times*, 13 mars 2026 ; Jennifer Rothman, « Grammarly Lawsuit Shows Existing Laws Can Combat Deepfakes », *Lawfare*, 9 avril 2026 (en ligne).

d'une présomption d'utilisation de telles œuvres entend lever[37]. Dans le cas de génération de propos, l'auteur ne pourra faire valoir que l'effet réputationnel, alors que la mention de l'usage de l'IA aurait pu éviter de porter atteinte à sa réputation. Dans tous les cas, une clarté de l'information en amont serait de nature à assurer une plus grande transparence dans un contexte où la distinction entre le vrai et le faux devient toujours plus complexe à percevoir.

Au regard de ces incertitudes, qui découlent d'une prise en compte des hypertrucages par le seul biais de la lutte contre la désinformation, **la mission ne peut que regretter que l'obligation de transparence imposée sur les textes manipulés ou générés par IA soit restreinte à la seule hypothèse des textes d'intérêt général**, dont une grande partie de l'effet est restreinte par l'exclusion – légitime – de ce champ des textes faisant l'objet d'un contrôle éditorial. **Elle estime qu'à défaut de modification à une échéance raisonnable du RIA sur ce point, il y a lieu de retenir une interprétation la plus large possible du champ d'application de l'article 50, paragraphe 4, sur les textes à caractère informatif**, sans que cela ne préjudicie à la liberté d'expression dès lors que le seul effet est l'obligation d'étiquetage.

1.2.2. Une exigence de ressemblance avec l'existant

Ce critère est probablement celui qui a donné lieu au plus grand nombre de commentaires, tant il revêt une dimension centrale dans la qualification d'un contenu synthétique comme hypertrucage. Il repose lui-même sur trois éléments, qui appellent une série d'observations.

a) Le point de référence est défini largement par le législateur européen quant à son objet : il s'agit non seulement des personnes, hypothèse qui vient immédiatement à l'esprit lorsqu'on évoque les hypertrucages, notamment du fait des atteintes aux droits de la personnalité, mais également les objets, lieux, entités ou événements. Dans la perspective poursuivie par le législateur européen de lutte contre la manipulation de l'information, ce champ très large apporte une garantie importante : **il s'agit, *in fine*, de viser tous les contenus qui ont une ressemblance avec la réalité**[38]. Si des risques d'une interprétation trop large ont été soulignés[39], ils semblent pouvoir être écartés du fait des autres conditions posées à la caractérisation de la notion – étant au demeurant rappelé que la qualification d'hypertrucage n'induit aucune illicéité.

Ce champ large peut également être considéré comme offrant une garantie aux secteurs culturels et créatifs, dans la mesure où, au-delà de l'énumération, sont couverts tous les types d'éléments. A cet égard, le terme « objet », qui correspond à tout ce qui s'offre à la perception[40] et non aux seuls biens matériels, paraît pouvoir faire l'objet d'une lecture extensive par son caractère relativement neutre, ce qui permettrait de couvrir la ressemblance avec les œuvres de l'esprit. Les lignes directrices de la Commission européenne sur l'article 50 du RIA sont d'ailleurs en ce sens (paragr. 113).

En revanche, le législateur européen a posé une restriction en précisant que ces personnes, objets, lieux, entités et événements de référence pour l'appréciation de la ressemblance doivent être « *existants* ». Il a été souligné que **ce terme peut en réalité prêter à interprétation et modifier profondément la portée de la définition de l'hypertrucage**[41], entre une interprétation littérale stricte, impliquant que l'élément préexistant doit être établi, une interprétation littérale plus ouverte, visant un rattachement à ce qui est susceptible d'être réel, et une interprétation téléologique qui serait conforme à l'objectif poursuivi par le RIA, reposant sur la prise en compte de la possibilité d'une perception interférant avec le réel.

Cette question est essentielle et se pose principalement pour les contenus générés par IA. En effet, s'agissant des contenus qui sont seulement manipulés par un système d'IA générative, le contenu initial est, par définition, existant. S'agissant des contenus générés, ce sont surtout deux lectures qui se présentent : soit un contenu est un hypertrucage s'il ressemble à l'existant, défini comme ce qui existe comme une réalité

---

[37] Proposition de loi relative à l'instauration d'une présomption d'exploitation des contenus culturels par les fournisseurs d'intelligence artificielle, présentée par Mmes Laure Darcos, Agnès Evren, MM. Pierre Ouzoulias, Laurent Lafon, Mmes Catherine Morin-Desailly et Karine Daniel, sénatrices et sénateurs, enregistrée le 12 décembre 2025.
[38] Mateusz Łabuz, « A Teleological Interpretation of the Definition of DeepFakes in the EU Artificial Intelligence Act—A Purpose-Based Approach to Potential Problems With the Word "Existing" », *Policy & Internet*, 2025, vol. 17, e435.
[39] Michael Veale, Frederik Zuiderveen Borgesius, « Demystifying the Draft EU Artificial Intelligence Act », *Computer Law Review International*, 2021, vol. 22, n° 4, p. 97-112.
[40] Définition du dictionnaire de l'Académie française, 9e éd.
[41] Voy. l'étude précise de Mateusz Łabuz, préc.

vivante ou vécue[42], soit cette qualification est retenue lorsque le contenu a un degré de ressemblance à la réalité tel qu'il est susceptible de faire penser qu'il représente un existant (illusion de réalité). Entre ces deux lectures, **la mission considère que la seconde est la plus à même de conférer son effet utile au règlement**. Elle observe d'ailleurs que c'est la position retenue par la Commission européenne dans ses lignes directrices sur l'article 50 du RIA (paragr. 113) : « *il est suffisant que les personnes, objets, places, entités ou événements simulés ressemblent à quelqu'un ou quelque chose qui peut exister ou pourrait avoir existé dans la réalité pour être considéré comme un hypertrucage* ».

Il est vrai que d'autres versions linguistiques peuvent conduire à retenir la première interprétation (« *wirklich* » en allemand ; « *reales* » en espagnol[43]). En outre, si la notion de ressemblance, ainsi qu'il sera évoqué ci-après, peut utilement s'inspirer des concepts du droit de la propriété intellectuelle, ce qui conduirait ici à retenir aussi la notion d'existant telle qu'elle s'applique en ce domaine, le parallèle paraît limité par l'objet même des législations en cause, qui est substantiellement différent, l'une étant destinée à protéger ce qui préexiste, l'autre à prévenir un risque de confusion, voire de manipulation et de tromperie.

En outre, **les autres critères de la définition de l'hypertrucage seraient largement dépourvus de portée si une interprétation restrictive de l'existant était retenue** : la question de la ressemblance « *sensible* » ne se poserait pas s'il s'agissait d'une stricte identité à une personne ou un objet existant ; la perception d'un contenu comme authentique ou véridique serait nécessairement impliquée par la mise en scène d'un contenu existant. En outre, **l'interprétation restrictive rendrait complexe la capacité à agir face à un hypertrucage non conforme à l'obligation de transparence** : toute action devrait assurer la démonstration de l'existence de la personne, de l'objet, de l'événement, ce que seule la victime directe serait raisonnablement en mesure de faire, alors que le préjudice subi peut dépasser cette seule victime. En outre, cette victime dispose, si elle est visée, d'outils juridiques pour agir, notamment à travers la garantie des droits de la personnalité ou les droits de propriété littéraire et artistique, de sorte que, par exemple, les hypertrucages qui donneraient vie à une personne purement imaginaire mais d'une parfaite ressemblance avec une personne susceptible d'être réelle, voire inspirée de personnes réelles mais mêlées – à l'image de Tilly Norwood[44] ou Lolita Cercel[45] – ne seraient guère atteignables, en dehors des procédures de sanction de l'article 99 du RIA (qui, comme on le verra dans la partie 7, ne visent que les déployeurs et pas les contenus, contre lesquels les possibilités d'action sont limitées).

Une telle approche étendue de la notion d'existant, incluant ce qui a pu être qualifié d' « hypertrucage fictionnel »[46], présente en outre l'avantage d'éviter les doutes que peut susciter l'hypothèse d'une ressemblance fortuite, dont les réponses au questionnaire de la mission ont souligné qu'elle soulevait des interrogations, et donc de l'incertitude.

b) Le contenu concerné doit en outre présenter une ressemblance avec ces éléments. Avant d'évoquer la notion même de ressemblance, il faut relever que la définition donnée par l'article 3, point 60, du RIA ne coïncide pas entièrement, sur ce point, avec le considérant 134 : ce dernier définit en effet l'hypertrucage comme présentant « *une ressemblance sensible* » (« *appreciably resembles* » en anglais) avec la réalité, cet adjectif « *sensible* » ne figurant pas dans les dispositions du règlement. Si ces dernières sont seules susceptibles d'avoir un effet juridique, cette discordance correspond également à une différence entre la définition de l'hypertrucage figurant dans le RIA et celle retenue, même si le terme d'hypertrucage n'est pas utilisé, par le DSA afin de définir les obligations des très grandes plateformes en ligne et des très grands moteurs de recherche en ligne à l'égard des risques systémiques liés aux contenus qu'ils hébergent ou auxquels ils facilitent l'accès : il est alors question d'un contenu qui « *ressemble nettement* » à la réalité[47]

---

[42] Définition du mot « *existant* » dans le dictionnaire de l'Académie française, 9e éd.

[43] En revanche, la version anglaise retient « *existing* ».

[44] Créatrice et actrice entièrement générée par IA, dont il est assumé qu'elle est inspirée de Scarlett Johansson et Natalie Portman.

[45] « Lolita Cercel, la chanteuse créée par IA en Roumanie, fascine et inquiète », *Le Monde Le mag*, 4 février 2026

[46] « *Fictional Deepfakes* », qui constitue une des quatre catégories d'hypertrucages identifiés par Jennifer Rothman, « Reframing Deepfakes », *Columbia Journal of Law & the Arts*, 2026, vol. 49, n° 4, p. 101.

[47] Art. 35 : « *1. Les fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne mettent en place des mesures d'atténuation raisonnables, proportionnées et efficaces, adaptées aux risques systémiques spécifiques recensés conformément à l'article 34, en tenant compte en particulier de l'incidence de ces mesures sur les droits fondamentaux. Ces mesures peuvent inclure, le cas échéant: (…) k) le recours à un marquage bien visible pour garantir qu'un élément d'information, qu'il s'agisse d'une image, d'un contenu audio ou vidéo généré ou manipulé, qui ressemble nettement à des personnes, à des objets, à des lieux ou à d'autres entités ou événements réels, et apparaît à tort aux yeux d'une personne comme authentique ou digne de foi, est reconnaissable lorsqu'il est présenté sur leurs interfaces en ligne,*

(« *appreciably resembles* » en anglais, ce qui confirme l'identité d'expressions). Si cet écart a pu être considéré comme étant le fruit d'une inattention ou d'une négligence[48], d'autant plus que le législateur européen s'est efforcé d'éviter l'écart de définition qui figurait initialement dans la proposition de RIA, il faut relever que toutes les versions linguistiques présentent ce même écart, de sorte qu'il peut être considéré comme un guide d'interprétation – et les lignes directrices de la Commission sur l'article 50 relèvent qu'il s'agit d'un « *ajout* » du considérant 134 (paragr. 113) et l'incluent dans la définition. Celui-ci, comme cela a été indiqué précédemment, pousse à faire une lecture de la définition de l'hypertrucage en vue de donner un effet utile au règlement : **la ressemblance doit être suffisamment marquée pour donner l'illusion de la réalité**.

La notion de ressemblance est centrale dans la définition de l'hypertrucage, d'autant plus que le degré de ressemblance est déterminant dans l'appréciation du risque tenant à la perception du contenu synthétique. Elle fait évidemment écho au terme qui est connu en droit de la propriété intellectuelle, plus précisément pour la caractérisation du délit de contrefaçon. Les cadre et objet sont différents, puisqu'il s'agit d'une part de caractériser l'atteinte à un droit et une infraction et, d'autre part, d'identifier un risque qui, en lui-même, ne caractérise pas une illicéité mais sert à déterminer la nécessité d'informer le destinataire quant au caractère artificiel du contenu généré. Elle permet toutefois d'apporter des éléments dans la compréhension de la notion, puisque c'est l'appréciation des ressemblances qui permet la qualification de la contrefaçon, et non les différences en tant que telles[49] : cette démarche – bien qu'elle soit parfois contestée[50] – paraît en elle-même adaptée à la qualification de l'hypertrucage. Si, en droit des marques, la contrefaçon est retenue lorsque les ressemblances sont telles qu'il en résulte un risque de confusion[51], ce dernier constitue ici un critère distinct. Pour autant, **la logique du droit d'auteur s'avère tout à fait pertinente, par la recherche des ressemblances, puis des différences**[52] (avant d'en venir au dernier critère du risque de confusion), les premières revêtant une importance particulière dès lors que c'est une ressemblance assez évidente qui est appréciée.

La notion de ressemblance en droit d'auteur est également intéressante en ce qu'elle permet de tenir compte des œuvres dérivées ou de celles qui empruntent à une œuvre protégée, même si la différence d'objectif induit un degré d'appréciation différent : la seule reprise d'éléments ne suffit pas à caractériser l'hypertrucage. Le cadre de prise en compte des idées dans le droit de la propriété littéraire et artistique constitue également un point de référence utile en ce qu'il permet de ne pas protéger les idées (hypothèse qui suppose que l'idée émerge non du prompt mais du contenu généré), tout en protégeant certaines mises en œuvre originales[53]. En revanche, au regard des considérations qui précèdent sur la notion d'*existant*, la jurisprudence relative aux ressemblances fortuites[54] n'aurait pas vocation à être transposée, d'autant que l'opacité du fonctionnement des systèmes d'IA générative (parfois qualifiés de « *boîte noire* »[55]) est susceptible de conduire à une invocation constante de cette fortuité, notamment par des déployeurs ne maîtrisant eux-mêmes pas les données d'apprentissage du système.

A la lumière de l'ensemble de ces éléments et au bénéfice des interprétations qui semblent pertinentes à la mission, celle-ci en déduit en outre que **la qualification d'hypertrucage devrait pourvoir**

---

et, en complément, la mise à disposition d'une fonctionnalité facile d'utilisation permettant aux destinataires du service de signaler ce type d'information. »

[48] Mateusz Łabuz, préc.

[49] Cass. Crim., 16 janvier 1957, Bull. crim. n° 25 ; Cass. 1re Civ., 4 février 1992, pourvoi n° 90-21.630, Bull. 1992, I, n° 42 ; Cass. 1re Civ., 15 juin 1994, pourvoi n° 92-19.824, Bull. 1994, I, n° 216.

[50] Voy. le débat doctrinal évoqué par Carine Bernault, Audrey Lebois, V° « Droit des auteurs – Contrefaçon et étendue du droit d'auteur (CPI, art. L. 122-4) », *JurisClasseur Propriété littéraire et artistique*, fasc. 267, 2024.

[51] Cass. Com., 4 octobre 1971, pourvoi n° 70-11.559, Bull. civ. IV, n° 225 ; Cass. Com., 7 octobre 1980, pourvoi n° 79-10.093, Bull. civ. IV, n° 327

[52] Sur le droit d'auteur : Pierre-Yves Gautier, Nathalie Blanc, *Droit de la propriété littéraire et artistique*, Paris, LGDJ-Lextenso, 2023, 2e éd., p. 644-645.

[53] Cass. 1re Civ., 17 juin 2003, pourvoi n° 01-17.650, Bulletin civil 2003, I, n° 148.

[54] Cass. 1re Civ., 16 mai 2006, pourvoi n° 05-11.780, Bull. 2006, I, n° 246 ; Cass. 1re Civ., 2 octobre 2013, pourvoi n° 12-25.941, Bull. 2013, I, n° 197

[55] Yavar Bathae, « The Artificial Intelligence Black Box and the Failure of Intent and Causation », *Harvard Journal of Law and Technology*, vol. 31, 2018, p. 889 ; Scott Lu, « Deepfakes Under Copyright Law – A Necessary Legal Innovation », *Southern Illinois University Law Journal*, 2024, vol. 48, p. 517. Les considérations parfois évoquées pour limiter la portée de ce phénomène de boîte noire (Kenneth S. Abraham, Catherine M. Sharkey, « Untangling AI Liability », *Virginia Public Law and Legal Theory Research Paper No. 2026-19*, 23 février 2026, à paraître à la *California Law Review*) ne remettent pas en cause l'analyse exposée ici.

inclure des contenus générés par IA « dans le style » d'un auteur, dès lors que la ressemblance est nette avec un contenu existant, sans qu'il y ait besoin d'avoir repris un élément particulier de cet auteur.

1.2.3. Le risque d'une perception erronée

Si la ressemblance est un élément déterminant de la qualification d'hypertrucage, l'évaluation de la possibilité d'une perception erronée de ce contenu comme authentique ou véridique constitue un élément moins aisé à apprécier et qui, *in fine*, **interroge sur la spécificité des hypertrucages parmi l'ensemble des contenus synthétiques générés par IA**.

Il faut d'abord relever que, conformément à la démarche du RIA, il s'agit d'un *risque*, et non de la qualification d'une intentionnalité : **c'est la potentialité du résultat qui importe**. Néanmoins, l'évaluation de ce risque est délicate, car, si on peut aisément comprendre que la perception « *par une personne* » **renvoie à une sorte de perception moyenne**, il n'en reste pas moins que les hypertrucages sont, en raison même de ce risque, de nature à constituer des pratiques d'IA interdites par l'article 5 du règlement ; ils peuvent en effet, y compris involontairement, être perçus par des personnes vulnérables d'une manière différente. Cette problématique fait directement écho à celle de l'appréciation du consommateur moyen en droit de la consommation[56], dont la variabilité n'a pas manqué d'interroger[57] et qui, si elle peut soulever des difficultés de lisibilité – et donc d'applicabilité – de la norme, permet de **procéder à une appréciation au regard du public visé**[58].

L'articulation entre les pratiques interdites et le champ d'application de l'obligation de transparence de l'article 50 du RIA sera évoquée ultérieurement, mais on peut d'ores et déjà souligner que l'existence même de ce risque, en particulier à l'égard des mineurs, implique, compte tenu de l'objectif de protection des droits fondamentaux affiché par le législateur européen[59], d'**écarter toute interprétation qui tendrait à réduire le champ des obligations nouvelles**. A cet égard, les enjeux liés aux secteurs culturels et créatifs revêtent une dimension systémique pour appréhender les hypertrucages, car l'impact que peut avoir l'emprunt d'éléments de ces secteurs, lorsqu'ils sont connus du grand public, en fait un moyen certain de manipulation, justifiant *a fortiori* une attention au respect des droits de ces personnes[60].

Sur le fond, ce critère s'intéresse à **la perception par le public, selon une approche objective**. Alors que les pratiques interdites visent le fonctionnement même du système d'IA, la notion d'hypertrucage recouvre les usages susceptibles d'être problématiques[61]. A ce titre, dans la mesure où cette qualification a pour seul effet d'imposer une obligation de transparence et ne porte pas, en elle-même, une appréciation sur l'illicéité de l'usage, elle est en principe exclusive d'une appréciation de l'intention du déployeur (ou de l'utilisateur à titre personnel, même si les obligations de l'article 50, paragraphe 4, ne s'appliquent pas dans ce cas). Il en résulte deux points d'attention.

D'une part, ce critère fait écho à d'autres éléments de qualification de pratiques interdites en droit de l'Union européenne ou en droit national. Celles-ci comportent alors un aspect d'intentionnalité et l'hypertrucage qui répondrait à l'ensemble de ces critères serait illicite. Il s'agit d'abord des diverses pratiques trompeuses et manipulatrices (entre autres, fausses informations en matière financière[62] ou

---

[56] Art. 5 de la directive 2005/29/CE sur les pratiques commerciales déloyales.

[57] Luis Gonzalez-Vaque, « La mise en œuvre de la directive 2005/29/CE sur les pratiques commerciales déloyales. Les consommateurs vulnérables sont-ils suffisamment protégés ? », *RDUE* 2013, n° 3, p. 471. Voy. plus généralement Guy Raymond, V° « Pratiques commerciales déloyales et agressives », *JurisClasseur Concurrence – Consommation*, fasc. 900, 2025.

[58] C'est la démarche prévue par la transposition de cette notion de consommateur moyen à l'article L. 121-1 du code de la consommation.

[59] Marion Ho-Dac, « La protection des droits fondamentaux dans l'AI Act », *RTD eur.*, 2024, p. 615 ; Julie Charpenet, « Les droits fondamentaux à l'épreuve de la normalisation de l'intelligence artificielle », *Dalloz IP/IT*, 2025, p. 591 ; Sarah Tabani, « Le règlement européen sur l'intelligence artificielle : problématiques choisies à l'issue des six premiers mois de l'entrée en application des premières obligations », *Rev. de l'UE*, 2025, p. 626.

[60] Ce dont la réglementation française tire d'ores et déjà certaines conséquences, par exemple à travers l'article D. 312-10 du code de la sécurité intérieure.

[61] Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1.

[62] Art. L. 465-3-2 et L. 465-3-3 du code monétaire et financier.

électorale[63], pratiques commerciales déloyales[64], faux, escroquerie[65]). Un rapprochement doit ensuite être fait avec les qualifications qui, pour apprécier le caractère illicite de la ressemblance (et donc le risque de confusion), tiennent compte du consentement à cette ressemblance : ce consentement induit, dans un modèle idéal-typique[66], que la confusion n'est qu'apparente puisque celui qui a consenti à l'utilisation de ses caractères a accepté celle-ci et approuve, par suite, le contenu. Tel est le cas des délits des articles 226-8 et 226-8-1 du code pénal ou du délit de contrefaçon. Pour autant, **pour la qualification de l'hypertrucage, la circonstance qu'il y ait eu un consentement à la manipulation des éléments d'identification par l'IA est sans incidence**. Cela est en cohérence avec la construction du RIA, qui se borne à imposer une obligation de transparence aux hypertrucages destinés à la bonne information du public : peu importe que David Beckham ait expressément consenti à ce qu'une IA soit utilisée pour lui faire tenir, dans le cadre de la diffusion de la campagne *Malaria No More*, des propos qu'il a validés mais dans des langues qu'il ne maîtrise pas et dont il n'a donc pas prononcé les mots, les personnes qui visionnent cette vidéo sont en droit de savoir – précisément – qu'une IA est venue se superposer pour prononcer le texte dans la langue du destinataire. Aussi, bien que certains répondants au questionnaire adressé aux parties prenantes par la mission aient pu s'interroger sur la prise en compte du consentement des personnes concernées, **la mission estime qu'il faut s'en tenir à une approche conforme à la logique générale du RIA, qui objectivise la qualification d'hypertrucage et ne préjuge aucunement de sa licéité**. Ce critère doit être apprécié plus vraisemblablement à la lumière du risque de confusion, tel qu'il a été développé également en droit de la consommation[67].

D'autre part, **ce critère de définition ouvre un risque de circularité qui pourrait aboutir à cette confusion avec la licéité et qui doit être évitée autant que cela est possible**. En effet, on pourrait estimer que le risque de perception erronée sur le caractère authentique ou véridique n'existe que pour autant que l'hypertrucage n'est pas clairement identifié. C'est d'ailleurs ce que prévoit l'article 226-8 du code pénal[68] et c'est en partie ce qu'avance la Commission européenne dans ses lignes directrices sur les pratiques d'IA interdites[69] (bien qu'elle se situe sur le seul terrain de la tromperie, alors que la tromperie peut persister malgré le marquage et l'étiquetage compte tenu d'un ensemble de pratiques manipulatrices). Il y a donc lieu de considérer que la qualification d'un contenu généré par IA comme étant hypertruqué doit être opérée sans tenir compte des obligations de transparence prévues par ailleurs, sauf à retirer sa substance à la définition de l'hypertrucage. En tout état de cause, cela souligne, d'une part, que la définition de l'hypertrucage n'est vraisemblablement pas, en elle-même, le support idoine de déploiement d'une législation substantielle et, d'autre part, que l'obligation de marquage devrait être affectée, en cas de méconnaissance, d'un effet substantiel sur le contenu lui-même, dépassant la seule sanction contre le déployeur (ce qui sera discuté dans la partie 7 s'agissant de l'effet à attacher au non-respect du marquage).

Une dernière interrogation structurante porte sur le sens des mots « *authentiques ou véridiques* », qui conduit à s'interroger sur les frontières de la ressemblance et de la perception dans le cadre de systèmes en constante évolution et aux fortes potentialités. Si on revient au sens même des termes « *authentique* » et « *véridique* », celui-ci désigne ce qui est conforme à la vérité et celui-là ce qui émane véritablement de l'auteur ou dont on a établi avec certitude la provenance[70]. Cette caractérisation est également déterminante pour le champ d'application de l'article 50 comme le montre le cas du cinéma : les IAG peuvent être utilisées dans un film pour de multiples fins, d'ampleurs très différentes, qu'il s'agisse d'une simple retouche ou de la modification, voire de la génération de scènes. Or, en entrant dans une salle de cinéma, le spectateur est pleinement conscient qu'il s'apprête à voir une œuvre de fiction : c'est un contrat tacite qui, de manière ancienne, postule l'acceptation du recours à des montages, des doublages, l'insertion d'effets spéciaux, voire des techniques de manipulation destinées à renforcer l'expérience vécue. On peut alors considérer qu'il n'y a aucun risque de perception à tort comme authentique ou véridique, de sorte que le

[63] Art. L. 97 du code électoral.
[64] Art. L. 121-1 et suivants du code de la consommation.
[65] Art. 441-1 et suivants et 313-1 et suivants, respectivement, du code pénal.
[66] Nous verrons dans la partie 3 les limites concrètes de cette approche.
[67] Cass. Com., 4 octobre 2016, pourvoi n° 14-22.245, Bull. 2016, IV, n° 128.
[68] A l'inverse, et c'est notable, du nouveau délit de l'article 226-8-1, ainsi que des dispositions de l'article 5 de la directive (UE) 2024/1385 du Parlement européen et du Conseil du 14 mai 2024 sur la lutte contre la violence à l'égard des femmes et la violence domestique.
[69] Doc. C(2025) 5052 final du 29 juillet 2025, paragr. 71 : « *Le marquage visible des hypertrucages et des dialogueurs réduit le risque de tromperie susceptible de survenir une fois que le contenu généré par l'IA est diffusé au public et diminue le risque d'effets d'altération préjudiciables sur la formation de l'opinion et des convictions et sur le comportement de la personne.* »
[70] Définitions du dictionnaire de l'Académie française, 9e éd.

film dont tout ou partie des scènes a été manipulé ou généré par IA pourrait échapper à la définition de l'hypertrucage. Néanmoins, cet exemple montre deux limites de cette approche et la **nécessité d'une démarche probablement plus casuistique, tenant compte du contexte et de l'objet de l'utilisation de l'IAG** :

- Le cas du doublage vocal s'inscrit dans une démarche visant à rendre accessible le film à un public plus large, ne comprenant pas la langue d'enregistrement initial. Le doublage offre une expérience plus immersive que le surtitrage mais il ne fait pas partie intégrante de la démarche créative, bien qu'il suppose la maîtrise de l'art de l'interprétation pour parvenir à rendre l'effet vocal nécessaire. Le doublage a lui-même pour objectif de rendre la scène plus authentique malgré le changement de langue. On peut alors pencher pour l'idée que, si le film n'est pas hypertruqué, le doublage (c'est-à-dire le son surajouté) constitue un hypertrucage s'il est fait appel à une technique susceptible d'être ainsi qualifiée. Il s'inscrit toutefois dans une œuvre de fiction et on pourrait considérer que l'hypertrucage n'est qualifié que si la voix doublée est artificielle et tend à reproduire la voix d'une personne identifiée (soit l'acteur originel, soit le doubleur habituel, dont la voix est connue du spectateur) ; ce ne serait alors pas le cas d'une voix entièrement artificielle ;
- On peut également envisager la situation spécifique des documentaires, parfois qualifiés de « documentaires de création », qui assument une démarche qui n'est pas purement informative et journalistique, faisant appel à la reconstitution de scènes, d'images, de sons. La démarche n'est pas née avec l'IA générative mais c'est alors le lieu et le contexte de visionnage qui deviennent déterminants dans la reconnaissance du risque d'une perception erronée de son authenticité : le risque qu'il y ait une mauvaise interprétation de la nature du documentaire au regard de ses modes de diffusion qualifie l'hypertrucage. La question prend également une autre dimension si le documentaire alterne des images originales (éventuellement d'archives) et d'autres manipulées ou générées par IA, sans qu'il soit possible, lors d'un visionnage dans des conditions standards, de distinguer les unes des autres. Si la qualification d'hypertrucage est liée à l'utilisation de l'IAG, on ne peut oublier qu'une telle pratique d'intégration d'images d'archives authentiques n'est pas innovante, y compris dans des films. Cela interroge sur cette qualification qui revient à tirer des conséquences du seul choix technique, le recours à l'IA étant une condition de cette qualification d'hypertrucage.

Pourtant, en prévoyant à l'article 50, paragraphe 4, une limitation de l'obligation d'étiquetage pour les contenus faisant partie d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue, le législateur européen a clairement entendu que ces contenus entrent dans le champ de l'hypertrucage.

Ces cas montrent que la définition juridique de l'hypertrucage ne peut se confondre, malgré les termes retenus, avec une approche sociologique[71] : la clarté et l'intelligibilité de la règle de droit ne peut se satisfaire d'une approche excessivement relativiste et contextualiste. Retenir le risque qu'un contenu soit perçu à tort comme authentique ou véridique implique donc de s'interroger sur la représentation avec le réel ou l'attribution réelle à un auteur. Le rapport d'un hypertrucage à la vérité est complexe : « *Si les deepfakes ne reprennent pas toujours la matérialité des supports de l'image, par exemple quand ce sont des visuels produits à partir de peintures, ils continuent à signifier possiblement pour celles et ceux qui les regardent, dans un système sémiologique visuel. Ainsi,* ***les deepfakes ne mentent pas mais ils ne disent pas non plus totalement la vérité*** »[72].

Il y a alors deux manières d'appréhender ce risque : soit par la prise en compte du degré de ressemblance, soit par l'intégration des risques liés à la pratique. La première méthode renvoie au critère précédent de la ressemblance, mais tient compte du poids relatif de l'ensemble des ressemblances par rapport aux différences. La seconde apprécie la condition à la lumière des différents risques auxquels les hypertrucages sont susceptibles d'exposer les personnes (utilisateurs et personnes dont les attributs sont repris) et la société, dont une étude du Parlement européen faisait la présentation suivante[73] :

---

[71] European Parliamentary Research Service, Scientific Foresight Unit, *Tackling deepfakes in European policy*, doc. PE 690.039, juillet 2021, p. 71.

[72] Nicole Pignier, « L'énonciation à l'épreuve de l'« I.A. ». Qu'est-ce qu'énoncer veut dire ? », *Interfaces numériques*, 2022, vol. 11, n°2.

[73] Etude précitée de juillet 2021, p. 29 (traduction des auteurs du présent rapport). Cette présentation était elle-même inspirée de l'article d'Aengus Collins, *Forged Authenticity : Governing Deepfake Risks*, Lausanne, EPFL International Risk Governance Center, 2019.

Dommage psychologique	Dommage financier	Dommage sociétal
(S)extorsion Diffamation Intimidation Harcèlement Perte de confiance	Extorsion Vol d'identité Fraude (par ex. assurance/paiement) Manipulation du cours des actions Atteinte à l'image de marque Dommage réputationnel	Manipulation de l'information Atteinte à la stabilité économique Atteinte au système judiciaire Atteinte à la science Erosion de la confiance Atteinte à la démocratie Manipulation des élections Atteinte aux relations internationales Atteinte à la sécurité nationale

Quelques exemples permettent de comprendre l'incidence de ce choix de méthode :

- S'agissant d'une reconstitution historique par IA, on peut se demander si le risque est caractérisé s'agissant d'un contenu audio reconstituant le discours du 18 juin du général de Gaulle[74] : dès qu'il est connu qu'il n'existe aucun enregistrement authentique de ce discours, une interprétation stricte du RIA pourrait porter à estimer que le risque n'est pas caractérisé concernant une reconstitution fidèle à la réalité historique[75]. Il en ira d'autant plus ainsi pour des images reconstituant des périodes plus anciennes qui n'ont pu être filmées. Néanmoins, l'effet utile du règlement implique manifestement de l'inclure, dans un contexte où, non seulement, il peut être délicat pour le spectateur moyen de déceler des manipulations qui seraient à cette occasion introduites, mais surtout où il est probable qu'un spectateur non averti ne soit pas conscient que l'image et le son qui lui sont présentés ont été générés par IA, quand bien même le texte serait authentique (voire *a fortiori* dans cette mesure) ;
- S'agissant d'une œuvre de fiction, comme un film, même si elle est ancrée dans le réel, il peut relever de l'objectif même de la production de lui donner une dimension d'authenticité et de réalité. Les conditions de sa diffusion, qui conduisent à ce que le spectateur sache d'emblée qu'il va être exposé à une œuvre de fiction, permettent alors de considérer qu'il n'existe pas de risque réel de manipulation, au-delà du procédé artistique. Mais les termes mêmes du RIA conduisent, ainsi qu'il a été dit, à l'inclure parmi les hypertrucages ;
- Plus délicat est le cas du documentaire qui utilise l'IAG pour reconstituer des visages et voix de témoins souhaitant conserver l'anonymat, sans perdre, contrairement aux techniques anciennes, l'expressivité du témoignage. Dans ce cas, le recours à l'IAG est en principe revendiqué et affiché, pour protéger tout risque qu'une personne ressemblante avec la personne générée ne puisse être inquiétée, de sorte que les enjeux de transparence sont intégrés au processus même. En l'absence d'intention manipulatrice, ni même, par définition au regard de l'objectif, de ressemblance avec les personnes existantes, la qualification d'hypertrucage semble devoir être retenue dès lors qu'il y a ressemblance avec une réalité globale, et en particulier celle des émotions que le recours à l'IAG permet précisément de conserver[76].

La prise en compte du dommage pouvant être causé doit conduire à intégrer dans la définition de l'hypertrucage, à titre de critère alternatif, le risque de manipulation en résultant. A cet égard, la lettre de mission invitait à s'interroger sur le point de savoir si l'intention de tromper pouvait être mobilisée dans la définition de l'hypertrucage. Il est, en tout état de cause exclu qu'il se substitue à cette approche objective du RIA dans la mesure où la prise en compte d'une intention dolosive, au demeurant potentiellement lourde à établir, est de nature, si elle était exigée en tout état de cause, à restreindre la portée de la définition de l'hypertrucage à des cas où il est par ailleurs sanctionné. En revanche, outre que la démonstration de l'intention de tromper devrait, en principe, suffire à caractériser le troisième critère de

[74] Cette reconstitution ayant été effectuée, en l'espèce, par le journal *Le Monde* avec des mentions assurant sa transparence. Voy. Charles-Henry Groult, Benoît Hopquin, « L'appel du 18 juin du général de Gaulle reconstitué pour la première fois », 18 janvier 2023.

[75] Voy. à ce propos les éléments développés par Pierre Saint-Germier, « Clones, filtres, et fakes. L'éthique de la conversion vocale à l'ère de l'intelligence artificielle », *L'étincelle. Le journal de la création à l'Ircam*, 2024, n° 24, p. 46.

[76] Réjane Hamus-Vallée, « Comment montrer sans révéler ? L'anonymat des entretenus filmés entre enjeux éthiques, scientifiques et artistiques », *Revue française des méthodes visuelles* [En ligne], 7, 2023.

la définition de l'hypertrucage, les risques manipulatoires liés à la technologie de l'IA générative[77] sont suffisamment importants pour conférer un large champ à l'obligation de transparence associée à cette qualification.

Néanmoins, le législateur ayant entendu intégrer les œuvres artistiques dans le champ de l'hypertrucage, limiter l'appréciation du critère du risque d'une perception erronée à l'intention constitue une inversion de la logique retenue par le RIA. Une approche plus empirique conduirait, à l'inverse, à tenir compte de l'intention afin de mieux appréhender le contexte d'exposition au contenu manipulé ou généré par IA pour qualifier l'hypertrucage : une telle approche aurait permis de limiter substantiellement la portée de l'obligation de transparence dans les secteurs culturels et créatifs, ce qui, compte tenu du contexte d'exposition à ces contenus, n'aurait pas privé d'effet utile le règlement. Pour autant, il ne semble pas que ce soit le choix fait par le RIA, en particulier du fait de l'insuffisante prise en compte des enjeux de ces secteurs dans la définition du régime juridique des hypertrucages. Il en résulterait en outre une difficulté à caractériser une telle intention, qui limiterait considérablement la portée du RIA. Au regard des termes du règlement, **l'intention de tromper ne peut être qu'un critère alternatif permettant de caractériser l'hypertrucage**, et non un critère exclusif d'appréciation du risque d'une perception erronée du contenu.

Comme cela a déjà été indiqué, la mission estime que la notion d'hypertrucage doit, pour conférer son effet utile au RIA, avoir une portée large. **Par conséquent, elle considère que l'appréciation du risque que le contenu soit perçu à tort comme véridique ou authentique devrait conduire à retenir cette qualification dès lors que ce risque est établi soit par les caractéristiques intrinsèques du contenu, du fait du degré d'une ressemblance avec un contenu véridique ou authentique, soit du fait de l'intention de manipuler le public à travers l'utilisation de ressemblances avec un tel contenu**. C'est, au demeurant, également la position proposée par la Commission européenne dans ses lignes directrices sur l'article 50 du RIA (paragr. 114).

[77] Ces risques seront développés dans la suite du présent rapport.

2. LA STRUCTURATION DES MARCHES

Les hypertrucages correspondent à un ensemble hétérogène de contenus synthétiques mobilisant des techniques différentes au service d'usages multiples. Ils s'intègrent dans une économie de la génération de contenus, elle-même sous-ensemble de l'essor de l'IA pour des applications diverses, certaines menaçantes ou répréhensibles, d'autres sources d'opportunités pour les filières culturelles.

2.1 Une nouvelle économie

La croissance attendue de l'intelligence artificielle et des hypertrucages est difficile à quantifier, du fait d'un marché assez peu lisible et du manque d'informations disponibles. Les estimations chiffrées présentées dans ce chapitre proviennent de trois familles de sources de fiabilité inégale : des publications académiques et institutionnelles indépendantes, des études sectorielles commandées par des acteurs concernés et des études de marché commerciales. Quelques chiffres convergents peuvent cependant fournir des ordres de grandeur.

2.1.1. Une forte croissance attendue

Le récent *Generative AI Outlook Report* du Joint Research Centre (JRC) de la Commission européenne[78] fournit un cadrage quantitatif du *positionnement européen dans l'écosystème mondial de l'IA générative*. Sur les 72 000 acteurs identifiés à l'échelle mondiale et leurs 149 000 activités (recherche, brevets, business), l'Union européenne ne représente que 7 % des acteurs, loin derrière la Chine (60 %) et les États-Unis (12 %), au coude-à-coude avec la Corée du Sud (6 %). Cette troisième place masque toutefois un profil européen atypique : alors que la moyenne mondiale est portée par l'activité business, l'UE répartit ses acteurs en 37 % business, 33 % innovation et 31 % recherche, avec une part de recherche significativement supérieure. L'Europe est ainsi le deuxième marché mondial pour les publications scientifiques dans le domaine de l'IA générative, derrière la Chine, mais ne représente que 2 % des dépôts de brevets prioritaires, contre 80 % pour la Chine et 7 % chacun pour les États-Unis et la Corée du Sud. Au sein de l'UE, l'activité se concentre en Allemagne (33 % des acteurs déposant des brevets), en France (12 %), aux Pays-Bas et en Espagne (9 % chacun). Cet écart entre excellence académique et faiblesse commerciale est aggravé par un déficit structurel de capital-risque vis-à-vis des États-Unis. Le rapport du JRC identifie en réponse les « *AI Factories* » (fabriques d'IA) et les « *Common European Data Spaces* » (espaces de données européens communs) comme leviers stratégiques destinés à combler l'écart d'infrastructure et à éviter un décrochage compétitif durable sur les couches technologiques en amont qui structurent en retour le marché des hypertrucages.

Le marché *mondial* de l'intelligence artificielle appliquée aux médias et au divertissement, quels que soient les usages, serait estimé à environ 12 milliards de dollars en 2025 et pourrait atteindre 68,8 milliards de dollars en 2036, soit un taux de croissance annuel composé (TCAC) d'environ 17 %[79]. **Le marché *européen* de l'IA appliquée à la création de contenus est estimé à 3,7 milliards de dollars en 2024, avec une croissance annuelle moyenne attendue de 31,8 % jusqu'en 2030 atteignant 19,6 milliards de dollars**[80].

Les effets économiques de l'IA générative ne se limitent pas à la compression des coûts : ils peuvent aussi modifier la demande et la répartition des revenus. Sur une plateforme chinoise d'externalisation artistique, une expérience naturelle liée à la diffusion soudaine d'un modèle de génération d'images de style animé montre que le prix moyen des commandes de dessins animés a chuté de 64 %, tandis que le nombre de commandes a augmenté de 121 % et que le revenu total du marché concerné a progressé de 56 %. Ce gain ne signifie pas que chaque créateur gagne davantage : il désigne le revenu agrégé généré sur la plateforme pour ce segment. La hausse provient surtout des commandes personnelles à bas prix, auparavant limitées par le coût de production, alors que les commandes commerciales restent relativement stables. Les principaux bénéficiaires sont les créateurs déjà présents sur la plateforme, qui remportent 97 % des

[78] Joint Research Centre, *Generative AI Outlook Report: Exploring the Intersection of Technology, Society and Policy*, JRC142598 (Publications Office of the European Union, 2025), https://publications.jrc.ec.europa.eu/repository/handle/JRC142598.

[79] Future Market Insights, 'AI in Media and Entertainment Market | Global Market Analysis Report - 2036', 2025, https://www.futuremarketinsights.com/reports/ai-in-media-and-entertainment-market.

[80] Grand View Research, *Europe Generative AI in Content Creation Market Outlook* (Grand View Research, 2025), https://www.grandviewresearch.com/horizon/outlook/generative-ai-in-content-creation-market/europe.

commandes après l'arrivée de l'IA, ainsi que la plateforme elle-même via ses commissions ; les usagers bénéficient, eux, de prix plus faibles. En revanche, l'étude ne montre pas une prise de contrôle du marché par de nouveaux entrants utilisant l'IA, ni une perte nette de commandes pour les créateurs établis, mais plutôt une baisse du revenu par commande compensée par une forte hausse du volume[81]. L'IA n'agit donc pas uniquement comme un facteur de compression des coûts ; en abaissant les barrières à l'entrée, elle stimule la demande et peut augmenter le volume total de revenus. Dans les médias d'information, 90 % des rédactions déclarent déjà utiliser l'IA dans la production de contenus, et une majorité anticipe des gains de productivité[82]. Des expériences à grande échelle sur des plateformes numériques montrent que l'utilisation d'outils génératifs pour produire des titres ou des métadonnées peut accroître la consommation, avec +7,1 % de visionnage et +4,1 % de durée de consommation dans certains cas[83].

Le marché de l'IA générative appliquée à la *musique* connaît une croissance forte, mais les estimations divergent fortement selon les cabinets. Le cabinet MRFR évalue ce marché à 2,38 milliards de dollars en 2024, avec une projection à 18,47 milliards de dollars d'ici 2034, soit un taux de croissance annuel composé (TCAC) de 22,7 %[84]. *Grand View Research* propose une fourchette plus faible de 570 millions de dollar en 2024, projeté à 2,8 milliards en 2030, avec un TCAC de 30 %[85]. *ResearchAndMarkets* donne une estimation proche de *Grand View Research* : 642,8 millions de dollars en 2024, projeté à 3 milliards de dollars d'ici 2030, TCAC de 29,5 %[86]. Cette divergence d'un facteur un à quatre selon les sources reflète des définitions de périmètres différentes.

Le marché mondial de l'IA dans la *production cinématographique* était estimé à 3,24 milliards de dollars en 2024, projeté à 23,54 milliards d'ici 2033, avec un TCAC de 25,4 %[87]. Le segment de la production domine avec 38 % de parts de marché, porté par l'automatisation des effets spéciaux (VFX), de la rotoscopie, du nettoyage d'images et du *tracking* caméra. Les studios déclarent des économies de 30 % sur les budgets VFX grâce aux technologies de type hypertrucage (rajeunissement, doublures numériques, synchronisation labiale)[88]. L'IA a également réduit le temps de montage jusqu'à 40 %, selon certaines sources sectorielles, permettant des cycles de production plus courts et des coûts réduits[89]. Le marché global des VFX, évalué à 10,7 milliards de dollars en 2024 et projeté à 18,75 milliards d'ici 2033[90], intègre de manière croissante l'IA dans ses pipelines. Le marché spécifique de l'hypertrucage IA (tous usages confondus) atteint 1,14 milliard de dollars en 2025, projeté à 8,11 milliards d'ici 2030 (TCAC de 48 %), dont 44,86 % dans le segment divertissement[91]. *Deep Market Insights* estime le marché mondial des plateformes de concerts virtuels à 3,85 milliards de dollars en 2024 et 12,45 milliards de dollars en 2030, avec des revenus issus de la billetterie numérique, des abonnements, du sponsoring, des NFT et du marchandisage virtuel[92].

2.1.2. La fabrique industrielle des hypertrucages

---

[81] Kaichen Zhang et al., 'The Impact of Generative Artificial Intelligence on Market Equilibrium: Evidence from a Natural Experiment', arXiv:2311.07071, preprint, arXiv, 10 October 2024, https://doi.org/10.48550/arXiv.2311.07071.

[82] 'Journalism, Media, and Technology Trends and Predictions 2025 | Reuters Institute for the Study of Journalism', 9 January 2025, https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2025; 'An Overview of the Launch of JournalismAI's Generating Change Report', accessed 4 June 2026, https://www.ftstrategies.com/en-gb/insights/journalismai-launch-event.

[83] Xinyi Zhang et al., 'The Value of AI-Generated Metadata for UGC Platforms: Evidence from a Large-Scale Field Experiment', arXiv.Org, 24 December 2024, https://arxiv.org/abs/2412.18337v1.

[84] Market Research Future, *Generative AI in Music Market Growth Report 2035* (Market Research Future, 2025), https://www.marketresearchfuture.com/reports/generative-ai-in-music-market-31943.

[85] Grand View Research, *Generative AI in Music Market (2024-2030): Size, Share & Trends Analysis Report*, GVR-4-68040-418-2 (Grand View Research, 2024), https://www.grandviewresearch.com/industry-analysis/generative-ai-in-music-market-report.

[86] Market Research Future, *Generative AI in Music Market Growth Report 2035*.

[87] Grand View Research, *AI in Filmmaking Market Size, Share & Trends Analysis Report, 2033* (Grand View Research, 2025), https://www.grandviewresearch.com/industry-analysis/ai-filmmaking-market-report.

[88] 'GenAI's Leading Role in Entertainment', Morgan Stanley, accessed 4 June 2026, https://www.morganstanley.com/insights/articles/ai-in-media-entertainment-benefits-and-risks.

[89] Market.us, *AI in Film Market Size, Share, Trends* (Market.us, 2025), https://market.us/report/ai-in-film-market/.

[90] IMARC Group, *Visual Effects (VFX) Market Size, Share, Trends and Forecast by Component, Product, Technology, Application, and Region, 2025-2033* (IMARC Group, 2025), https://www.imarcgroup.com/visual-effects-market.

[91] Mordor Intelligence, *Deepfake AI Market Size & Share Analysis: Growth Trends & Forecasts* (Mordor Intelligence, 2025), https://www.mordorintelligence.com/industry-reports/deepfake-ai-market.

[92] Deep Market Insights, *Virtual Concert Platform Market Research Report* (Deep Market Insights, 2025), https://deepmarketinsights.com/report/virtual-concert-platform-market-research-report; Visionary Data Reports, *Virtual Concert Platform Market* (Visionary Data Reports, 2025).

Derrière ces estimations macro-économiques globales et parfois divergentes, l'écosystème technique et industriel permettant la fabrication des hypertrucages doit faire l'objet d'une attention spécifique. De manière visible, des individus manipulent des outils et partagent les résultats, fascinants ou inquiétants, de leurs expériences sur les réseaux sociaux. Mais ces « *face swap* » (substitution de visage) ou « *voice cloning* » (clonage de voix) « amateurs » produits par le grand public ne sont jamais créés *ex nihilo* ; ils sont le fruit d'un écosystème beaucoup moins visible ; l'articulation entre techniques, entreprises et usagers caractérise la « fabrique » des hypertrucages.

Pour mieux comprendre le cadre de cet écosystème, un travail de recherche (Butraud, 2026) a établi une base de données originale à partir de *Crunchbase*[93]. Cette base généraliste fait ressortir en son sein un ensemble de 89 061 entreprises liées à l'IA, intervenant à différents niveaux de la chaîne de valeur : développement de modèles de génération, fourniture d'API, services d'inférence, outils de création et de diffusion. Afin de cartographier cet écosystème, les 89 061 entreprises ont été classées grâce à un algorithme de factorisation textuelle sur les descriptions de ces sociétés, permettant de les grouper sur la base de leur similarité. Le marché a ainsi été structuré en 32 clusters thématiques.

Afin d'isoler spécifiquement les acteurs de la fabrique des contenus synthétiques, 5 clusters (parmi les 32) ont été extraits : le « *Gaming* » (jeu vidéo), les médias et le divertissement, le design et la visualisation, les assistants virtuels (XR), le marketing. Ces 5 clusters regroupent 7 942 entreprises réparties dans 112 pays. **La génération de contenus (dont les hypertrucages font partie) représente ainsi environ 10 % de l'écosystème mondial de l'IA en termes de nombre d'entreprises (et non de volume d'activités ou de chiffre d'affaires des acteurs).** Cette part est nettement plus élevée au Japon (12,5 % en raison de la place de l'industrie graphique) mais identique aux Etats-Unis (8,8 %) et en Europe (8,1 %).

Sur le plan géographique, et toujours en termes de nombre global d'entreprises dans la génération de contenus, l'observation confirme une domination écrasante des Etats-Unis, qui rassemblent 3 022 entreprises dédiées aux contenus synthétiques contre 1 371 pour l'Europe. Le continent asiatique s'impose comme un pôle de second plan, porté par l'Inde avec 460 entreprises, suivi par le Japon (333 entreprises) et la Chine (225 entreprises).

L'analyse des métropoles montre que l'expertise culturelle historique permet de fracturer l'hégémonie américaine sur des briques technologiques très spécialisées :

- Londres s'impose comme le leader mondial incontesté du cluster « *Gaming* » avec 27 entreprises, devançant les hubs américains.
- Tokyo, se hisse à la première place mondiale (à égalité avec Londres) sur le secteur du Design graphique et de la Visualisation (68 entreprises) devant San Francisco et New York.

La recherche menée met en évidence un financement différencié selon les clusters. **Bien que la génération de contenus bénéficie d'une visibilité médiatique écrasante, elle est structurellement sous-financée par rapport au reste de l'IA.** A titre d'exemple, le financement médian d'une entreprise américaine de génération de contenus synthétique chute à 2,5 millions de dollars, contre 3,75 millions pour une entreprise d'IA « autre ». Les entreprises semblent donc plus « frileuses » à investir dans la génération de contenus que dans le reste de l'IA (agroalimentaire, dispositifs, médicaux, formation, analyse de données ..).

De plus, le **marché des hypertrucages se caractérise par une bicéphalie économique** : **deux marchés coexistent et se développent en parallèle, avec des dynamiques, des externalités et des modèles économiques très différents. Le premier est un marché d'entreprise à entreprise (B2B) en forte croissance**, dominé par les usages « *corporate* » : avatars marketing, doublage IA multilingue, restauration d'archives, formation interne, prototypage publicitaire. McKinsey[94] estimait que 75 % des entreprises ayant déployé l'IA générative à l'échelle l'utilisent pour la création de contenu, avec des gains de productivité documentés de 30 à 60 % sur les fonctions concernées. Deloitte[95] rapportait que 22 % des organisations

[93] Base de données internationale qui recense des informations détaillées sur les entreprises (secteur d'activités, financement, acquisitions, localisation, etc.) et leurs écosystèmes économiques. Document de travail.
[94] Alex Singla et al., *The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value* (McKinsey & Company, 2024), https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024.
[95] Deborshi Dutt et al., *The State of Generative AI in the Enterprise: Now Decides Next* (Deloitte AI Institute, 2024), https://www.deloitte.com/gr/en/services/consulting/research/state-of-generative-ai-in-enterprise.html.

ayant adopté l'IAG mentionnaient la création de contenus comme premier cas d'usage. Goldman Sachs[96] estimait l'effet de productivité de l'IAG à 7 % du PIB mondial sur dix ans, dont une portion non négligeable concentrée sur les industries créatives. Les enquêtes d'usage déclaratif d'Adobe indiquent que la majorité des créateurs professionnels pensent que les outils d'IA générative peut faire gagner du temps et de l'argent et aide à produire de nouvelles idées[97]. **Le second est un marché grand public** mêlant des usages anodins ou créatifs (filtres TikTok, applications de transformation faciale, Wombo, Reface) à des usages malveillants à grande échelle.

L'asymétrie entre ces deux marchés est documentée mais reste imprécise faute de mesures unifiées. **En volume d'usages, le pôle grand public semble dominer** : les applications de « *face-swap* » (substitution de visage), d'avatars et de filtres IA atteignent des centaines de millions de téléchargements (plus de 200 millions pour Wombo et plus de 300 millions pour Reface) tandis que les contenus abusifs circulent aussi à grande échelle. Les estimations publiques disponibles mentionnent toutefois plutôt des millions d'hypertrucages : certaines sources évoquent une hausse de 500 000 à 8 millions d'hypertrucages en deux ans ; par exemple, Security Hero estimait 95 820 vidéos hypertruquées en ligne en 2023[98]. **En valeur monétaire identifiable, le rapport s'inverse** : le marché d'entreprise à entreprise (B2B) représente l'essentiel des revenus traçables (estimés entre 1 et 3,5 milliards de dollars en 2025[99]). Toutefois, ces chiffres ne mesurent pas strictement le marché B2B de la production d'hypertrucages culturels mais plutôt un ensemble mêlant logiciels de génération, avatars, voix synthétiques, services, détection, authentification et usages de sécurité. Le marché B2B capte des revenus identifiables via abonnements, licences, API, services d'avatars, doublage ou détection, tandis que le marché grand public combine monétisation plus diffuse (publicité, abonnements freemium, achats in-app) et externalités négatives beaucoup moins bien comptabilisées.

2.1.3. L'émergence du marché des hypertrucages en tant que service (« deepfake as a service », DFAAS)

Les motivations des acteurs pour fabriquer des hypertrucages sont diverses ; certains développent des usurpations d'identité à des fins d'escroquerie financière (promotion de cryptomonnaies ou d'investissements risqués par exemple). La production massive de contenus hypertruqués à visée commerciale peut également avoir pour but d'alimenter des sites internet à vocation lucrative. De tels services, monétisés notamment par la publicité en ligne, sont susceptibles d'attirer un nombre conséquent de visiteurs ; **d'autres acteurs, enfin, monétisent directement une multitude de services et un véritable marché des hypertrucages en tant que service émerge.**

Rappelons tout d'abord que la chaîne de valeur d'un système d'IA peut être segmentée en trois principaux blocs [100]. En amont, des ressources sont nécessaires (main d'œuvre qualifiée, données en quantité, puissance de calcul et de stockage). Le développement correspond quant à lui aux phases de pré-entraînement, d'affinage sur des données spécialisées et d'ancrage du modèle dans l'actualité avec des données fraîches. Enfin, en aval, les modèles sont déployés et commercialisés. La phase de développement n'est pas source de valorisation en elle-même, c'est la phase de déploiement qui l'est, lorsque le modèle s'intègre dans un système commercialisable (via des activités de création de logiciels et applications pour interagir avec les modèles).

---

[96] Goldman Sachs, 'Generative AI Could Raise Global GDP by 7%', 5 April 2023, https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent.

[97] Adobe, *State of Creativity Report 2024* (Adobe, 2024), https://www.adobe.com/content/dam/cct/creativecloud/business/teams/whitepapers/dotcom-116728/Adobe_StateofCreativity_Report_2024_FV_UE.pdf; Adobe Communications Team, 'Adobe's AI and the Creative Frontier Study Reveals Creators' Views on the Opportunities and Risks of Generative AI', Adobe Blog, 8 October 2024, https://blog.adobe.com/en/publish/2024/10/08/adobes-ai-creative-frontier-study-reveals-creators-views-opportunities-risks-generative-ai.

[98] '2023 State Of Deepfakes: Realities, Threats, And Impact | Security Hero', accessed 4 June 2026, https://www.securityhero.io/state-of-deepfakes/.

[99] DataInsightsMarket, *AI Dubbing Tools Market's Evolution: Key Growth Drivers 2026–2034* (DataInsightsMarket, 2025), https://www.datainsightsmarket.com/reports/ai-dubbing-tools-1372047; Mordor Intelligence, *Deepfake AI Market Size & Share Analysis: Growth Trends & Forecasts*; Future Market Insights, 'AI in Media and Entertainment Market | Global Market Analysis Report - 2036'.

[100] Thomas Hoppner and Luke Streatfeild, 'ChatGPT, Bard & Co.: An Introduction to AI for Competition and Regulatory Lawyers', *An Introduction to AI for Competition and Regulatory Lawyers (February 23, 2023)* 9 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4371681.

Ces activités monétisables sont déployées par deux grands types d'entreprises[101]. D'un côté les opérateurs qui agissent sur le marché – dominé par un oligopole d'entreprises américaines – du développement des modèles de fondation. Par ailleurs, des opérateurs tiers, moyennant paiement aux développeurs de modèles de fondation, réalisent leurs propres applications ou spécialisent des modèles pour des services qu'elles facturent à leur tour à leurs clients.

Les hypertrucages en tant que service qui se développent sont un dérivé du modèle « SAAS » (« *sofware as a service* » (logiciel à la demande) dans lequel un logiciel, hébergé sur le cloud, est disponible pour différents services) appliqué aux hypertrucages. L'émergence du marché des hypertrucages en tant que service s'inscrit dans une dynamique plus large identifiée dès 2021 par l'étude STOA[102], qui distinguait quatre tendances structurelles aujourd'hui pleinement confirmées. D'abord, l'apparition de *plateformes d'offre et de demande* dédiées : des sites de marché (« *marketplaces* »), souvent hébergés sur des forums spécialisés ou sur des plateformes de messagerie, où des utilisateurs commandent à des opérateurs anonymes des hypertrucages sur mesure, y compris à des fins d'imagerie intime non consentie. Ensuite, une *marchandisation accélérée* des outils eux-mêmes : les logiciels avancés de génération sont librement distribués, accompagnés de tutoriels et de documentation, tandis que des applications smartphone et des bots de messagerie permettent à des utilisateurs sans compétence technique de générer des hypertrucages en quelques clics. Troisièmement, l'émergence de véritables entreprises offrant des hypertrucages en tant que service (« *deepfake-as-a-service companies* ») : l'étude STOA citait déjà Synthesia et Rephrase comme exemples précurseurs en 2021, et le marché s'est depuis structuré. Enfin, une *réduction continue des données* d'entrée nécessaires ; là où les premières architectures exigeaient des centaines de photographies ou des heures d'enregistrement, les modèles contemporains se contentent d'une seule image ou de quelques secondes de voix (cf. chap. 6), ce qui rend tout individu disposant d'une présence en ligne potentiellement cible. Ces quatre tendances se sont matérialisées en quatre ans et continuent de structurer l'offre.

Le modèle économique repose sur une redistribution de la chaîne de valeur audiovisuelle. **Les plateformes DFaaS cherchent à gagner de l'argent en transformant des modèles complexes d'IA générative en services accessibles** par crédit (« *pay per use* »), par abonnement, licences professionnelles ou selon une logique de monétisation indirecte via la publicité, démocratisant ainsi l'accès à des technologies de pointe auparavant exclusivement réservées aux très grosses entreprises. Les clients (studios, agences, médias, entreprises de formation) y gagnent en rapidité, en réduction des coûts et en capacité de personnalisation. Les fournisseurs d'infrastructure et de modèles, à l'image de plateformes comme Hugging Face pour l'écosystème de l'IA, captent une partie de la valeur en hébergeant, distribuant ou rendant déployables les modèles sous forme d'API[103]. La valeur ne vient donc pas seulement de la production d'un hypertrucage particulier, mais de l'automatisation et de la standardisation de tout un processus de création audiovisuelle.

La puissance et l'accessibilité de ces outils facilite les usages malveillants à grande échelle. Des publications malveillantes diffusées sur les réseaux sociaux utilisent l'image de personnalités et renvoient parfois vers des sites frauduleux faisant notamment la promotion de sites de cryptomonnaies. Les **hypertrucages à caractère sexuel** sont quant à eux **loin d'être des dérives marginales et s'inscrivent dans cette économie de DFaaS** fondée sur l'accessibilité technique, l'intermédiation par des plateformes et la captation massive d'attention. Ils fournissent en effet des outils de visibilité et de mise en relation entre créateurs d'hypertrucages et usagers tout en permettant de personnaliser les demandes. La valeur économique de ces services ne réside pas seulement dans l'hébergement de contenus visibles, mais dans l'animation de communautés capables de produire, personnaliser et monétiser des hypertrucages à la demande. MrDeepFakes, fermé en mai 2025, illustre ce modèle : le site hébergeait près de 70 000 vidéos, comptait environ 650 000 membres, générait des millions de visites mensuelles et permettait à ses utilisateurs d'échanger des conseils techniques ou de commander des hypertrucages payants sur mesure,

---

[101] Joëlle Farchy and Bastien Blain, *Rémunération Des Contenus Culturels Utilisés Par Les Systèmes d'IA*, 2025, https://strossburi.eu/wp-content/uploads/2025/05/CSPLA_Rapport-economique_Remuneration-des-contenus_Version-provisoire_Mai-2025.pdf.

[102] VMariëtte van Huijstee et al., *Tackling Deepfakes in European Policy*, EPRS_STU(2021)690039, STOA: Panel for the Future of Science and Technology (European Parliamentary Research Service, 2021), https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039.

[103] Une API (Application Programming Interface) est une interface qui permet à deux logiciels de communiquer entre eux. L'API est donc une sorte de prise de branchement logicielle : elle permet d'utiliser une technologie sans avoir à la construire soi-même. C'est par exemple une manière pour un logiciel d'interagir avec un modèle d'IA sans avoir à passer par un *chat.*

visant des célébrités mais aussi des personnes de leur entourage[104]. Depuis sa fermeture, le marché ne disparaît pas mais se fragmente : CBS News rapporte que la fermeture du site déplace ces communautés vers des espaces moins visibles, notamment Telegram, tandis que *The Guardian* note que des créateurs continuent à proposer leurs services après la disparition de MrDeepFakes[105]. Des enquêtes de Bellingcat et du *Guardian* ont par ailleurs relié des services comme ClothOff, Nudify, Undress et DrawNudes à un réseau opaque de domaines, comptes Telegram, sociétés-écrans et prestataires dont la piste mène notamment vers la Biélorussie et la Russie[106]. L'explosion de l'imagerie intime non consentie, très fortement genrée — les deepfakes à caractère sexuel viseraient presque toujours des femmes ou des jeunes filles — engendre des coûts considérables : coûts psychologiques pour les victimes, actions judiciaires, demandes de retrait, modération, gestion de crise et risques réputationnels ou commerciaux pour les personnalités publiques ciblées[107].

L'analyse précédente peut être synthétisée en distinguant cinq maillons de la chaîne de valeur des hypertrucages, de la production des modèles en amont jusqu'à la diffusion auprès du public en aval. Chaque maillon se caractérise par un degré de concentration, un modèle économique, une captation de valeur distincts.

[104] 'Deepfakes pornographiques : l'un des principaux sites ferme, un pharmacien canadien identifié', Pixels,Deepfakes, *Le Monde*, 7 May 2025, https://www.lemonde.fr/pixels/article/2025/05/07/deepfakes-pornographiques-l-un-des-principaux-sites-ferme-un-pharmacien-canadien-identifie_6603814_4408996.html.

[105] 0, 'AI-Generated Porn Site Mr. Deepfakes Shuts down after Service Provider Pulls Support - CBS News', 5 May 2025, https://www.cbsnews.com/news/ai-generated-porn-site-mr-deepfakes-shuts-down/; Michael Safi et al., 'Revealed: The Names Linked to ClothOff, the Deepfake Pornography App', Technology, *The Guardian*, 29 February 2024, https://www.theguardian.com/technology/2024/feb/29/clothoff-deepfake-ai-pornography-app-names-linked-revealed.

[106] Safi et al., 'Revealed'; Kolina Koltai, 'Behind a Secretive Global Network of Non-Consensual Deepfake Pornography', *Bellingcat*, 23 February 2024, https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/.

[107] 'Addressing Deepfake Image-Based Abuse | eSafety Commissioner', 24 July 2024, https://www.esafety.gov.au/newsroom/blogs/addressing-deepfake-image-based-abuse.

Tableau — La chaîne de valeur des hypertrucages et la captation de la valeur

Maillon	Acteurs typiques	Concentration	Modèle économique	Captation de valeur
Modèles fondationnels	OpenAI, Anthropic, Google, Meta, Mistral, Stability AI, Black Forest Labs	Extrême (moins de dix acteurs)	Vente d'API, licences entreprise	Très élevée par unité, coûts fixes massifs
Plateformes DeepFake as a service B2B	Synthesia, HeyGen, D-ID, Rephrase (Adobe), Hour One, Deepdub, ElevenLabs	Moyenne (quelques dizaines d'acteurs majeurs)	Abonnement, paiement à l'usage	Élevée, marges en croissance
Applications B2C grand public	Wombo, Reface, FaceApp, plusieurs milliers d'apps mobiles	Faible à moyenne	Freemium, publicité, achats in-app	Faible par unité, significative en agrégat
Marketplaces et forums spécialisés	Bots Telegram, forums Discord, plateformes anonymes	Hautement fragmentée, partiellement illicite	Transactions opaques, économie souterraine	Mal documentée
Plateformes de diffusion	YouTube, Spotify, Deezer, TikTok, Meta, X	Très élevée (oligopole mondial)	Publicité, abonnement	Très élevée, captation indirecte via l'audience

2.1.4. Des acteurs présents dans les industries culturelles

Dans leurs interactions avec les industries culturelles, on retrouve à la fois des développeurs de modèles de fondation qui proposent eux-mêmes des outils grand public et des services et d'autres entreprises qui s'appuient sur des modèles de fondation développés par les premiers pour proposer leurs propres services. Là aussi, certaines de ces sociétés se concentrent sur le marché grand public avec des interfaces simples et automatisées. D'autres se positionnent sur le marché professionnel en développant des services pour les studios, les plateformes ou les agences.

Des pôles de spécialisation apparaissent :

- Génération de textes : ChatGPT (outil d'OpenAI), Le Chat (outil de Mistral, désormais dénommé Vibe), Claude (outil d'Anthropic).
- Génération d'images et de vidéos : Midjourney, Runway, DALL-E, Pika, Luma, Synthesia, Rephrase, Sora (outil d'OpenAI), Hedra, Nano banana Pro (outil de Gemini - Google), Veo 3 (Google), Vibes (outil de Meta).
- Génération de musique : Suno, Udio, Mubert, AIVA, Boomy.
- Clonage vocal et synthèse de voix : ElevenLabs, Respeecher, Wellsaid, Voicemaker.
- Outils de postproduction et de doublage : Flawless, Papercut, HeyGen, Deepdub.

Ces plateformes représentent une révolution en abaissant les barrières techniques et financières de la production audio-visuelle. **Les hypertrucages en tant que service nouent une forme de collaboration différente de celle, classique, de sous-traitance entre deux acteurs ; il ne s'agit plus d'une prestation ponctuelle mais de plateformes qui automatisent l'ensemble du processus et proposent des outils clés en main** : Respeecher qui permet de restaurer la voix d'acteurs malades ou de ressusciter des figures

historiques, Synthesia pour les avatars vidéos et le doublage multilingue, facilitant l'exportation internationale des contenus à moindre coûts, D-ID pour l'animation et le rajeunissement facial offrant de nouvelles libertés scénaristiques aux réalisateurs, Deep brain pour les avatars interactifs et contenus éducatifs rendant l'apprentissage beaucoup plus immersif, Hour one pour la création d'avatars réalisés à des fins professionnelles, médias ou d'apprentissage en ligne (« *e-learning* »), Rephrase pour la production automatisée de vidéos marketing à partir d'avatars synthétiques, racheté par Adobe en 2023 et intégré depuis à la suite Adobe Express...

Les bandes-annonces conceptuelles ou bandes-annonces de fans constituent un autre usage de l'hypertrucage culturel : elles ne visent pas seulement à tromper, mais à matérialiser l'imaginaire des fans en donnant une forme audiovisuelle à des suites, relances (« *reboots* ») ou auditions (« castings ») fictives. Le phénomène précède l'IA générative : dès les débuts de YouTube, des montages amateurs comme Titanic 2, Jack's Back de VJ4rawr2 imaginaient le retour de Jack Dawson, retrouvé congelé au fond de l'océan puis ramené à la vie dans le monde contemporain ; cette vidéo aurait cumulé environ 53 millions de vues avant d'être bloquée par la 20th Century Fox[108]. Depuis l'arrivée des outils d'IA générative, le genre s'est industrialisé : des chaînes comme KH Studio ou Screen Culture produisent en série des bandes-annonces fictives mêlant extraits existants, images générées et voix synthétiques, parfois présentées comme bande-annonce initiale ou conceptuelle. Certaines vidéos obtiennent des millions de vues et peuvent être prises pour des annonces officielles, comme l'illustre une fausse bande-annonce de James Bond avec Henry Cavill et Margot Robbie, vue plus de 2 millions de fois[109]. Cette économie crée une tension pour les ayants-droit : selon Deadline, repris par The Verge et le Washington Post, certains studios, dont Warner Bros Discovery, Sony Pictures et Paramount, ont demandé à YouTube de rediriger vers eux une partie des revenus publicitaires de ces vidéos plutôt que d'en exiger systématiquement le retrait[110]. Ce choix révèle un arbitrage ambigu entre captation de revenus et protection de la marque, des artistes et du contrôle promotionnel des œuvres.

Dans la musique, les hypertrucages vocaux et visuels sont devenus à la fois des outils d'engagement communautaire et des objets de conflit juridique. Les fans produisent et diffusent des reprises (« AI *covers* »), remix ou montages vocaux générés par IA sur TikTok, YouTube et Spotify, parfois dans une logique d'hommage ou de promotion informelle, mais aussi en dehors de tout accord avec les ayants-droit. Le cas de Heart on My Sleeve, faux duo Drake/The Weeknd créé avec des voix générées par IA, illustre cette tension : devenu viral sur TikTok et les plateformes musicales, le titre a été retiré après intervention d'Universal Music Group[111]. Des formes visuelles similaires apparaissent aussi dans les cultures de fans, comme les faux selfies avec célébrités ou idoles, qui mettent en scène une intimité fictive et brouillent la frontière entre rencontre réelle, relation parasociale et fantasme numérique[112].

Face à ces usages, plusieurs acteurs cherchent à déplacer les pratiques d'imitation vers des environnements sous licences. Grimes a ainsi lancé Elf.Tech, qui autorise des tiers à utiliser une version IA de sa voix contre un partage 50/50 des redevances du producteur pour les sorties commerciales ; Holly Herndon avait déjà expérimenté un modèle proche avec Holly+, présenté comme un "jumeau" vocal IA gouverné par des règles de consentement et de partage[113]. Ce modèle se professionnalise avec des plateformes comme Kits.AI ou Voice-Swap, qui proposent des voix vérifiées, consenties et rémunérées, transformant la voix en service

---

[108] Jake Kanter, 'Inside YouTube's Weird World Of Fake Movie Trailers — And How Studios Are Secretly Cashing In On The AI-Fueled Videos', *Deadline*, 28 March 2025, https://deadline.com/2025/03/inside-youtube-fake-movie-trailers-1236352406/.

[109] 'James Bond Trailer Featuring Henry Cavill, Margot Robbie Receives 2 Million Views on YouTube — Unfortunately, It's Fake', People.Com, accessed 4 June 2026, https://people.com/fake-james-bond-trailer-henry-cavill-margot-robbie-receives-2-million-views-youtube-8634904.

[110] Herb Scribner, 'Fake Movie Trailers Were an Art Form. Then Came the AI Slop.', *The Washington Post*, 28 April 2025, https://www.washingtonpost.com/entertainment/movies/2025/04/28/fake-movie-trailers-ai-youtube/.

[111] Vincent Fagot, 'Drake and The Weeknd's fake AI song raises questions about respect for intellectual property', Economy,Music, *Le Monde*, 26 April 2023, https://www.lemonde.fr/en/economy/article/2023/04/19/drake-and-the-weeknd-s-fake-ai-song-raises-questions-about-respect-for-intellectual-property_6023514_19.html.

[112] 'Want Your Own "Celebrity Elevator Selfie"? Here's How TikTokers Are Doing It', 17 October 2025, https://dailydot.com/celebrity-elevator-tiktok-trend/.

[113] Jazz Monroe, 'Grimes Unveils Software to Mimic Her Voice, Offering 50-50 Royalties for Commercial Use', Pitchfork, 2 May 2023, https://pitchfork.com/news/grimes-unveils-software-to-mimic-her-voice-and-announces-2-new-songs/; Evan Minsker, 'Holly Herndon's AI Deepfake "Twin" Holly+ Transforms Any Song Into a Holly Herndon Song', Pitchfork, 14 July 2021, https://pitchfork.com/news/holly-herndons-ai-deepfake-twin-holly-transforms-any-song-into-a-holly-herndon-song/.

avec licence. YouTube a testé une logique similaire avec Dream Track, permettant à certains créateurs de Shorts de générer des extraits musicaux de 30 secondes avec les voix IA d'artistes volontaires comme John Legend, Sia ou Demi Lovato. Cette logique d'encadrement prend désormais la forme de plateformes fermées ou semi-fermées. En 2026, Spotify et Universal Music Group ont annoncé un accord permettant aux abonnés Premium de créer des reprises et remix IA à partir d'artistes participants, dans un cadre fondé sur le consentement, le crédit et la compensation[114]. **Le modèle de « jardin clos »** est encore plus explicite dans l'accord UMG–Udio : la plateforme annoncée pour 2026 doit permettre de personnaliser, écouter en flux (« streamer ») et partager de la musique générée par IA dans un environnement sous licence et protégé, entraîné sur des œuvres autorisées[115].

Plus largement, un marché se structure autour de catalogues d'actifs biométriques (voix, visages, corps, avatars) utilisables dans la musique, la publicité, la formation, la communication ou l'audiovisuel. ElevenLabs a par exemple lancé une place de marché de voix iconiques destinées à la publicité et aux contenus de marque, tandis que Synthesia utilise des acteurs pour produire des avatars vidéo, avec des contrats de licence sur leur image[116]. Dans l'audiovisuel américain, les négociations SAG-AFTRA ont consacré l'importance économique des répliques numériques: les doubles numériques d'acteurs ou de figurants doivent faire l'objet d'un consentement et d'une rémunération spécifiques[117]. Enfin, la publicité s'approprie aussi cette logique : des technologies de placement virtuel permettent d'insérer des marques ou produits en post-production dans des contenus déjà tournés, transformant chaque scène en inventaire publicitaire potentiellement monétisable[118].

2.2 Des usages au service des professionnels de l'information et de la création

On le comprend, les hypertrucages sont donc « fabriqués » pour de multiples usages et selon des logiques économiques bien différentes. Les hypertrucages à usage manifestement illicite (répliques de contenus dans un but frauduleux, contenus pédocriminels, usurpation d'identité, subversion des informations de provenance effectuée sciemment et dans un but de tromperie, de propagande, d'ingérence étrangère) soulèvent des préoccupations majeures ; contrairement aux représentations dominantes, les hypertrucages ne sont cependant pas exclusivement associés à des usages malveillants, nocifs ou illégaux en particulier dans le domaine de l'information et de la création.

A bas bruit, **de nombreux milieux professionnels de l'information et de la création ont en effet désormais recours aux hypertrucages. La création d'hypertrucages apparaît comme une « opportunité » en améliorant les capacités créatives ou en résolvant certaines contraintes techniques et budgétaires.** Certaines collaborations étroites entre entreprises technologiques et acteurs de contenus, en nécessitant un encadrement contractuel strict pour respecter les droits des personnes ou de leurs ayants droit peuvent donc conduire, pour certains acteurs, à des exploitations positives des hypertrucages, pour des raisons essentiellement économiques, artistiques ou éditoriales.

Sur le plan économique, le doublage et la localisation vidéo constituent l'un des secteurs où les économies liées à l'IA générative sont les plus directement observables, même si les estimations de marché restent très hétérogènes. The Business Research Company évalue le marché mondial des outils de doublage IA à 0,98 milliard de dollars en 2024 et 1,16 milliard de dollars en 2025, soit une croissance annuelle de 18,1 % ; d'autres cabinets donnent des périmètres très différents, allant de 794,3 millions de dollars en 2023 chez

---

[114] Allison Wallace, 'Spotify and Universal Music Group Announce Landmark Licensing Agreements for Fan-Made Covers and Remixes', *Spotify*, 21 May 2026, https://newsroom.spotify.com/2026-05-21/universal-music-group-spotify-licensing-agreements-fan-made-covers-remixes/.

[115] Mallory Hinkle, 'UNIVERSAL MUSIC GROUP AND UDIO ANNOUNCE UDIO'S FIRST STRATEGIC AGREEMENTS FOR NEW LICENSED AI MUSIC CREATION PLATFORM', *UMG*, 30 October 2025, https://www.universalmusic.com/universal-music-group-and-udio-announce-udios-first-strategic-agreements-for-new-licensed-ai-music-creation-platform/.

[116] Dan Milmo and Dan Milmo Global technology editor, 'AI Avatar Generator Synthesia Does Video Footage Deal with Shutterstock', Technology, *The Guardian*, 10 April 2025, https://www.theguardian.com/technology/2025/apr/10/ai-avatar-generator-synthesia-video-footage-shutterstock.

[117] 'Generative AI in Movies and TV: How the 2023 SAG-AFTRA and WGA Contracts Address Generative AI | Perkins Coie', accessed 5 June 2026, https://perkinscoie.com/insights/blog/generative-ai-movies-and-tv-how-2023-sag-aftra-and-wga-contracts-address-generative?utm_source=chatgpt.com.

[118] 'Advertisers Can Buy Product Placement Ads through The Trade Desk', 17 June 2025, https://www.thetradedesk.com/press-room/ttd-unveils-ai-tool-rembrand.

Market.us à 1,97 milliard de dollars en 2025 chez DataInsightsMarket[119]. Ces écarts suggèrent qu'il faut éviter de présenter un chiffre unique comme stabilisé. L'effet économique principal tient plutôt à la baisse des coûts unitaires et des délais : selon Speeek, un film standard de 90 minutes coûterait 4 500 à 27 000 dollars à doubler selon les méthodes traditionnelles, contre 45 à 900 dollars avec des outils IA, grâce à l'automatisation de la transcription, de la traduction, de la synthèse vocale et parfois de la synchronisation labiale[120]. Les plateformes de localisation vidéo comme HeyGen annoncent désormais la traduction de vidéos en 175+ langues et dialectes, avec clonage ou préservation de la voix, synchronisation labiale et conservation des expressions du locuteur ; ces promesses doivent toutefois être distinguées de la qualité effectivement obtenue selon les langues, les accents, le type de contenu et le degré de révision humaine[121].

Dans les usages professionnels, les premiers déploiements concernent surtout les contenus à fort besoin de localisation (formation interne, *marketing*, communication professionnelle, apprentissage en ligne (*e-learning*) et contenus courts pour plateformes numériques). Coca-Cola illustre cette logique d'industrialisation multilingue : sa campagne Santa, développée avec Azure AI Speech, a été lancée en 26 langues dans 43 marchés en 60 jours, mais l'entreprise ne communique pas sur des économies budgétaires précises[122]. Dans les médias d'information et l'audiovisuel, l'IA de doublage avance aussi mais sous contrôle éditorial. Reuters a lancé en 2025 des vidéos comportant des voix IA en espagnol et portugais, en collaboration avec ElevenLabs : les vidéos anglaises sont traduites et vocalisées par IA, puis contrôlées et éditées par des producteurs Reuters[123]. Cette hybridation se retrouve dans le cinéma : *The Brutalist* a utilisé la technologie Respeecher pour affiner la prononciation hongroise de certains dialogues, sans remplacer les performances des acteurs selon le réalisateur[124]. Le doublage IA doit donc être présenté moins comme un remplacement intégral déjà généralisé que comme une mutation rapide des flux de travaux (« *workflows* ») de localisation, avec des gains de coûts et de délais très importants pour les contenus standardisés, mais avec des limites artistiques, juridiques et sociales pour les œuvres de fiction, les voix d'acteurs et les contenus sensibles (cf infra).

Tilly Norwood, personnage virtuel généré par IA, alimente le fantasme à terme d'une vie d'artiste totalement inventée, ici avec sa filmographie, sa vie personnelle agitée ou des interactions étroites avec ses fans. Mais dans les films et autres œuvres de fiction audiovisuelle, l'accélération technologique multiplie déjà les cas de recours aux hypertrucages, tant dans la production que dans la diffusion. Ils s'étendent progressivement à d'autres genres : magazines, émissions de divertissement. Si les décors de *Metropolis* (1927) nécessitaient des construction colossales, l'IA permet aujourd'hui de générer des environnements d'ampleur comparable sans coûts disproportionnés. Les hypertrucages permettent de créer ou d'améliorer des effets spéciaux réalistes par exemple pour représenter des cascades spectaculaires ou générer des figurants intégrés de façon crédible à des décors complexes.

Sur le plan éditorial, les hypertrucages servent, par exemple, à préserver l'anonymat d'un témoin en modifiant son visage ou sa voix[125]. Dans des documentaires ou reportages sensibles, les hypertrucages peuvent être utilisés à la place du simple floutage pour modifier l'apparence et la voix de témoins. Cette technique offre un anonymat plus efficace tout en empêchant les techniques de « défloutage » et en

---

[119] The Business Research Company, *Artificial Intelligence (AI) Dubbing Tools Global Market Report 2025* (The Business Research Company, 2025), https://www.researchandmarkets.com/reports/6075298/artificial-intelligence-ai-dubbing-tools; 'AI Dubbing Tools Market', *Market.Us*, n.d., accessed 5 June 2026, https://market.us/report/ai-dubbing-tools-market/; DataInsightsMarket, *AI Dubbing Tools Market's Evolution: Key Growth Drivers 2026–2034*.

[120] Alexey Kulyasov, 'AI Dubbing 2025: How Technology Is Transforming the Video Localization Market', Speeek.Io Blog, 31 July 2025, https://speeek.io/en/blog/ai-dubbing-2025.

[121] 'HeyGen AI Video Translator : Translate Videos into 175+ Languages', HeyGen, accessed 5 June 2026, https://www.heygen.com/translate.

[122] Microsoft Customer Stories, '60 Days to Launch: Coca-Cola Reaches Millions with Immersive Campaign Built on Azure', Microsoft Customer Stories, 15 April 2025, https://www.microsoft.com/en/customers/story/22668-coca-cola-company-azure-ai-and-machine-learning.

[123] Reuters Communications, 'Reuters Launches AI-Voiced Packaged Video Content in Spanish and Portuguese', Media Center, *Reuters*, 15 July 2025, https://www.reuters.com/media-center/reuters-launches-ai-voiced-packaged-video-content-in-spanish-and-portuguese/.

[124] Andrew Pulver, 'The Brutalist and Emilia Perez's Voice-Cloning Controversies Make AI the New Awards Season Battleground', Film, *The Guardian*, 20 January 2025, https://www.theguardian.com/film/2025/jan/20/the-brutalist-and-emilia-perezs-voice-cloning-controversies-make-ai-the-new-awards-season-battleground.

[125] Remo Gramigna, 'Preserving Anonymity: Deep-Fake as an Identity-Protection Device and as a Digital Camouflage', *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 37, no. 3 (2024): 729–51, https://doi.org/10.1007/s11196-023-10079-y.

préservant l'authenticité émotionnelle et narrative des témoignages. Certaines techniques d'altération visuelle ou vocale peuvent donc être utilisées à des fins de protection plutôt que de manipulation. Elles permettent par exemple non seulement d'anonymiser des témoins dans des reportages, mais aussi de protéger l'identité de personnes vulnérables ou de sécuriser des contenus sensibles. Dans ce cadre, les technologies associées aux hypertrucages deviennent des outils de protection et de responsabilité éditoriale, contribuant à renforcer le journalisme d'investigation et à garantir la sécurité des sources.

Les hypertrucage*s* permettent aussi de gérer des impossibilités : recréer un animal dans une scène (évitant les interdictions de la loi sur la protection animale), reconstituer une séquence sans témoin pour illustrer des faits divers, ou encore réinsérer un comédien absent lors d'un tournage interrompu pour maladie. Ils rajeunissent des acteurs, les ressuscitent, ou prolongent un plan trop court en « inventant » des images manquantes. Dans certains cas, un acteur absent ou décédé peut être remplacé à l'écran par un avatar numérique, qu'il s'agisse d'un personnage totalement virtuel ou d'un clone numérique de l'acteur original. De même, à la télévision ou à la radio, des animateurs peuvent être remplacés par des clones vocaux ou visuels, permettant de maintenir une continuité dans la diffusion d'émissions. L'entreprise Respeecher a collaboré avec des équipes de production pour recréer la voix synthétique de Michael York afin de réaliser un film d'animation, l'état actuel de sa voix étant devenu trop différent de la version originale[126]. Des entreprises de VFX envisagent également d'utiliser la technologie pour recréer des icônes historiques comme Nelson Mandela à des fins documentaires ou éducatives[127].

Les hypertrucages offrent l'occasion de reconstitutions immersives de scènes. Dans les documentaires, les auteurs mobilisent les hypertrucages pour reconstituer un évènement historique non filmé, illustrer un témoignage par l'image et le son, ou restaurer la voix d'un personnage historique et/ou décédé. Dans les médias et l'audiovisuel, ces technologies peuvent également enrichir les formes de narration, en permettant de reconstituer des contextes historiques, d'animer des archives ou d'insérer un journaliste dans un décor reconstitué afin de renforcer la dimension pédagogique et immersive d'un reportage. Les documentaires peuvent intégrer des reconstitutions de lieux, d'événements ou de situations historiques à partir d'images et de vidéos manipulées par IA. Cela permet d'illustrer des faits pour lesquels il n'existe pas de captations originales, tout en facilitant la compréhension par le public. Le cas de *Welcome to Chechnya* a montré qu'une technologie proche de l'hypertrucage pouvait servir à protéger l'identité de témoins vulnérables tout en conservant leur expressivité ; à l'inverse, les recommandations de l'Archival Producers Alliance soulignent le risque que des images ou archives synthétiques brouillent la confiance du public dans le documentaire si leur statut n'est pas explicité[128]. Un autre exemple notable est celui du programme pionnier de Thierry Ardisson, *Hôtel du temps*[129], dans lequel la société Mac Guff a travaillé en partenariat avec 3e Œil Productions pour recréer de manière réaliste l'apparence et la voix de personnalités disparues comme Dalida, Coluche ou Jean Gabin diffusés entre 2022 et 2024. Le programme illustre un usage culturel et patrimonial de l'hypertrucage, mais aussi les questions éthiques liées au consentement *post mortem*, à la fidélité documentaire et au statut des images recréées.

L'intelligence artificielle constitue de plus un outil majeur pour la restauration et la valorisation des archives audiovisuelles. Elle permet d'améliorer la qualité visuelle et sonore de contenus anciens sans en modifier substantiellement la nature : super-résolution d'images, stabilisation vidéo, suppression d'artefacts techniques, débruitage, colorisation assistée. Ces techniques permettent d'adapter des archives filmées en définition standard aux exigences actuelles de diffusion en haute définition, tout en préservant leur valeur historique. La colorisation d'archives audiovisuelles constitue un autre exemple. Bien qu'elle soit documentée et crédible, rien ne garantit que le résultat reproduise fidèlement les couleurs originales tout en en donnant le sentiment ; la colorisation s'apparente en cela aux hypertrucages. Très chronophage et coûteuse par les méthodes traditionnelles, la colorisation est en train de devenir plus accessible par l'utilisation d'outils IA dédiés. Si leur maturité n'est pas encore complètement satisfaisante, ces outils offrent

[126] 'How Respeecher Restored Michael York's Voice for an Inspirational Healthcare Initiative', accessed 5 June 2026, https://www.respeecher.com/case-studies/respeecher-gives-voice-michael-york-healthcare-initiative.
[127] Chad, 'The Nelson Mandela Foundation Authorizes a Definitive Documentary Series on Nelson Mandela's Life. Nelson Mandela's Own Story in His Own Words, Narrated in His Voice.', *Joburgstyle Online*, 20 May 2024, http://www.joburgstyle.co.za/the-nelson-mandela-foundation-authorizes-a-definitive-documentary-series-on-nelson-mandelas-life-nelson-mandelas-own-story-in-his-own-words-narrated-in-his-voice/.
[128] 'David France | Identity Protection with Deepfakes', *MIT Open Documentary Lab*, n.d., accessed 5 June 2026, https://opendoclab.mit.edu/presents/david-france-identity-protection-deepfakes/.
[129] 'Hôtel du Temps | FranceTvPro.fr', accessed 5 June 2026, https://www.francetvpro.fr/contenu-de-presse/16034879.

des opportunités intéressantes pour valoriser des archives en noir et blanc, la colorisation devenant souvent un préalable pour toucher un jeune public.

L'IA générative favorise ainsi l'accessibilité et la diffusion des œuvres culturelles auprès de publics élargis : traduction vocale d'interprétations existantes, génération de doublages dans différentes langues, production de versions adaptées à des publics spécifiques. Ces usages contribuent à la circulation internationale des œuvres et à leur valorisation économique. Ils peuvent également soutenir l'inclusion, en facilitant l'accès aux contenus pour des personnes en situation de handicap ou pour des publics non francophones. Des médias mobilisent ces techniques pour immerger un journaliste dans un décor d'époque ou illustrer des travaux historiques, scientifiques ou culturels. Dans ce cas de figure, l'hypertrucage ne correspond pas à une manipulation trompeuse, mais à un outil de médiation culturelle permettant de rapprocher le public d'un passé ou d'une mise en contexte difficile à illustrer. Dans l'édition, l'usage de l'IA pour restaurer ou améliorer des documents anciens (photographies, parutions) peuvent relever de l'hypertrucage. Ces techniques ouvrent de nouvelles possibilités d'accessibilité, notamment pour les personnes souffrant de troubles de la parole ou de la lecture.

Sur le plan plus directement créatif, des usages parodiques ou pédagogiques se multiplient. Sur les réseaux sociaux, des personnalités publiques apparaissent dans des situations absurdes. Cette dimension divertissante est confirmée par d'autres travaux[130] : l'étude menée sur X montre que, malgré une croissance rapide et durable, la majorité des contenus synthétiques relève de l'humour, de la satire et de l'expérimentation créative. Selon leur analyse, les usages politiques – tweets se référant explicitement à des acteurs ou enjeux politiques – sont minoritaires (~40%). De plus, les contenus les plus viraux sont humoristiques et non politiques (64%). Une autre étude montre que les hypertrucages sur TikTok et YouTube sont avant tout conçus pour stimuler une forme d'imagination créative et non pour tromper[131]. Ce constat doit cependant être nuancé. Avec la mécanique de l'amplification algorithmique, les contenus activement mis en avant par les plateformes obéissent à une autre logique. Les algorithmes de recommandation étant conçus pour maximiser l'engagement, les plateformes multiplient les contenus extrêmes voire dégradants pour attirer l'attention des utilisateurs au détriment de contenus purement divertissants pourtant très largement présents[132] (cf. *supra*, la dichotomie valeur/volume).

Dans la musique, plus d'un tiers des 1500 producteurs professionnels interrogés utilisent déjà des outils d'IA dans leur flux de travail : 36,8 % des producteurs interrogés déclarent utiliser l'IA dans leur flux de travaux (« *workflow* »)[133], tandis que 51 % des créateurs de moins de 35 ans déclarent avoir utilisé l'IA selon un rapport GEMA/SACEM relayé par Complete Music Update[134]. Les maisons de disques et les artistes explorent ensemble de nouveaux espaces de création permettant de faire « *revivre des chanteurs disparus, d'automatiser la création de podcasts ou d'histoires audio avec des outils text-to-speech, de créer des morceaux inédits…* ». Un exemple emblématique est la chanson *Now and Then* des Beatles (2023), pour laquelle des techniques d'intelligence artificielle ont été utilisées afin d'isoler et de restaurer la voix de John Lennon à partir d'une démo ancienne, permettant d'achever le morceau avec les contributions des autres membres du groupe. De manière similaire, certaines maisons de disques expérimentent des adaptations vocales, des traductions ou des reconstitutions artistiques avec l'accord des artistes concernés. Plusieurs majors (Universal, Sony, Warner) expérimentent la traduction vocale contrôlée (par exemple, faire chanter un artiste dans une autre langue en conservant son timbre), la restauration de voix historiques et des collaborations assistées par IA lorsque les contrats et le consentement sont explicitement définis. Des spectacles comme ABBA Voyage, qui met en scène des avatars numériques photoréalistes du groupe dans une salle dédiée à Londres, prolongent les expériences de « résurrection » d'artistes. De nouveaux formats permettent de monétiser l'image et la voix d'artistes à une échelle globale, y compris pour des artistes

---

[130] Giulio Corsi et al., 'The Spread of Synthetic Media on X', *Harvard Kennedy School Misinformation Review* 5, no. 3 (2024): 1–19.

[131] Carl Öhman, 'Who Are the Targets of Deepfakes? Evidence From Flagged Videos on TikTok and YouTube', *Journal of Digital Social Research* 7, no. 2 (2025): 18–33, https://doi.org/10.33621/jdsr.v7i255529.

[132] Kaitlyn Regehr et al., 'Normalizing Toxicity: The Role of Recommender Algorithms for Young People's Mental Health and Social Wellbeing', *Frontiers in Psychology* 16 (November 2025), https://doi.org/10.3389/fpsyg.2025.1523649.

[133] Rare Connections, 'AI Music Statistics and Facts: How Producers Use AI', 2025, https://www.rareconnections.io/ai-music-statistics.

[134] Chris Cooke, '71% of Music Creators Fear Multi-Billion Dollar Music AI Business Could Stop Songwriters from Earning a Living, Says New Report', Complete Music Update, 31 January 2024, https://completemusicupdate.com/71-percent-of-music-creators-fear-multi-billion-dollar-music-ai-business-could-stop-songwriters-from-earning-a-living-says-new-report/; Rare Connections, 'AI Music Statistics and Facts: How Producers Use AI'.

décédés ou retirés de la scène, tout en soulevant des questions juridiques complexes (voir chapitre 3). De plus, des « artistes synthétiques » entièrement virtuels expérimentent de nouveaux formats de performance dans les métavers.

Parfois, les éléments artificiels sont intégrés de manière invisible, comme de classiques effets spéciaux ; parfois au contraire, la production synthétique est mise en avant. L'usage de l'IA comme outil créatif devient alors un élément constitutif de l'œuvre, voire un argument esthétique et commercial. C'est l'approche revendiquée par le *World Artificial Intelligence Film Festival*, dédié aux films réalisés à partir de technique d'IAG. Le festival présente une sélection de courts métrages entièrement ou partiellement réalisés à l'aide de modèles de génération d'images, de voix ou de vidéos afin de montrer que l'IA peut être un vecteur de créativité et soutenir une nouvelle génération d'artistes.

Quels que soient les secteurs, les technologies de génération et de transformation des contenus ouvrent la voie à de nouveaux formats artistiques et culturels : œuvres hybrides, performances virtuelles, collaborations entre artistes et intelligences artificielles, formats interactifs. Ces usages participent à l'émergence de nouvelles pratiques culturelles et à la transformation des industries créatives. L'hypertrucage peut donc devenir un outil d'expérimentation artistique et un levier d'innovation, à condition d'être utilisé dans un cadre éthique, transparent et respectueux des droits des personnes.

Pour les professionnels de la création et de l'information, dans tous les secteurs, **le recours aux hypertrucages est donc désormais une réalité.** Cette situation n'est pas problématique en soi puisque, comme nous venons de le voir, lorsqu'il est consenti et encadré, **le recours à l'hypertrucage peut servir des objectifs légitimes et stimuler la créativité à l'instar de ce que les trucages ont toujours fait dans l'histoire des industries culturelles**. **Certains enjeux liés au recours à de tels outils suscitent cependant des craintes** [135] qui se sont accentuées avec la généralisation de l'accès du grand public à des outils techniques. **Les hypertrucages deviennent en effet problématiques dès lors qu'ils atteignent la réputation et les revenus individuels de personnes physiques et morales ou se substituent globalement, dans certaines filières, à des emplois, fragilisant ainsi les modèles économiques des industries culturelles.**

2.3 Fragilisation des industries culturelles

Face à la multiplication des hypertrucages, les effets réputationnels dont sont victimes des personnes physiques ou morales, voire des Etats, s'accompagnent bien souvent de conséquences économiques. Au-delà des atteintes à leurs activités individuelles, les hypertrucages entraînent par ailleurs, au niveau des filières culturelles dans leur ensemble, des pertes de revenus et des effets d'éviction.

2.3.1. Réputation individuelle et conséquences économiques

Les relations entre fournisseurs d'IA et professionnels de la création sont marquées par des tensions, en particulier autour de l'usage non autorisé de la voix ou de l'apparence. Les hypertrucages peuvent également imiter le style reconnaissable d'un artiste, ou générer de fausses œuvres attribuées à un créateur connu. Ces pratiques soulèvent des questions en lien avec la contrefaçon, bien que les œuvres générées ne soient pas des copies exactes mais des créations nouvelles inspirées du style original. Les hypertrucages, lorsqu'ils représentent des personnes réelles ou identifiables, soulèvent également des enjeux majeurs en matière de droits de la personnalité, notamment le droit à l'image ou le respect de la vie privée (cf. chap.3).

Ces défis sont exacerbés par la capacité des modèles d'IA à générer des contenus réalistes mais purement synthétiques, susceptibles d'être perçus comme représentant des artistes, auteurs ou célébrités, sans leur consentement. Toute utilisation non autorisée des attributs de la personnalité d'un artiste-interprète pour créer du contenu généré par IA a ainsi des conséquences sur sa réputation, en captant l'attention des communautés de fans, en perturbant des campagnes promotionnelles pourtant planifiées, et en générant des revenus fondés sur sa notoriété, mais à son insu. Cette utilisation non autorisée des attributs de la personnalité cible souvent des artistes qui sont confirmés sans être nécessairement des superstars, via la mise à disposition de modèles de clonage destinés à produire des répliques numériques de leurs performances.

[135] Robert Chesney and Danielle Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *California Law Review* 107, no. 6 (2019): 1753.

Une affaire fortement médiatisée a opposé Scarlett Johansson à l'entreprise OpenAI, suspectée d'avoir intégré dans l'un de ses produits une voix synthétique nommée Sky, ressemblant fortement à celle de l'actrice Scarlett Johansson – qui avait justement prêté sa voix à celle d'une intelligence artificielle dans le film *Her*. L'actrice a condamné cette ressemblance vocale et dénoncé un usage non consenti[136]. Bien qu'OpenAI ait nié toute imitation volontaire, l'entreprise a suspendu l'utilisation de cette voix après la controverse[137]. Ce cas a agi comme un catalyseur dans le débat public sur les clones vocaux et les hypertruages d'artistes, mettant en lumière les enjeux de consentement et de réputation. Des organisations comme Les Voix ou UVA (United Voice Artists) s'opposent ainsi fermement au développement de technologies de clonage vocal ou de doublage automatisé par IA lorsqu'elles sont déployées sans accord préalable des artistes. Dans le domaine musical, la plupart des artistes s'opposent par défaut à l'usage de leur voix pour réaliser des chansons par hypertrucage. Mais de rares exceptions existent, comme la chanteuse électro canadienne Grimes (cf. *supra*) qui a annoncé publiquement permettre à quiconque d'utiliser sa voix pour des créations musicales générées avec IA, à condition de partager les bénéfices à parts égales.

Des hypertrucages véhiculent également de fausses informations attribuées à des groupes de médias, à leurs dirigeants, ou aux personnalités incarnant ces entreprises (en particulier les présentateurs des journaux télévisés, journalistes/chroniqueurs). Des « arnaques aux célébrités » se multiplient : faire dire à quelqu'un des choses qu'il n'a jamais dites, détourner une vidéo, produire des contenus diffamatoires ou incitant à la haine, etc. Des journalistes sont ciblés dans le cadre d'ingérences étrangères et de campagnes de désinformation. Une entreprise peut également mobiliser un hypertrucage pour diffuser des informations mensongères sur un produit concurrent afin d'en altérer la réputation.

Dans le domaine de la publicité, l'usage d'hypertrucages au sein de clips promotionnels a tantôt pu être autorisé (par exemple par Bruce Willis, mais sans pour autant céder durablement le droit d'exploiter son clone numérique) ou tantôt dénoncé (par exemple par Tom Hanks concernant des publicités pour une mutuelle dentaire réalisée sans son consentement). L'existence de contenus sur lesquels la publicité affichée porte préjudice à l'image de la marque (*brand safety*) a des impacts sur les effets de la publicité pour la marque et sur les revenus publicitaires dans leur ensemble. A contrario, lorsqu'une entreprise reconnaît l'usage d'hypertrucages, son image et sa réputation peuvent en sortir renforcés[138].

Des contenus de plus en plus réalistes, créent donc de la confusion dans l'esprit du public autour de l'image officielle d'un artiste, d'un groupe, d'une marque en laissant croire à l'authenticité et à l'attribution réelle des prestations, ce qui porte directement atteinte à sa réputation avec les conséquences économiques associées.

2.3.2. Des pertes de revenus pour la filière

La croissance constatée des marchés (cf. *supra*) ne s'effectue pas toujours au profit de l'amont de la filière culturelle, c'est-à-dire des créateurs. Il n'existe à ce jour aucune estimation consolidée des pertes économiques liées spécifiquement aux usages des hypertrucages dans les industries culturelles ; les données disponibles mêlent souvent fraude générale, désinformation et usages spécifiquement culturels. Des études fournissent cependant quelques estimations au niveau sectoriel, pour l'ensemble des effets de l'IAG et non pour les seuls hypertrucages.

L'étude Goldmedia, commandée conjointement par la SACEM et la GEMA auprès de 15 073 créateurs et éditeurs en France et en Allemagne, estime un déficit potentiel de 27 % dans les revenus des créateurs musicaux d'ici 2028, en l'absence d'un système de rémunération pour les apports humains utilisés dans

[136] Dawn Chmielewski et al., 'Scarlett Johansson Says OpenAI Chatbot Voice "eerily Similar" to Hers', Technology, *Reuters*, 21 May 2024, https://www.reuters.com/technology/scarlett-johansson-says-openai-chatbot-voice-eerily-similar-hers-2024-05-21/.
[137] 'How the Voices for ChatGPT Were Chosen', OpenAI, accessed 5 June 2026, https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/.
[138] Greig Powers et al., 'To Tell or Not to Tell: The Effects of Disclosing Deepfake Video on US and Indian Consumers' Purchase Intention', *Journal of Interactive Advertising* 23, no. 4 (2023): 339–55, https://doi.org/10.1080/15252019.2023.2260399.

l'entraînement des modèles[139]. Pour la France et l'Allemagne seules, cela représenterait un écart de 950 millions d'euros en droits d'auteur en 2028, et un total cumulé de 2,7 milliards d'euros de pertes d'ici cette date[140]. Parmi les créateurs interrogés, 71 % estiment que l'IA les privera de revenus et menacera leur avenir, tandis que 64 % considèrent que les risques dépassent les opportunités[141] (Confédération Internationale des Sociétés d'Auteurs et Compositeurs). Le ratio est donc paradoxal : le marché IA-musique croît de manière importante (×10 en 5 ans) mais les créateurs humains risquent de ne pas en bénéficier.

L'étude économique internationale de la CISAC élargit ce constat à l'échelle mondiale. Elle estime que 24 % des revenus des créateurs musicaux et 21 % de ceux du secteur audiovisuel pourraient être menacés d'ici 2028 en raison de l'essor des contenus générés par IA[142] . Cela correspond à une perte cumulée d'environ 22 milliards d'euros sur cinq ans (2023–2028) pour les industries musicales et audiovisuelles, dont environ 10 milliards d'euros pour la musique seule. Les projections indiquent que les contenus musicaux et audiovisuels générés par IA pourraient représenter un marché de 64 milliards d'euros d'ici 2028, contre environ 3 milliards aujourd'hui, pouvant couvrir jusqu'à 20 % des revenus des plateformes de streaming et 60 % des bibliothèques musicales.

Les coûts indirects sont également importants. Les procès se multiplient aboutissant parfois, in fine, à des accords commerciaux. Les majors Universal Music Group, Warner Music Group et Sony Music Entertainment ont engagé des poursuites contre Suno et Udio, deux startups de génération musicale par IA, pour violation de droits d'auteur à grande échelle. Outre l'utilisation des œuvres pour l'entraînement des modèles en amont, des litiges portent également sur le clonage vocal non autorisé (cas Drake, The Weeknd), sur les avatars d'artistes et sur l'utilisation de voix de célébrités sans consentement (cas Scarlett Johansson c. OpenAI). Ces contentieux représentent des coûts et créent une incertitude réglementaire qui freinent les investissements. Les grèves de Hollywood en 2023 ont également eu des répercussions économiques. Les grèves hollywoodiennes d'une durée de 118 jours, ont eu un impact sur 45 000 emplois et coûté environ 6,5 milliards de dollars à la Californie du Sud[143]. L'IA et la duplication numérique étaient au cœur des revendications et des concessions obtenues, avec plus d'1 milliard de dollars en nouvelles compensations[144].

Certes, de nouveaux métiers émergent. Des ingénieurs en apprentissage automatique (« *machine learning* ») sont chargés de la modélisation des visages, des voix ou des corps ; des animateurs 3D et superviseurs d'effets visuels (VFX) sont responsables de l'intégration réaliste des contenus synthétiques dans les séquences finales. A côté de ces profils qui renouvellent le travail de la post-production, apparaissent de nouveaux métiers hybrides à la frontière entre création, ingénierie et direction artistique. L'entraîneur d'avatars (« *avatar trainer* ») ajuste les modèles de visages synthétiques en fonction des attentes narratives et esthétiques ; l'entraîneur de voix synthétique (« *synthetic vocal coach* ») supervise la restitution vocale et son adéquation émotionnelle ; le concepteur de pipeline IA (« *AI pipeline designer* ») conçoit des chaînes de production compatibles avec l'IA pour une intégration fluide…. Ces fonctions encore marginales il y a peu de temps, deviennent centrales dans les productions audiovisuelles à gros budget et se diffusent progressivement vers les productions intermédiaires.

D'autres métiers se sentent d'ores et déjà globalement menacés : traducteur, photojournalistes, illustrateurs, figurants, porte d'entrée vers le métier de comédien, designers ou graphistes voient leurs

---

[139] Goldmedia, *AI and Music: A Study on the Impact of Artificial Intelligence on the Music Sector* (Goldmedia, 2024), https://www.goldmedia.com/fileadmin/goldmedia/Studie/2023/GEMA-SACEM_AI-and-Music/AI_and_Music_GEMA_SACEM_Goldmedia.pdf.

[140] Liz Pelly, 'As AI-Made Music Explodes, Deezer Lays out Strategy to Identify AI Tracks and Weed out Illegal and Fraudulent Content on Its Platform', Music Business Worldwide, 6 June 2023, https://www.musicbusinessworldwide.com/as-ai-made-music-explodes-deezer-lays-out-strategy-to-identify-ai-tracks-and-weed-out-illegal-and-fraudulent-content-on-its-platform/.

[141] CISAC, 'Global Economic Study Shows Human Creators' Future at Risk from Generative AI', CISAC Newsroom, 4 December 2024, https://www.cisac.org/Newsroom/news-releases/global-economic-study-shows-human-creators-future-risk-generative-ai.

[142] CISAC, 'Global Economic Study Shows Human Creators' Future at Risk from Generative AI'; GEMA, 'Study: AI and Music', GEMA News, 30 January 2024, https://www.gema.de/en/news/ai-study.

[143] Dominic Patten D'Alessandro Anthony, 'The Strike Is Over! SAG-AFTRA & Studios Reach Tentative Deal On New Three-Year Contract', *Deadline*, 9 November 2023, https://deadline.com/2023/11/sag-strike-ends-actors-studios-deal-contract-1235566470/.

[144] 'Tentative Agreement Reached', SAG-AFTRA, 8 November 2023, https://www.sagaftra.org/tentative-agreement-reached.

pratiques actuelles remises en cause. En ce qui concerne les métiers de l'écriture, comme le journalisme, une étude estime que les métiers de reporters et journalistes sont parmi les plus exposés aux systèmes d'IA[145]. De même, pour les artistes-interprètes, des contenus reprenant tout ou une partie des attributs de leur personnalité et de leur talent (apparence, voix, attitude, patte artistique, etc.) peuvent être perçus, à tort, comme authentiques.

La généralisation de l'IA générative pour le doublage audiovisuel (séries télévisées, films ou encore jeux vidéo) est au cœur de discussions entre les syndicats, les commanditaires et les studios pour poser le cadre des conditions du recours à l'IA et garantir, dans les contrats d'artistes, la preuve du consentement. La crainte exprimée est que certains commanditaires de doublage (studios de cinéma, plateformes et diffuseurs) substituent massivement l'IA à la voix humaine dans les prochaines années. Cette question du doublage est emblématique ; le déploiement de l'IAG est source d'économies déjà observables (cf supra 2.2) mais une mutation de l'activité à moyen terme pourrait fragiliser certains studios français au profit notamment des États-Unis comme le souligne une note sur le métier des comédiens de doublage à l'heure de l'IA, émanant de l'observatoire des métiers lancé par Audiens, l'Afdas et le Centre national du cinéma et de l'image animée.

2.3.3. Des effets d'éviction

Pour certaines professions, les effets d'éviction s'exercent d'abord par une concurrence par les prix, dans la mesure où l'IA permet, on l'a vu, de produire plus vite à moindre coût que des humains. Les contenus synthétiques (images, sons archives …), moins coûteux et plus rapides à produire, dévalorisent le travail des créateurs humains, entraînent une baisse des tarifs des prestations créatives et, à terme, menacent certains métiers. Pour réduire les coûts, un producteur peut décider de solliciter les outils d'IA pour générer un morceau de harpe joué « à la manière » d'une musicienne, ou avec une voix « imitant » celle d'une chanteuse, plutôt que d'engager et rémunérer ces dernières. Les studios de livres audios utilisent de plus en plus de catalogues de voix IA, qui sont beaucoup moins coûteux qu'un narrateur humain. Les entreprises comme Suno, Udio ou encore Mubert, permettent de créer instantanément des morceaux libres de droits, remplaçant de fait certains compositeurs par des générateurs IA. Cette dynamique ne se limite plus à des startups indépendantes ; de très grands groupes déploient désormais leurs propres outils de création générative (à l'image de Google avec Gemini) tandis que d'autres, au contraire, abandonnent ces segments de marché jugés peu rentables (cf. l'outil Sora de génération de vidéos abandonné par OpenAI en mars 2026). Cette situation soulève directement la question de la viabilité d'une spécialisation dans la génération de contenus et renvoie à la frilosité relative des investisseurs dans ce segment de l'IA (cf. *supra*).

Les effets d'éviction s'exercent également par les quantités. La surabondance de contenus générés par les systèmes d'IA risque en effet de saturer le marché et, par conséquent, de réduire la découvrabilité des œuvres humaines. C'est notamment le cas de la prolifération de clones vocaux non autorisés, en particulier sur YouTube et TikTok, qui détourne l'attention des enregistrement officiels. Le risque est que la production de contenus via l'hypertrucage domine le marché de l'information et de la création en volume, « pollue » les offres en ligne et se banalise chez les consommateurs qui ne sont pas avertis et ont du mal à distinguer les contenus humains (moins nombreux) des contenus synthétiques (voir chap. 4). Cette pratique détourne l'attention du public et par conséquent une partie des revenus qui auraient pu bénéficier aux créateurs humains[146]. Les estimations de la proportion frauduleuse d'écoutes vont de 1 à 3 % [147] à 7 % (Deezer, 2022)[148]. Les revenus mondiaux de la musique enregistrée ont atteint 29,6 milliards de dollars en 2024, dont le streaming représente plus de 20 milliards de dollars[149]. Même avec l'estimation basse de 1 à 3 % de

[145] Tyna Eloundou et al., 'GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models', arXiv:2303.10130, preprint, arXiv, 21 August 2023, http://arxiv.org/abs/2303.10130.
[146] Harnoorvir Singh Josan, *AI and Deepfake Voice Cloning: Innovation, Copyright and Artists' Rights*, Digital Policy Hub Working Paper (Centre for International Governance Innovation, 2024), https://www.cigionline.org/publications/ai-and-deepfake-voice-cloning-innovation-copyright-and-artists-rights/.
[147] Centre national de la musique, *Stream Manipulation: 2021* (Centre national de la musique, 2023), https://cnm.fr/wp-content/uploads/2023/01/2023_-CNM-_Manipulation-des-streams_ENG.pdf.
[148] Identity Music, 'Artificial Streaming: Fake Streams, Fraud & Fines', 10 February 2026, https://identitymusic.com/blog/artificial-streaming-fake-streams-fraud-fines.
[149] World Intellectual Property Organization, 'IFPI Looks at a Decade of Digital Transformation in the Music Industry', *WIPO Magazine*, 23 April 2025, https://www.wipo.int/en/web/wipo-magazine/articles/ifpi-looks-at-a-decade-of-digital-transformation-in-the-music-industry-73661.

fraude, cela représente 200 à 600 millions de dollars détournés annuellement. Avec l'estimation haute de 7 %, ce chiffre dépasserait 1,4 milliard de dollars par an.

Ainsi, des **artistes « fantômes » déferlent sur les plateformes de streaming musical,** avec des enregistrements entièrement générés par IA sous des noms fictifs, en offrant un semblant d'authenticité et d'identité réelle (génération par IA de pochettes d'albums, de bio fictives, de photos crédibles de profils...). Les revenus issus d'artistes fantômes sur les plateformes musicales (Deezer, Spotify, …) et audiovisuelles (YouTube) sont devenus un phénomène préoccupant en raison de leur ampleur et de leur caractère expansionniste ; Deezer estime ainsi (le 11 septembre 2025) recevoir plus de 30 000 titres entièrement générés par IA chaque jour, soit plus de 28 % de ses livraisons quotidiennes. C'est pourquoi un détecteur de « musiques *deepfakes* », a été mis en place[150] (cf. Chap.6). La prolifération de contenus fantômes s'accompagne souvent d'un autre phénomène, celui des « fermes à clics » qui tentent de simuler des écoutes (« *streams* ») faites par des auditeurs humains.

Un cas emblématique illustre la mécanique de cette fraude. Le musicien américain Michael Smith a été inculpé en 2024 pour avoir organisé une fraude massive aux redevances ; il a ainsi utilisé l'IA afin de créer des centaines de milliers de morceaux, puis des robots (« *bots* ») (environ 10 000 comptes) pour générer environ 661 440 écoutes par jour, détournant plus de 10 millions de dollars en redevances[151]. Ce cas révèle la vulnérabilité du modèle « *market centric* » (centré sur le marché) ; les millions de dollars détournés ont été soustraits non aux plateformes – pas directement tout du moins car ces pratiques érodent la relation de confiance avec les majors et les artistes – mais au montant global destiné aux artistes humains.

Ce cas illustre également la facilité avec laquelle les outils génératifs permettent d'industrialiser la fraude musicale, de la génération de contenus « *slop* » (de faible qualité et produits en grande quantité) en passant par la création d'identités, jusqu'à la consommation artificielle par des robots capables de simuler des comportements humains pour contourner les systèmes anti fraudes (création de pauses, de playlists…). Le groupe *The Velvet Sundown* a accumulé plus d'un million d'écoutes sur Spotify en 2025 avant d'être identifié comme groupe fantôme synthétique. Cependant 80 % de la fraude détectée se situe non pas dans les tops d'écoutes mais dans une sorte de « *longue traîne artificielle* »[152]. Au lieu de propulser un seul titre vers des millions d'écoutes ce qui déclencherait des alertes, il devient plus judicieux de générer des quantités massives de titres « écoutés » quelques milliers de fois seulement ce qui globalement a un effet massif d'éviction au détriment des créateurs humains.

Plus récemment encore, des contenus générés par IA sont apparus sur les pages d'authentiques artistes, sur des plateformes de streaming ou, le plus souvent, sur YouTube et Tik Tok, en usurpant l'identité, le nom, l'image, la voix de ces artistes. Le secteur musical n'est pas le seul concerné. En 2026, l'écrivain Julien Blanc-Gras a ainsi découvert sur Amazon un ouvrage écrit par une intelligence artificielle mais disponible à la vente comme faisant partie de sa propre bibliographie.

Conclusion

Ce chapitre a montré que **la structuration du marché des hypertrucages s'organise autour de dynamiques divergentes entre marché grand public et usages professionnels et entre usages légitimes et malveillants.** L'étude *Tackling Deepfakes in European Policy* projetait en 2021 un scénario à cinq ans pour l'évolution des risques liés aux hypertrucages[153]. Sa prédiction centrale était que l'usage malveillant des hypertrucages passerait d'une situation de forte gravité et faible probabilité (à l'époque, quelques cas isolés et médiatisés comme une tentative de coup d'État au Gabon ou la pornographie non consentie ciblant des célébrités féminines) à une situation de forte gravité et probabilité modérée à élevée, marquée par la diffusion grand public des outils et la normalisation de leur usage dans les pratiques sociales

[150] Darius Afchar et al., 'Detecting Music Deepfakes Is Easy but Actually Hard', arXiv:2405.04181, preprint, arXiv, 22 May 2024, https://doi.org/10.48550/arXiv.2405.04181.
[151] Luis Prada, '"Musician" Arrested for Using AI Songs and a Bot Army to Scam Spotify for Millions in Royalties', VICE, 6 September 2024, https://www.vice.com/en/article/ai-songs-spotify-scam/; U.S. Attorney's Office, Southern District of New York, 'North Carolina Musician Charged With Music Streaming Fraud Aided By Artificial Intelligence', 4 September 2024, https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence.
[152] Centre national de la musique, *Manipulation Des Écoutes En Ligne* (Centre national de la musique, 2023), https://cnm.fr/etude-manipulation-des-ecoutes-en-ligne/.
[153] Vvan Huijstee et al., *Tackling Deepfakes in European Policy*.

courantes. Quatre ans plus tard, cette prédiction est largement validée. Sumsub[154] rapporte une multiplication par plus de dix des tentatives de fraude à l'identité utilisant des hypertrucages entre 2022 et 2024, avec une concentration sectorielle sur les services financiers, les cryptomonnaies et les plateformes d'emploi. Pindrop[155] documente une multiplication équivalente des cas de fraude vocale en environnement bancaire sur la même période. Europol[156] institutionnalise le constat d'une économie criminelle organisée autour des hypertrucages (fraude au dirigeant, *deepfake-as-a-service* criminel, *voice phishing*). Le *Government Accountability Office* américain propose un cadre fédéral équivalent[157]. L'étude STOA insistait également sur **le caractère à double usage (« *dual-use* ») de la technologie, c'est-à-dire son couplage indissociable entre usages légitimes et usages malveillants qui rend nécessaire une régulation fine.**

Les industries culturelles n'échappent pas à cette dualité entre opportunités et risques. Elles sont de plus, comme nous allons le voir, confrontées à des problématiques juridiques complémentaires du fait de leur spécificité.

[154] Sumsub, *Identity Fraud Report 2025–2026* (Sumsub, 2025), https://sumsub.com/fraud-report-2025/.
[155] Pindrop, *2024 Voice Intelligence & Security Report* (Pindrop, 2024), https://www.pindrop.com/press-release/pindrop-unveils-voice-intelligence-security-report/.
[156] Europol, *Steal, Deal and Repeat: Internet Organised Crime Threat Assessment (IOCTA) 2025* (Europol, 2025), https://www.europol.europa.eu/cms/sites/default/files/documents/Steal-deal-repeat-IOCTA_2025.pdf.
[157] U.S. Government Accountability Office, *Science & Tech Spotlight: Combating Deepfakes*, GAO-24-107292 (U.S. Government Accountability Office, 2024), https://www.gao.gov/products/gao-24-107292.

3. L'ENCADREMENT JURIDIQUE DES HYPERTRUCAGES : UN ENSEMBLE DISPARATE DE REGLES N'ASSURANT PAS UNE PROTECTION UNIFORME

En se bornant à harmoniser les règles de transparence applicables aux hypertrucages[158], le législateur européen a fait le choix de renvoyer pour l'essentiel l'encadrement de ces contenus aux autres législations européennes et nationales susceptibles de s'appliquer. L'hypertrucage ne constitue ainsi une qualification juridique pertinente qu'à cette fin et n'exclut pas que d'autres qualifications s'appliquent pour les besoins des dispositions en cause. Une telle complémentarité est indispensable car l'obligation de transparence ne saurait épuiser l'ensemble des enjeux auxquels nos sociétés sont confrontées du fait de l'apparition de ces contenus[159].

Cette logique, qui s'explique aussi par le caractère nouveau d'un type de contenu qui s'inscrit dans un ensemble plus large de contenus générés ou manipulés par l'IA, n'est pas propre à l'Union européenne. Des interrogations liées à l'application par superposition des différentes législations se posent également dans les autres Etats membres, mais également au Royaume-Uni[160] ou aux Etats-Unis[161]. La doctrine américaine a nourri – et continue à le faire – d'importantes réflexions sur la capacité à mobiliser face aux hypertrucages les règles du *copyright* et du *right of publicity*, dans un contexte où la liberté d'expression et la section 230 du *Communications Decency Act* constituent des limites importantes aux actions et évolutions juridiques envisageables. Si de nombreuses initiatives législatives ont été engagées, que ce soit au niveau des Etats ou de l'Etat fédéral[162], le débat y demeure largement ouvert. De l'ensemble de ces initiatives et débats, certains éléments sont de nature à nourrir la réflexion pour le cadre européen et plus spécifiquement français, bien qu'il faille garder à l'esprit les différences structurelles entre les textes applicables des deux côtés de l'Atlantique et les limites de leur transposition. Malgré ces différences, des éléments de ce débat américain constituent des points de repère utiles : il s'agit en particulier du cadre posé par le *US Copyright Office* pour des évolutions législatives répondant aux questions soulevées par les hypertrucages[163], des réflexions liées aux enjeux de la patrimonialisation de certains droits, notamment ceux de la personnalité[164], ou encore des discussions sur la protection de la personne par le *copyright*[165], qui ont pris corps par le dépôt, avec succès, par l'acteur américain Matthew McConaughey auprès du *US Patent and Trademark Office* de huit marques portant sur sa voix, certaines expressions emblématiques et des séquences audiovisuelles courtes, de même que par la chanteuse Taylor Swift [166].

Sans être exposés en détail, ces éléments ont nourri la réflexion de la mission, qui a par ailleurs pu échanger avec plusieurs spécialistes américaines du sujet. Il sera en revanche fait état plus en détail du projet de loi présenté par le gouvernement danois sur cette question, dans la mesure où, malgré là aussi des différences de cadre juridique, les enjeux de conformité au droit de l'Union européenne s'y retrouvent.

Après une étude du kaléidoscope que constituent les principaux textes susceptibles d'être mobilisés face à la diffusion d'hypertrucages (3.1), les possibilités d'évolution de ceux-ci seront évoquées en tenant compte des enjeux contractuels (3.2).

---

[158] Art. 1er, paragr. 2, point *d*, du RIA.

[159] Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1.

[160] Voy. par ex. John Zerilli, « Appropriation of Personality in the Era of Deepfakes », *in* Ernest Lim, Phillip Morgan (éd.), *Private Law and Artificial Intelligence*, Cambridge University Press, 2024, p. 227.

[161] Robert Chesney, Danielle K. Citron, « Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security », *California Law Review*, 2019, vol. 107, p. 1753.

[162] Marylou Le Roy, « L'encadrement des systèmes d'intelligence artificielle par les Etats-Unis – Bref panorama et convergence avec l'Europe », *Communication Commerce électronique*, n° 6, juin 2024, étude 8.

[163] *Copyright and Artificial Intelligence*, Part 1: Digital Replicas, juillet 2024.

[164] Michael Goodyear, « Dignity and Deepfakes », *Arizona State Law Journal*, 2025, vol. 57, p. 931.

[165] Jennifer Rothman, « Copyrighting People », *Journal of Copyright Society*, 2025, vol. 72, n° 1, p. 1.

[166] Jane Ginsburg, Graeme Austin, « Deepfakes in Domestic and International Perspective », *Columbia Public Law Research Paper*, n° 5127178, 2025 ; Pierre Sirinelli, Stéphane Prévost-Boyard, « For sure », *Dalloz IP/IT*, 2026, p. 53.

3.1. Les potentialités du droit positif : un ensemble de dispositions disparates couvrant certains aspects des hypertrucages

Les différents textes pouvant être mobilisés face aux hypertrucages ont déjà fait l'objet de présentations[167]. Au regard de l'objet de la présente mission, il s'agit moins d'en refaire un état exhaustif que d'évaluer la capacité des acteurs des secteurs culturel et créatif à mobiliser ces outils pour répondre de manière appropriée aux enjeux soulevés par les hypertrucages, et donc d'en apprécier les limites. Les questions posées par ces contenus pour ces acteurs doivent distinguer ce qui relève de la protection des personnes visées par l'hypertrucage, de la situation des objets, lieux, entités ou événements concernés.

Il faut néanmoins au préalable faire un constat d'ensemble : pour traiter des hypertrucages, il est nécessaire de solliciter un ensemble de plusieurs normes, ayant des finalités et des champs d'application différents[168]. Ceux-ci se recoupent en partie, mais laissent des doutes et insuffisances qu'il s'agit principalement d'identifier. Cela résulte du fait que ces textes ont été conçus dans une perspective différente de celle des hypertrucages, et, ainsi que cela a déjà été indiqué, il semble à la mission que, **s'il est nécessaire de répondre aux enjeux soulevés par ceux-ci pour les secteurs culturels et créatifs, il serait risqué de développer des dispositions qui n'auraient pour champ d'application que les hypertrucages** : les contours incertains de la notion conduiraient à laisser des zones d'incertitudes importantes et une telle approche pourrait se heurter au développement des techniques futures.

3.1.1. La protection des personnes face aux hypertrucages

Envisager la protection des personnes dont les caractéristiques sont susceptibles d'être reprises dans un hypertrucage conduit à identifier trois points d'entrée, qui y revêtent une dimension particulière : la protection de l'image, celle de la voix, et celle du style et de la réputation.

a) La protection de l'image

L'image de la personne, entendue comme étant, au regard de la présente mission, la représentation qu'en fait un contenu généré ou manipulé par IA (c'est-à-dire une image de la personne), est protégée par plusieurs séries de dispositions, dont il s'agit surtout, ici, d'évaluer les limites au regard de la problématique des hypertrucages.

La première de ces protections est celle du droit civil, et en particulier du droit à l'image. Bien que ne figurant pas expressément à l'article 9 du code civil, **le droit à l'image est largement reconnu par le droit français parmi les droits de la personnalité** consacrés depuis la loi du 17 juillet 1970. Ce droit, exclusif et absolu de la personne sur son image[169], porte sur la captation, la conservation, la reproduction et l'utilisation de l'image, ce qui permet de couvrir toute la chaîne d'édition et de diffusion d'un hypertrucage[170], tant en ce qui concerne les fournisseurs que les déployeurs. Il protège notamment contre toute utilisation dans un sens volontairement dévalorisant de l'image d'une personne, justifiant que le juge prenne toutes mesures propres à faire cesser l'atteinte aux droits de la personne[171]. La manière dont ce droit à l'image est invocable à l'égard d'hypertrucages représentant des personnes est particulièrement intéressante du fait des conditions qui l'encadrent : il y a méconnaissance de ce droit, en effet, en l'absence de consentement à l'utilisation de l'image[172], lorsque la personne est identifiable[173]. Cette condition fait l'objet d'une appréciation relativement souple, même si celle-ci bénéficie surtout aux personnes célèbres : les juges du fond ont ainsi pu admettre qu'un animateur de télévision, dont l'image était utilisée dans le

---

[167] Jonathan Elkaim, « Quand les acteurs donnent de la voix face à l'IA », *La Semaine Juridique Entreprise et Affaires*, n° 37, 12 septembre 2024, 1260 ; Valérie-Laure Benabou, « Doublage des voix : les professionnels face à l'IA », *Le club des juristes*, 11 février 2026 (en ligne) ; Christian Le Stanc, « L'IA, l'image, la voix », *Propriété industrielle*, n° 3, mars 2026, repère 3.

[168] Cette situation n'est pas propre à la France ou l'UE, puisqu'on la retrouve aux Etats-Unis : voy. Robert Chesney, Danielle Citron, « Deep Fakes : A Looming Challenge for Privacy, Democracy, ans National Security », *California Law Review*, 2019, vol. 107, p. 1753.

[169] Cass. 2e Civ., 30 juin 2004, pourvoi n° 02-19.599, Bull., 2004, II, n° 340.

[170] Cass. 1re Civ., 2 juin 2021, pourvoi n° 20-13.753, au Bull.

[171] Cass. 1re Civ., 16 juillet 1998, pourvoi n° 96-15.601, Bull. civ. I, n° 259.

[172] Par ex. Cass. 2e Civ., 10 mars 2004, pourvoi n° 02-16.354, Bull. 2004, II, n° 118 ; 1re Civ., 20 mars 2007, pourvoi n° 06-10.305, Bull. 2007, I, n° 125.

[173] Cass. 1re Civ., 21 mars 2006, pourvoi n° 05-16.817, Bull. 2006, I, n° 170 ; 1re Civ., 5 avril 2012, pourvoi n° 11-15.328, Bull. 2012, I, n° 86.

cadre d'une campagne publicitaire sur internet, pouvait invoquer son droit à l'image pour l'utilisation non consentie de son image, de son nom et de sa voix donnés à un sosie à la ressemblance frappante, censé le représenter[174].

Le développement de ce droit, bien qu'important, **n'en demeure pas moins sujet à débats, notamment en ce qui concerne les conditions de sa contractualisation et, plus largement, de sa patrimonialisation**. Bien qu'une évolution du droit français en ce sens soit régulièrement souhaitée par la doctrine[175], en cohérence avec l'évolution d'autres droits, notamment en référence au *right of publicity*[176] mais selon des modalités qui sont elles-mêmes débattues[177], la Cour de cassation – à l'inverse des juges du fond – s'en tient à une grande réserve, son arrêt principiel du 11 décembre 2008[178] étant même parfois qualifié d'alambiqué[179]. Or, si une cession contractuelle du droit à l'image est ainsi admise par la jurisprudence, avec une rémunération afférente, l'absence de patrimonialisation a pour effet l'extinction pour l'essentiel de ce droit au décès de la personne[180], l'absence de toute invocabilité par des tiers (à l'inverse des droits reconnus aux OGC en droit de la propriété intellectuelle) et l'absence de tout droit voisin, à l'inverse du droit de la propriété intellectuelle invoqué dans cette même affaire de 2008.

Une autre limite du droit à l'image, particulièrement notable au regard des hypertrucages, tient à la conciliation qui doit être opérée avec d'autres droits et libertés fondamentaux[181], en particulier la liberté d'expression, le droit à l'information et la liberté de la création artistique[182]. Dans l'équilibre qui doit être recherché par le juge entre ces droits, la liberté de création artistique est d'une intensité évidemment moindre que ce qui relève de la liberté de l'information[183], mais les frontières sont nécessairement moins nettes et, en réalité, c'est la question même de ce qu'est l'art à l'heure de l'IA générative qui est soulevée, ainsi que celle de la faculté de fabriquer, par un simple prompt, un contenu utilisant l'image d'une personne[184]. L'une des difficultés tient aux rôles distincts des fournisseurs et des déployeurs, posant par exemple la **question de l'imputation de la responsabilité** de l'inclusion de l'image d'une personne dans un contenu selon que le prompt utilisé par le déployeur se réfère expressément ou non à la personne. Il semble à tout le moins que la liberté de la création artistique ne devrait pas pouvoir être invoquée de manière abusive et qu'une démarche de création artistique devrait pouvoir être démontrée pour bénéficier d'une atténuation du droit à l'image alors que les outils permettent aisément, bien que parfois involontairement, des atteintes à ce droit.

Si le droit civil constitue un outil central dans la protection de l'image de la personne, **le droit pénal est également susceptible d'être mobilisé**, en particulier dans le cadre du délit prévu et réprimé par l'article 226-8 du code pénal. Cette infraction, récemment adaptée aux hypertrucages par la loi SREN, vise à protéger la représentation de la personne, de sorte que, **si le seul montage réalisé sans le consentement de la personne n'est pas sanctionné, sa simple diffusion ou le seul fait qu'il soit porté à la connaissance d'un tiers par quelque voie que ce soit rend l'auteur passible d'une sanction**. Le législateur a donc souhaité élargir le champ de cette infraction tout en maintenant que l'infraction vise le diffuseur, et non à l'auteur en tant que tel[185] : si ce choix a pu, déjà au regard de la rédaction antérieure à la

---

[174] TGI Nanterre, 23 mars 2007, *CCE* 2007, comm. 75, obs. Agathe Lepage. Voy. Agathe Lepage, V° « Droits de la personnalité », *Encyclopédie de droit civil*, Dalloz, juillet 2022, paragr. 250.

[175] Notamment Théo Hassler, « Quelle patrimonialisation pour le droit à l'image des personnes ? Pour une recomposition du droit à l'image », *Légipresse*, 2007, II, p. 123 ; P. Tafforeau, « L'être et l'avoir ou la patrimonialisation de l'image des personnes », *CCE* 2007, étude 9. *Contra*, L. Marino, « Les contrats portant l'image des personnes », *CCE*, 2003, étude 7.

[176] Jennifer Rothman, *The Right of Publicity. Privacy Reimagined for a Public World*, Harvard University Press, 2018.

[177] Voy. à ce sujet Julie Mattiussi, *L'apparence de la personne physique. Pour la reconnaissance d'une liberté*, LEH Edition, collection Thèses, 2018, p. 74-75.

[178] Cass. 1re Civ., 11 décembre 2008, pourvoi n° 07-19.494, Bull. 2008, I, n° 282 ;

[179] Agathe Lepage, préc., paragr. 157. Voy. également Grégoire Loiseau, « La crise existentielle du droit patrimonial à l'image », *Dalloz*, 2010, p. 450.

[180] Cass. 1re Civ., 22 octobre 2009, pourvoi n° 08-10.557, Bull. 2009, I, n° 211.

[181] Par ex. Cass. 1re Civ., 30 septembre 2015, pourvoi n° 14-16.273, Bull. 2015, I, n° 228.

[182] Dont on rappellera qu'elle a été consacrée par le législateur à l'article 1er de la loi n° 2016-925 du 7 juillet 2016 relative à la liberté de la création, à l'architecture et au patrimoine.

[183] Cass. 1re Civ., 9 juillet 2003, pourvoi n° 00-20.289, Bull. 2003, I, n° 172.

[184] Voy. à ce sujet les réflexions de Jim Gabaret, *L'art des IA*, PUF, 2025.

[185] Evan Raschel, « Retour sur les principales dispositions répressives de la loi SREN : *deepfake* et bannissement numérique », *Légipresse*, 2025, p. 21

loi SREN, être regretté[186], il présente un avantage important à l'heure de l'IA générative, puisqu'il permet de lever la – délicate[187] – question de la responsabilité que le système d'IA pourrait avoir dans la création du contenu pour s'intéresser à la seule action nécessitant une intention personnelle. En outre, si les développements des capacités des IA génératives pour produire des contenus d'un très grand réalisme est de nature à rendre *de facto* inopérante la réserve prévue par le législateur du contenu dont il serait évident qu'il ne constitue pas un montage, la question de l'image représentée et de son identification peut s'avérer floue lorsque des traitements algorithmiques successifs sont appliqués sur l'image de la personne : la question se pose de savoir jusqu'à quel point la personne est encore représentée lorsque son image est soumise à un filtre[188]. Il y a certes une appréciation d'espèce, qui n'est pas inédite (on pense à la question de la pixélisation des images déjà abordée en jurisprudence[189]) mais réinterroge l'exception d'évidence, que la diffusion successive des contenus peut rendre délicate à apprécier par chaque déployeur.

Par ailleurs, le droit pénal trouve sa limite dans le fait que la mention d'une intervention de l'IA fait obstacle à la qualification de l'infraction : **il s'agit donc d'un outil sanctionnant des utilisations non mentionnées, mais pas d'un instrument permettant de manière générale de préserver les intérêts des acteurs des secteurs de la culture et de la création face à des utilisations non consenties**.

Un troisième ensemble de protections peut être trouvé dans le droit des données à caractère personnel (le volet pénal de ce droit n'étant toutefois pas abordé ici), dont l'application est réservée par le RIA – et le règlement omnibus en cours de négociation pourrait modifier le RIA sur ce point afin de renforcer ce maintien des garanties du RGPD. L'image constitue une donnée personnelle identifiante, dont il est admis qu'elle constitue une donnée à caractère personnel au sens du règlement général sur la protection des données[190], outre la qualification de donnée biométrique qu'elle peut avoir si elle satisfait aux critères de la définition posée par l'article 4, point 14, du RGPD. L'utilisation de données biométriques par les systèmes d'IA est largement restreinte par l'article 5 du RIA, mais ces cas d'interdiction, de même que les restrictions prévues par le RGPD n'intéressent pas directement le secteur de la culture et de la création[191], de sorte que la qualification de donnée biométrique n'y a pas de portée utile décisive et que l'enjeu est de déterminer l'effectivité de la protection générale offerte par le RGPD à l'égard de l'image des personnes. A cet égard, **l'un des enjeux porte sur la licéité du traitement en l'absence de consentement de la personne** mais la question se pose différemment selon que l'on s'intéresse au fournisseur du SIA ou au déployeur : la protection des données intervient essentiellement au stade du recueil et du traitement des données, et revêt une prégnance moins forte s'agissant d'un contenu qui, étant généré, peut avoir utilisé des éléments d'identification sans être identifiant d'une personne physique ou alors d'une manière qui rend difficile la mobilisation de ces normes[192]. En outre, dans le cas du fournisseur, celui-ci se borne à récolter des images qui sont disponibles par ailleurs (à supposer qu'elles soient légales, la problématique étant alors différente), ce qui est susceptible de relever du régime de l'article 6, paragraphe 4, du RGPD ; pour le second, la question est plus prégnante, mais les libertés d'expression, de communication ou de la création sont susceptibles d'être invoquées pour établir la licéité du traitement – ce qui devrait également conduire à ce que celui qui invoque un tel fondement soit dûment tenu d'en justifier. Sous ces réserves, la personne concernée – elle seule et tant qu'elle est vivante – pourra exercer ses droits auprès des opérateurs concernés : **la protection de l'image comme donnée personnelle est donc susceptible d'être réelle, mais son effectivité n'est pas toujours évidente**.

---

[186] Voy. à cet égard Jean-Claude Planque, Blandine Cloez, Léa Lelièvre, « Du porno trop faux pour être vrai ! », *Dalloz*, 2024, p. 973, ainsi que l'Avis sur la protection de l'intimité des jeunes en ligne de la Commission nationale consultative des droits de l'homme, A – 2025 – 1, du 23 janvier 2025.

[187] Sur ce point, voy. Pierre-François Laslier, *Réseaux sociaux numériques et responsabilité pénale*, thèse, éd. Mare Martin, 2026 et « La responsabilité pénale du fait de l'intelligence artificielle : une approche par les risques ? », *Dalloz*, 2025, p. 1064.

[188] Voy. Agathe Lepage, « Le droit au respect de la vie privée », *in* Bernard Teyssié, *Les métamorphoses du droit des personnes*, LexisNexis, 2023, p. 217.

[189] Cass. 1re Civ., 21 mars 2006, pourvoi n° 05-16.817, Bull. civ. I, n° 170.

[190] CJUE, 11 décembre 2014, *Ryneš*, C-212/13, point 22 ; 14 février 2019, *Buivids*, C-345/17, points 31-32.

[191] Dans le cadre du RGPD, les restrictions liées à l'utilisation de données biométriques portent sur les traitements utilisant ces données aux fins d'identifier une personne physique de manière unique (art. 9, paragr. 1). Au demeurant, les pratiques prohibées par le RIA en matière d'intelligence artificielle concernent aussi les systèmes d'identification ou de catégorisation biométrique.

[192] Felipe Romero Moreno, « Generative AI and deepfakes: a human rights approach to tackling harmful content », *International Review of Law, Computers & Technology*, vol. 38, 2024, p. 297.

Le cumul de ces trois ensembles de législations montre que **l'image d'une personne est dans l'ensemble protégée face à une utilisation non consentie dans le cadre d'un hypertrucage, sous réserve que la liberté du fournisseur ou du déployeur ne prime, tandis que le droit de la propriété intellectuelle n'est pas, sur l'image elle-même, invocable**, du fait de la jurisprudence constante de la Cour de cassation depuis 2008[193]. **Il en résulte, précisément, une protection moindre que celle que peut offrir ce droit, compte tenu de l'absence de pleine patrimonialisation du droit à l'image**, ce qui conduit à s'interroger sur d'éventuelles évolutions à cet égard, qui seront évoquées par la suite (point 3.2 ci-après).

Le droit de la propriété intellectuelle n'est toutefois pas absent, en particulier pour les secteurs culturels et créatifs : dans les mêmes conditions que celles qui seront évoquées ci-après pour la voix, mais avec une fréquence moindre à ce jour (bien que l'apparition de Val Kilmer dans le récent film *As Deep as the Grave* sorti un an après son décès[194] ou le faux film représentant un combat entre Tom Cruise et Brad Pitt[195] ouvrent des perspectives encore largement indéterminées), l'image de l'artiste dans l'interprétation est protégée, ce qui inclut les droits moraux attachés[196]. **C'est néanmoins la seule interprétation qui bénéficie ainsi de garanties**. Le récent dépôt par Taylor Swift auprès de l'USPTO d'une image d'elle afin de se protéger[197] – et prévenir l'utilisation qui en a été faite lors de la campagne présidentielle américaine de 2024 – met en lumière ces carences. Ce dépôt rappelle l'enregistrement comme marque du visage de Marlene Dietrich en Allemagne, admis par le *Bundesgerichtshof* en 2008[198] et s'inscrit dans la continuité de précédents, y compris en Europe où, selon une étude de 2024, l'EUIPO aurait enregistré 57 marques de portraits masculins et 23 de portraits féminins[199]. Toutefois, il ne s'agit pas d'une acceptation générale de l'EUIPO : dans d'autres affaires, elle a refusé des enregistrements de visages, dans la mesure où elle tient compte du caractère distinctif du portrait et de la notoriété suffisante de la personne au sein de l'Union européenne[200]. Demeure actuellement pendant un renvoi de la seconde chambre d'appel de l'EUIPO devant la grande chambre pour savoir si l'enregistrement du visage d'un chanteur et présentateur néerlandais pouvait être accepté[201].

En tout état de cause, dans l'attente d'une position de la grande chambre, il faut être conscient que **cette démarche dans le cadre du droit de la propriété intellectuelle**, qui peut aller jusqu'à la recherche de protection d'un « jumeau numérique », laquelle soulève d'autres enjeux, y compris sur le rapport à la création[202], **ne pourra offrir qu'une protection qui s'avèrera limitée à des cas exceptionnels**, intéressant certes les personnes les plus exposées des secteurs de la culture et de la création, mais ne pourra répondre aux enjeux massifs et systémiques posés par le développement des hypertrucages. Par ailleurs, elle n'est pas en phase avec la protection de l'interprétation, telle qu'elle est habituellement conçue : celle-ci est en effet attachée à l'exécution dans sa matérialité même, plus qu'aux caractéristiques intrinsèques, indépendamment de la fixation.

b) <u>La protection de la voix</u>

Il est généralement admis que le droit à la voix est un dérivé du droit à l'image, étant présenté comme « *l'image sonore* » de la personne[203]. La voix est expressément mentionnée à l'article 226-8 du code

---

[193] Arrêt précité de la première chambre civile du 11 décembre 2008.
[194] « L'acteur Val Kilmer bientôt ressuscité dans un film grâce à l'intelligence artificielle », *Le Figaro*, 19 mars 2026.
[195] « L'improbable bagarre entre Brad Pitt et Tom Cruise au sommet d'un immeuble met en rage Hollywood », *Le Figaro*, 16 février 2016.
[196] CJUE, gr. ch., 3 septembre 2014, *Deckmyn et Vrijheidsfonds*, C-201/13.
[197] Voy. à ce sujet, Leticia Caminero, « Taylor Swift trademark strategy: a model for artist IP protection », *WIPO Magazine*, 19 août 2025 et actualisé au regard des dépôts intervenus le 1er mai 2026 (en ligne).
[198] Bundesgerichtshof, 24 avril 2008, *Marlene Dietrich Bildnis I*, I ZB 21/06.
[199] Barna Arnold Keserű, « Trademark protection for faces? A comprehensive analysis on the benefits and drawbacks of trademarks and the right to facial image », Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 2024, vol. 15, n° 1, p. 88.
[200] Voy. par ex., EUIPO, 16 novembre 2017, aff. R 2063/2016-4.
[201] Aff. R 0050/2024-2, renvoi du 26 septembre 2024.
[202] A titre d'éléments de réflexion, voy. par ex. Roberto Minerva *et al.*, « Artificial Intelligence and the Digital Twin: An Essential Combination », *in* N. Crespi, A.T. Drobot, R. Minerva (éd)., *The Digital Twin*, Springer, 2023, p. 299 ; Seohoo Yi, « Avatar beyond replication: digital human thought twins as co-creators through AI mind design in Eastern and Western thought », *AI 1 Society*, 2026.
[203] Daniel Bécourt, *Image et vie privée*, L'Harmattan, 2004, p. 185.

pénal[204] ; elle est usuellement admise comme étant incluse dans le champ de l'article 9 du code civil au titre des droits de la personnalité[205] et elle est considérée comme une donnée à caractère personnel au sens du RGPD[206], susceptible également de constituer une donnée biométrique.

Pour autant, **la jurisprudence n'a jamais expressément consacré un droit sur la voix** et les illustrations sont plus rares, se concentrant sur les juges du fond, avec des affaires anciennes notamment, concernant Léon Zitrone, Claude Piéplu ou Maria Callas. La jurisprudence civile de la Cour de cassation ne lui a pas donné son autonomie par rapport au droit à la vie privée, retenant la circonstance d'une captation intervenue en méconnaissance de ce droit dans un lieu privé[207]. La prise en compte de la voix retrouve néanmoins une attention particulière au regard de l'IA générative et des multiples usages qui ont pu soulever des difficultés : génération de musiques entièrement par IA, bouleversant l'équilibre du secteur musical ; doublage des films, avec la réutilisation de voix antérieurement enregistrées, au risque de remettre en cause à long terme l'existence même de la profession ; utilisation de voix d'acteurs célèbres sans leur consentement par des systèmes d'IA comme dans le cas de Scarlett Johansson… En conséquence, la question retrouve également une nouvelle dimension juridique : aux Etats-Unis, des tentatives ont été faites pour protéger la voix (cas de Matthew McConaughey ou de Taylor Swift[208]) ; en France, la question trouve une nouvelle vigueur devant les prétoires, avec notamment un récent arrêt de la cour d'appel de Paris qui a reconnu une atteinte illicite au droit de l'artiste Grand Corps Malade à sa voix, qualifiée d' « *attribut sonore et élément d'identification d'une personne* » qui « *constitue un des attributs de sa personnalité et bénéficie de la protection instituée par l'article 9 du code civil* », de sorte que chacun dispose d'un droit exclusif et absolu sur sa voix[209]. Le pourvoi en cassation concernant cet arrêt n'ayant pas, à la date du présent rapport, été jugé, la mission s'en est tenue à la position de la cour d'appel.

Cette différence d'intensité de la jurisprudence entre l'image et la voix tient aux **particularités de la voix qui soulèvent des enjeux spécifiques et partiellement distincts**.

La première tient au fait que **le droit de la propriété littéraire et artistique peut ici trouver, plus aisément que l'image, un certain rôle de protection**, à la fois en ce que les œuvres orales sont protégées par l'article L. 112-1 du code de la propriété intellectuelle, et en ce que l'interprétation donnée par un artiste-interprète bénéficie également d'une protection au titre des droits voisins de l'article L. 212-1 de ce code[210]. Cette protection, qui garantit le consentement de l'interprète et des titulaires de droits voisins, est large, indépendante de la qualité du support de fixation[211] et permet au titulaire d'un droit exclusif sur un phonogramme de s'opposer à l'utilisation par un tiers d'un échantillon sonore, même très bref, de son phonogramme aux fins de l'inclusion de cet échantillon dans un autre phonogramme, à moins qu'il ne soit inclus sous une forme modifiée et non reconnaissable à l'écoute[212]. Elle est toutefois partielle : elle bénéficie certes largement aux acteurs des secteurs de la culture et de la création mais elle suppose pour être efficace l'utilisation de voix reprises d'enregistrements d'œuvres orales ou d'interprétation, alors même que la voix des artistes peut être enregistrée dans le cadre d'autres interventions non protégées à l'un ou l'autre de ces titres (une interview par exemple). Même s'il est possible de considérer que c'est la manière dont les attributs de la personnalité de l'artiste sont utilisés dans le cadre d'une performance qui intéresseront principalement les créateurs d'hypertrucages[213], c'est une limite importante dans la prévention de la manipulation, quelqu'en soit le but et l'usage. Les limites de la notion d'interprétation face

[204] Pour une application : Cass. Crim., 21 avril 2020, pourvoi n° 19-81.507, au Bull.
[205] Agathe Lepage, préc., paragr. 131.
[206] Dans le cas des jouets connectés : déc. n° MED-2017-073 de la CNIL du 20 novembre 2017 (*Communication commerce électronique*, 2018, comm. 15). Voy. également « A votre écoute », livre blanc de la CNIL sur les assistants vocaux, 7 septembre 2020.
[207] Cass. 1re Civ., 6 octobre 2011, pourvoi n° 10-21.822 , Bull. civ. I, n° 161.
[208] « Matthew McConaughey fait breveter son image pour la protéger de l'IA », *Le Monde*, 15 janvier 2026 ; « Taylor Swift veut faire de sa voix une marque déposée pour la protéger des modèles d'IA », *Le Monde*, 28 avril 2026
[209] CA Paris, pôle 5, chambre 2, 17 octobre 2025, RG n° 23/11112.
[210] Voy. à ce sujet Aline Vignon-Barrault, V° « Droit à réparation. Responsabilité fondée sur la faute. Atteintes aux droits de la personnalité. Les droits de la personnalité », *JurisClasseur Responsabilité civile et Assurance*, fasc. 133-20.
[211] Cass. 1re Civ., 24 septembre 2009, pourvoi n° 08-11.112, Bull. 2009, I, n° 184.
[212] CJUE, gr. ch., 29 juillet 2019, *Pelham e.a.*, C-476/17.
[213] Valérie-Laure Benabou, « Deepfakes and Private Rights in the Perspective of EU Law : Is It Necessary to Intervene ? », *Columbia Journal of Law & The Arts*, 2026, vol. 49, n° 4, p. 729.

aux usages de la voix par les systèmes d'IAG ont déjà été soulignées dans un précédent rapport de mission du CSPLA[214].

La seconde concerne la difficulté particulière qui peut exister à identifier une voix. Certes, il existe un ensemble d'éléments qui permettent d'individualiser une voix, mais l'attribution en tant que telle à une personne – condition nécessaire à l'invocation des droits de la personne – n'est pas évidente, à l'exception de voix disposant de certaines particularités – tel était le cas dans l'affaire précitée concernant l'artiste Grand Corps Malade, l'arrêt soulignant la nécessité que la voix reproduite rende la personne identifiable, bien qu'en l'espèce ce ne fût pas un enjeu du débat contentieux. Au demeurant, certaines auditions menées par la mission ont relevé le caractère délicat, même pour les experts judiciaires, du rattachement d'une voix à une personne précise. Si, s'agissant de l'image, l'identification de la personne suppose la réunion d'un ensemble d'éléments établissant la ressemblance accessible au juge, les éléments d'identification d'une voix le sont moins aisément. C'est d'ailleurs la raison pour laquelle l'acteur McConaughey a déposé certaines intonations ou expressions qui lui sont habituellement attachées et qui constituent, en quelque sorte, une part de son identité d'acteur.

Cette capacité à établir la preuve d'une utilisation d'une voix d'une personne est un enjeu désormais majeur au regard des capacités des systèmes d'IAG ; elle révèle une limite inhérente à la voix qui est la détermination d'un caractère identificatoire, supposant une particularité suffisamment individuelle pour assurer l'identification. A cet égard, une simple intonation ne semble pas de nature à permettre cette identification à elle seule, mais la prise en compte d'un ensemble d'éléments n'est pas toujours de nature à assurer l'identification de la personne titulaire d'une voix (sans compter les changements potentiels de celle-ci au cours de la vie) : on comprend à cet aune la tentative de dépôt de voix comme marque en propre – en rupture avec la logique du droit de la propriété intellectuelle sur ce point qui ne nécessite pas d'enregistrement préalable, réinterrogeant ce principe[215] – tandis que la réutilisation d'éléments vocaux est par ailleurs permise au titre du pastiche[216].

Ainsi, malgré une protection élargie par rapport à l'image au bénéfice du droit de la propriété intellectuelle, **la voix repose sur un équilibre d'autant plus délicat que l'émergence de l'IA générative recouvre deux hypothèses : celle, immédiate, d'utilisation d'une voix préalablement enregistrée ; celle, plus indirecte, de génération d'une voix « nouvelle » au bénéfice de l'apprentissage d'enregistrements qui peuvent, eux-mêmes, être identifiables**.

c) <u>Au-delà de l'image et de la voix : le style et la réputation</u>

Les nombreuses manipulations de l'image et de la voix d'une personne susceptibles d'être effectuées par les outils d'IA générative soulèvent des questions au-delà de ces deux attributs immédiats, en particulier en ce qui concerne la protection du « style », autrement dit ce qui participe de l'identité artistique d'une personne (dessiner, écrire, chanter, interpréter à la manière de…) ou, pour reprendre les termes de Paul Valéry, « *la manière dont quelqu'un s'exprime, quoi qu'il exprime, cette manière étant révélatrice de sa nature, abstraction faite de sa pensée actuelle* »[217]. L'enjeu est celui, déjà abordé par une mission du CSPLA, des « *faux stricto sensu* »[218]. En conséquence, ces manipulations soulèvent également des enjeux réputationnels, qui ne seront évoqués ici qu'en ce qu'ils découlent du style, mais dont on mentionnera que, par ailleurs, ils disposent de nombreuses réponses dans le droit positif (droit civil, droit de la concurrence, droit de la presse, droit de la consommation…).

Paradoxalement, le style interprétatif d'une voix dispose d'une protection potentiellement plus assurée à ce jour, bien que, compte tenu des limites actuelles de l'interprétation de ces dispositions[219], elle

---

[214] *Rapport de mission sur les assistants vocaux et autres agents conversationnels*, établi par Célia Zolynski, Karine Favro et Serena Villata, mars 2023, not. p. 43-45.

[215] Question soulevée de manière générale au regard des outils numériques, notamment en droit américain, alors que ce pays avait rompu avec la logique formaliste : Stef Van Gompel, « Formalities in the digital era : an obstacle or opportunity ? », *in* L. Bently, U. Suthersanen, P. Torremans (éd.), *Global Copyright. Three Hundred Years Since the Statute of Anne from 1709 to Cyberspace*, Edward Elgar, 2010, p. 395 et, dans le même ouvrage, Jane Ginsburg, « The US experience with formalities : a love/hate relationship », p. 425.

[216] La définition de cette notion vient d'être précisée par la CJUE : gr. ch., 14 avril 2026, *Pelham*, C-590/23.

[217] Paul Valéry, V° « Styles », *Vues*, éd. De la Table ronde, 1948.

[218] Rapport de mission sur les faux artistiques, établi par Tristan Azzi et Pierre Sirinelli, 2024.

[219] Rapport de mission du CSPLA précité sur les assistants vocaux.

nécessite un élargissement de la portée des articles L. 212-2 et L. 212-3 du code de la propriété intellectuelle[220]. S'il n'est pas possible de définir les droits voisins au-delà de ce que prévoit la directive du 22 mai 2001, **il demeure une marge d'appréciation pour assurer (et ainsi répondre à une inquiétude des artistes-interprètes) que la reprise et la modification des fixations ainsi protégées bénéficient de cette protection lors de leur diffusion**, en cohérence avec l'appréciation stricte des cessions effectuées[221]. A cet égard, il peut être envisagé que, par un pas supplémentaire, le style interprétatif, lorsqu'il identifie spécifiquement une personne, soit à tout le moins reconnu, en particulier lorsqu'il est particulièrement distinctif : il est en effet nécessaire, comme l'a rappelé récemment la Cour de justice pour définir la notion de *pastiche* que les éléments pris en compte soient protégés par le droit d'auteur (et les droits voisins ici) et qu'ils permettent de reconnaître la personne concernée pour celui qui en connaît l'œuvre[222]. Néanmoins, une telle évolution devrait être suffisamment encadrée pour préserver la liberté de création, le droit d'auteur ayant vocation à la soutenir et non à l'entraver et devant être interprété dans cette perspective[223] ; le contexte du litige est alors susceptible de constituer un élément déterminant.

En revanche, le style d'une œuvre n'est en principe pas protégé par le droit d'auteur, puisqu'il n'y a pas, par définition, reproduction d'une œuvre préexistante, condition nécessaire à l'invocation du droit d'auteur[224]. Le développement des IAG a mis en lumière, y compris aux Etats-Unis[225], les limites du droit d'auteur pour protéger cet aspect. Le fonctionnement de ces systèmes ouvre certes la possibilité qu'au-delà du style, des éléments mêmes d'une œuvre protégée soient repris de manière suffisante pour qu'une contrefaçon soit caractérisée[226], mais cette hypothèse ne suffit pas à couvrir l'ensemble de la problématique. C'est l'un des enjeux du contentieux très récemment initié en Allemagne par Penguin Random House contre OpenIA, en ce qu'il propose de générer de nouveaux manuscrits inspirés d'œuvres au catalogue de cet éditeur, sans même que l'utilisateur demande une telle ressemblance[227] : l'objectif est d'essayer d'agir contre le fournisseur du système d'IA sur le terrain de l'entraînement, terrain sur lequel une juridiction de Munich avait déjà retenu une violation par OpenAI de la loi allemande sur le droit d'auteur du fait du moissonnage réalisé par ChatGPT[228]. C'est également une question au cœur de l'action de groupe engagée à l'encontre de l'outil Grammarly et sa fonction « *Expert Review* », qui offre des avis inspirés du style d'auteurs célèbres[229] : l'un des axes de ce contentieux s'appuie sur la protection du nom par le *right of publicity*, mais l'inadéquation du cadre juridique et procédural face à ce type d'outil a été relevée, même si cette procédure est l'occasion de réinterroger les contours de ce droit[230].

En l'absence de consécration aujourd'hui d'un droit de propriété incorporelle sur la notoriété[231], c'est alors sur le terrain du parasitisme qu'une solution peut être recherchée[232], bien que cela impose de s'extraire du cadre « naturel » de protection des œuvres et des outils, y compris processuels, qu'il offre – mais dont la jurisprudence admet la simultanéité des actions[233]. A cet égard, **la mission estime que le style n'a pas nécessairement vocation à être protégé en toute circonstance** : l'imitation d'un style n'est pas nouvelle, même si elle prend une ampleur considérable avec l'IAG compte tenu de la facilité de son utilisation, ce dont il résulte un risque accru de déstabilisation de l'image de l'auteur

---

[220] En sens, voy. également Valérie-Laure Benabou, « Doublage des voix : les professionnels face à l'IA », *Le club des juristes*, 11 février 2026 (en ligne).
[221] Cass. 1re Civ., 6 mars 2001, pourvoi n° 98-15.502, Bull. 2001, I, n° 58.
[222] CJUE, gr. ch., 14 avril 2026, *Pelham*, C-590/23. Rappr. également CJUE, 29 juillet 2019, *Pelham GmbH et autres*, C-476/17.
[223] Voy. récemment Cass. 1re Civ., 8 février 2023, pourvoi n° 21-24.980, non publié.
[224] Voy. le rapport de mission de Tristan Assi et Pierre Sirinelli, précité, p. 14.
[225] Benjamin Sobel, « Elements of Style : Copyright, Similarity, and Generative AI », *Harvard Journal of Law & Technology*, 2024, vol. 38, n° 1, p. 49.
[226] *Ibid.*
[227] « OpenAI accusé d'avoir enfreint les droits d'auteur de livres pour enfants en Allemagne », *Le Figaro*, 31 mars 2026 ; « Penguin to sue OpenAI over ChatGPT version of German children's book », *The Guardian*, 31 mars 2026.
[228] Voy. le communiqué de presse du tribunal de Munich : https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php
[229] Julia Angwin, « Why I'm Suing Grammarly », *the New York Times*, 13 mars 2026.
[230] Jennifer Rothman, « Grammarly Lawsuit Shows Existing Laws Can Combat Deepfakes », *Lawfare*, 9 avril 2026 (en ligne).
[231] Sur ce débat, voy. Jean-Michel Bruguière et Bérengère Gleize, *Droit des personnes*, éd. Dalloz, Sirey Université, 2e éd., 2025, p. 305.
[232] Voy. par ex. en ce sens Michel Vivant, Jean-Michel Bruguière, *Droit d'auteur et droits voisins*, Dalloz, 2024, 5e éd., p. 191.
[233] Cass. Com., 8 mars 2026, pourvoi n° 24-17.016, au Bull.

original. Or, précisément, **le cadre du parasitisme permet**, sur le fondement de l'article 1240 du code civil, **d'appréhender des comportements fautifs par lesquels un agent économique s'immisce dans le sillage d'un autre afin de tirer profit, sans rien dépenser, de ses efforts et de son savoir-faire**[234].

Certes, par deux arrêts du 26 juin 2024, la chambre commerciale de la Cour de cassation a restreint les conditions de reconnaissance de la faute de parasitisme[235]. Toutefois, elle juge qu'un tel comportement est caractérisé lorsque celui qui se prétend victime de tels actes identifie la valeur économique individualisée qu'il invoque, ainsi que la volonté d'un tiers de se placer dans son sillage. Dans ce cadre, elle retient que le savoir-faire et les efforts humains et financiers propres à caractériser une valeur économique identifiée et individualisée ne peuvent se déduire de la seule longévité et du succès de la commercialisation du produit et, les idées étant de libre parcours, le seul fait de reprendre, en le déclinant, un concept mis en œuvre par un concurrent ne constitue pas, en soi, un acte de parasitisme[236]. C'est le détournement de la notoriété, des investissements ou du savoir-faire par un tiers qui est ainsi protégé, mais il s'agit d'une faute intentionnelle : cela suppose que, dans le cadre d'un contenu généré par IA, cette intention puisse être établie. On peut potentiellement identifier ici une difficulté, mais **l'intention est celle du comportement parasitaire, ce qui devrait, pour l'essentiel, exclure l'hypothèse d'un contenu de nature parasitaire sans que cela ait été la demande du déployeur** : pour la victime potentielle, la question du prompt sera indifférente, puisque c'est au regard du contenu que le caractère parasitaire est apprécié ; les conditions de génération ne peuvent que relever d'un appel en garantie potentiel à l'encontre du fournisseur du système d'IAG, la charge probatoire de la victime du contenu portant sur l'utilisation parasitaire.

Ce cadre est évidemment très contraint, mais la mission estime que, sous réserve de tenir compte de la réalité des usages d'IAG qui pourront se développer, et notamment de la différence d'approche qui pourrait être retenue selon que l'obligation de transparence a été respectée ou non (son respect n'excluant toutefois pas un comportement parasitaire), **il assure le maintien d'une protection du style qui soit encadrée, sans faire obstacle à la création** – ce qui serait le cas si un style était protégé en toutes circonstances.

Au terme de l'étude de ces éléments, il apparaît que **la protection des personnes face à des contenus hypertruqués se caractérise par son caractère épars et inégal**. Le développement des systèmes d'IA générative et la faculté de générer aisément des hypertrucages soulignent les limites de cette protection : si à l'égard des fournisseurs de ces systèmes, la réglementation sur les données personnelles peut constituer un outil pertinent, à l'égard des déployeurs – et donc face à la diffusion d'un contenu hypertruqué – la situation est plus complexe. L'image et la voix sont essentiellement protégés au titre des droits de la personnalité, droits certes absolus mais qui ne sont invocables que par la personne et dans un cadre civil de droit commun. Si la voix peut bénéficier d'une protection complémentaire au titre du droit d'auteur et des droits voisins, elle présente des enjeux probatoires complexes, qui rappellent toutefois un élément important pour la mission, également valable s'agissant du style : même si la facilité à générer des hypertrucages soulève une question quant au caractère massif des risques, cela ne saurait induire que toute image, toute voix ou tout style soit, par principe, protégé. Non seulement la protection doit être limitée par les libertés en cause, mais le principe de la protection ne peut porter que sur ce qui est identifiable et générateur de valeur potentielle. C'est dans cette perspective que seront abordés les réflexions du point 3.2 sur d'éventuelles évolutions législatives ou jurisprudentielles.

3.1.2. La protection des objets, lieux, entités et événements face aux hypertrucages

La définition de l'hypertrucage recouvrant d'autres hypothèses que les personnes, quelques observations – relativement brèves dans la mesure où la mission s'est essentiellement intéressée aux enjeux liés aux personnes – peuvent être formulées concernant les objets, lieux, entités et événements.

S'agissant de l'hypertrucage d'un objet, au bénéfice de la perspective élargie de cette notion retenue dans le chapitre 1, les enjeux pour les secteurs culturels et créatifs sont essentiellement en lien avec la propriété intellectuelle et les données d'apprentissage des systèmes d'IA générative, en particulier au bénéfice de l'exception de moissonnage. Il est renvoyé à cet égard aux réflexions de la précédente mission

---

[234] Cass. Com., 27 juin 1995, pourvoi n° 93-18.601, Bull. civ. IV, n° 193 ; Com., 26 janvier 1999, pourvoi n° 99-22.457, inédit ; Com., 3 juillet 2001, pourvois nos 98-23.236, 99-10.406, Bull. civ. IV, n° 132.

[235] Voy. dans le *Recueil annuel des études 2025* de la Cour de cassation, l'étude « La faute de parasitisme : vers des conditions plus restrictives » par Mélanie Bessaud et Laure Compte, p. 85.

[236] Cass. Com., 26 juin 2024, pourvoi n° 23-13.535 et pourvoi n° 22-17.647, 22-21.497 (deux arrêts), au Bull.

menée par les professeures Bensamoun et Farchy[237] et à la proposition de loi relative à l'instauration d'une présomption d'exploitation des contenus culturels par les fournisseurs d'intelligence artificielle en cours d'examen au Parlement[238]. La diffusion d'un contenu modifié (y compris par IA) demeure la reproduction d'une œuvre protégée[239], notamment du fait que la fourniture au public par téléchargement, pour un usage permanent, d'une telle œuvre constitue une communication au public au sens de la directive 2001/29 sur le droit d'auteur et les droits voisins[240].

En ce qui concerne les entités, le droit des marques et le droit de la concurrence – y compris le parasitisme – devraient trouver à s'appliquer et cette hypothèse n'appelle pas d'observation particulière.

En revanche, il peut s'avérer plus difficile de cerner les enjeux d'un hypertrucage portant sur des événements ou des lieux. Deux volets peuvent être identifiés.

Selon toute vraisemblance, il s'agit pour l'essentiel de prémunir les personnes face à de fausses informations (générales[241]). À cet égard, le dispositif actuel repose essentiellement sur deux outils à la portée limitée : l'article 27 de la loi du 29 juillet 1881 sur la liberté de la presse qui réprime « *la publication, la diffusion ou la reproduction de nouvelles fausses, de pièces fabriquées, falsifiées ou mensongèrement attribuées à des tiers, lorsque, faite de mauvaise foi, elle aura troublé la paix publique, ou aura été susceptible de la troubler* », et l'article L. 163-2 du code électoral, issu de la loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, qui permet, seulement à l'approche d'une échéance électorale, de saisir le juge des référés du tribunal judiciaire pour faire cesser la diffusion d'informations dont il est manifeste qu'elles sont inexactes ou trompeuses et de nature à altérer la sincérité du scrutin et qui sont diffusées de manière délibérée, artificielle ou automatisée et massive, par le biais d'un service de communication au public en ligne. Ces dispositions ont donc un champ particulièrement limité, en conséquence de la nécessaire conciliation avec la liberté d'expression et il semble, en tout état de cause, délicat d'aller au-delà par un dispositif général autre que des obligations de transparence[242].

Par ailleurs, on rappellera qu'outre la protection de la vie privée, il existe un droit à l'image sur les biens au bénéfice de leurs propriétaires, bien qu'il soit d'une portée réduite[243]. Contrairement au droit d'une personne sur son image, le préjudice doit être caractérisé, ce qui rend son invocation plus complexe mais, outre le cas de l'utilisation à des fins commerciales[244], il est possible d'envisager des hypothèses où la manipulation de l'image du bien d'une personne serait de nature à lui causer un trouble anormal, notamment lorsque le contenu a pour but de tromper le destinataire. En revanche, une autre limite, potentiellement forte dans le cadre des IA génératives, tient au fait qu'un préjudice à ce titre ne peut être retenu si le bien concerné ne constitue qu'un élément accessoire de l'image litigieuse[245] : le droit de la propriété intellectuelle offre là aussi une protection plus effective dès lors que l'incorporation d'une œuvre dans un autre contenu peut conduire à la qualification d'œuvre composite, nécessitant l'autorisation des titulaires de droit conformément aux articles L. 113-2 et L. 113-4 du code de la propriété intellectuelle. La protection est en revanche plus forte concernant les biens publics[246], ce qui permet de mieux les protéger à l'égard de manipulations par IA.

Ainsi, dans le cas des objets, lieux, entités et événements, la protection apparaît également inégale : si le droit de la propriété intellectuelle trouve pleinement son expression pour protéger les œuvres qui relèvent de ce champ, la protection contre des contenus hypertruqués représentant des lieux ou événements apparaît nettement plus limitée, exposant à un risque de désinformation, renforçant l'importance d'une bonne application des obligations de transparence posées par le RIA.

---

[237] CSPLA, rapport de mission relative à la rémunération des contenus culturels utilisés par les systèmes d'intelligence artificielle, 2025.
[238] Proposition de loi présentée par M. Laure Darcos et plusieurs autres sénateurs, texte n° 220, déposé le 12 décembre 2025.
[239] CJUE, 1er décembre 2011, *Painer*, C-145/10 ; CJUE, 22 janvier 2015, *Art & Allposters International*, C-419/13.
[240] CJUE, gr. ch., 19 décembre 2019, *Nederlands Uitgeversverbond et Groep Algemene Uitgevers*, C-163/28.
[241] N'est pas abordée ici la législation sur les fausses informations en matière financière.
[242] Sauf, évidemment, à identifier des enjeux d'intérêt général aussi fondamentaux.
[243] Cass. Ass. plén., 7 mai 2004, pourvoi n° 02-10.450, Bull. A.P., n° 10 : « *le propriétaire d'une chose ne dispose pas d'un droit exclusif sur l'image de celle-ci ; qu'il peut toutefois s'opposer à l'utilisation de cette image par un tiers lorsqu'elle lui cause un trouble anormal* ».
[244] A titre d'exemple : Cass. Com., 28 juin 2012, pourvoi n° 10-28.716, inédit.
[245] Cass. 1re Civ., 25 janvier 2000, pourvoi n° 98-10.671, Bull. civ. I, n° 24.
[246] CE, Ass., 13 avril 2018, *Etablissement public du domaine national de Chambord*, n° 397047, au Rec.

3.2. Mobiliser les droits des acteurs des secteurs de la culture et de la création : des enjeux législatifs et contractuels

A la lumière de ces éléments et au regard des incidences des hypertrucages pour les secteurs créatifs et culturels, la mission a poursuivi sa réflexion sur la manière de répondre à ces enjeux en s'intéressant d'abord à la possibilité de mobilier l'outil conventionnel et, ensuite, aux évolutions juridiques qui pourraient être nécessaires. **Ces réflexions sont exclusivement centrées sur les hypertrucages concernant les personnes, qui constituent le volet le plus sensible, au regard du champ de la mission, dans les secteurs culturel et créatif.**

3.2.1. Le cadre contractuel : un vecteur porteur d'importantes potentialités

Les outils contractuels constituent d'ores et déjà un cadre de réflexion pour traiter les enjeux des hypertrucages. Leur mobilisation **repose toutefois sur un postulat : celui du consentement de la personne** à l'utilisation d'éléments de sa personnalité ou de son patrimoine (droits d'auteur et droits voisins) par une IA générative, ce qui suppose que l'utilisation des données correspondantes repose elle-même sur un consentement. Les réponses au questionnaire de la mission et les auditions qu'elle a menées ont souligné une difficulté importante à cet égard, alors que le respect des oppositions (« *opt out* ») à la fouille de textes et de données, exprimées en application de la directive du 17 avril 2019 sur le droit d'auteur et les droits voisins[247], n'est pas évident face à des systèmes dont le fonctionnement n'est pas toujours transparent[248] et dont la capacité de désapprentissage interroge, alors même que cette exception de fouille n'a pas été conçue en vue de l'entraînement de ces systèmes[249]. Les réflexions qui suivent sur le terrain contractuel, tout en encourageant au développement de ces outils, ne méconnaissent pas les pratiques non respectueuses de ce consentement, qui seront prises en compte au titre du cadre législatif.

Par ailleurs, il ne peut être fait abstraction de **deux facteurs importants**. Le premier tient au fait que **ne peut être contractualisé que ce que la loi permet**, ce qui renvoie aux contours des différents droits reconnus par le législateur : un pan substantiel de la question soulevée par les hypertrucages tient à l'exploitation de contenus moissonnés par les systèmes d'IA sans aucune autorisation préalable d'une quelconque manière. Le second concerne la **dimension extraterritoriale des risques** liés aux pratiques d'hypertrucage : si cela ne saurait être un motif de renoncement à l'action, cette circonstance est de nature à moduler la pertinence des solutions envisageables.

a) La contractualisation individuelle

Dans son champ, le droit de la propriété intellectuelle offre des capacités contractuelles relativement fortes, notamment en ce qu'il permet d'impliquer d'autres personnes que le détenteur du droit d'auteur et organise les conditions d'exploitation *post mortem*. **De telles possibilités ne sont pas ouvertes, à ce jour, lorsque les éléments à protéger relèvent des droits de la personnalité** : aussi absolus et exclusifs soient-ils, ceux-ci relèvent du droit commun des obligations et sont exclusivement attachés à la personne concernée. En outre, les contrats régis par le code de la propriété intellectuelle prévoient des garanties, qui relèvent, à défaut de cadre spécifique pour les droits de la personnalité, de la liberté contractuelle – avec les risques que cela présente pour des parties en situation de déséquilibre[250] – ce qui explique les tentatives de soumettre les contrats portant sur le droit à l'image au formalisme prévu par le code de la propriété intellectuelle, sans que la Cour de cassation ait donné suite depuis 2008.

Or, dans ce cadre, **les contrats de droit commun portant sur les attributs de la personnalité soulèvent des difficultés tenant à la prise en compte des potentialités d'exploitation des attributs de la personnalité dans des hypertrucages**. Les auditions menées par la mission ont en effet souligné l'absence de prise en compte de ces enjeux par des contrats contenant des clauses d'exploitation illimitée dans le temps et quant à l'objet de cette exploitation. Cette question se pose, au demeurant, tant en ce qui

[247] Directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE.
[248] Voy. le rapport de la mission du CSPLA relative à la mise en œuvre du règlement européen sur l'intelligence artificielle, établi par Alexandra Bensamoun, décembre 2024.
[249] Nicola Lucchi, Serra Hunter, *Generative AI and copyright : Training, Creation, Regulation*, étude juridique pour la commission des affaires juridiques du Parlement européen, doc. PE 774.095.
[250] Grégoire Loiseau, *Le droit des personnes*, Ellipses, 2de éd., 2020, p. 301.

concerne les contrats antérieurs à l'émergence et au développement des systèmes d'IA générative que des contrats plus contemporains, dans lesquels la rémunération n'aurait pas été adaptée. Elle est d'autant plus forte pour les acteurs des secteurs culturels et créatifs que ces contrats revêtent en général la forme de contrats d'adhésion auxquels les intéressés ne peuvent apporter de modification. La solution retenue par la chanteuse Holly Herndon, qui a mis à disposition le clone numérique de sa voix (Holly+) aux fins d'usage créatif à condition qu'il soit traçable, approuvé et qu'une partie des revenus générés lui soit reversée[251], n'est guère reproductible à l'ensemble du secteur.

Pour les contrats « passés », l'absence de pleine patrimonialisation du droit à l'image permet certes d'agir en cas d'atteinte à la dignité de la personne ou à l'ordre public, mais la portée de ces deux réserves est limitée face à l'ampleur des potentialités qui se présentent. Toutefois, **il ne paraît pas admissible, compte tenu de l'importance que revêtent les droits de la personne, en particulier de leur caractère absolu et exclusif, qu'un contrat conclu à une époque où les modalités d'exploitation actuelle ne pouvaient être anticipées, puisse entièrement lier la personne** et qu'une protection au-delà de son décès ne soit réellement accessible. A cet égard également, le droit commun des contrats est moins protecteur que celui des contrats spéciaux en matière de propriété intellectuelle[252]. Pour l'essentiel, il sera délicat d'identifier un motif viciant le consentement de la personne lors de la conclusion du contrat : des arrêts d'espèce ont certes admis la remise en cause du consentement déterminé par une erreur substantielle sur la rentabilité de l'activité concernée[253] mais les circonstances étaient alors particulières car il est nécessaire que la rentabilité économique soit expressément dans le champ contractuel et constitue donc une caractéristique essentielle du contrat, déterminante du consentement[254]. L'équilibre contractuel ne semble guère pouvoir être plus pertinent, pour l'essentiel, dès lors que le cocontractant a conscience de la part d'aléa avant la signature, sauf – puisqu'il s'agit en général de contrats d'adhésion – à pouvoir mobilier le caractère abusif du contrat au sens de l'article 1171 du code civil. Il n'est toutefois pas évident de considérer que cette disposition suffise à régler le sort de l'exploitation d'une voix ou d'une image dès lors que, d'une part, cet article exclut expressément que le déséquilibre significatif puisse porter sur l'objet principal du contrat ou sur l'adéquation du prix à la prestation. Elle a néanmoins commencé à être mobilisée dans le cadre des contrats de services numériques, afin de conforter la protection des données à caractère personnel, mais la jurisprudence doit encore être confortée sur ce point[255].

Une telle voie ne permet toutefois que de faire face *a posteriori* à certaines situations abusives, question majeure pour les contrats antérieurs aux outils d'IAG. Pour ceux-ci, c'est dans le cadre du droit actuel que la réflexion doit porter : nonobstant les différences importantes existantes, à l'image de recommandations qui ont été formulées aux Etats-Unis pour s'appuyer sur le *right of publicity*[256], il semble également pertinent en droit français de **s'appuyer sur le caractère exclusif et absolu des droits de la personnalité, qui, demeurant essentiellement extra-patrimoniaux, ne devraient pas avoir pu faire l'objet de droits d'exploitation dépassant ce qui était raisonnablement connu à la date de la conclusion**. Cela est cohérent avec la place centrale qu'occupe la volonté individuelle dans les droits de la personnalité[257]. Tant sur le terrain du droit à la vie privée que des droits de la personnalité protégés par le même article 9 du code civil, la jurisprudence est au demeurant attentive à limiter la portée du consentement qui a pu être donné[258]. Toutefois, pour la mission, cette perspective devrait probablement être renforcée face à des outils aux potentialités qui ne peuvent être entièrement maîtrisées : à cet égard, la reconnaissance d'un consentement tacite[259] devrait être abordée avec prudence face à l'exploitation des

[251] Cas évoqué par Pierre Saint-Germier, « Clones, filtres, et fakes. L'éthique de la conversion vocale à l'ère de l'intelligence artificielle », *L'étincelle. Le journal de la création à l'Ircam*, 2024, n° 24, p. 46.

[252] On rappellera que selon la jurisprudence de la Cour de justice, il est admis qu'une nouvelle autorisation du titulaire des droits soit requise lorsque la communication au public s'effectue selon des modalités ou à destination de publics non prévus initialement (CJCE, 9 avril 1987, *Basset c/ SACEM*, aff. 402/85 ; CJUE, gr. ch., 22 juin 2021, *Youtube et Cyando*, aff. C-682/18 et C-683/18).

[253] Cass. Com., 4 octobre 2011, pourvoi n° 10-20.956, inédit ; Com., 12 juin 2012, pourvoi n° 11-19.047, inédit.

[254] Cass. 1re Civ., 21 octobre 2020, pourvoi n° 18-26.761, au Bull.

[255] Voy. à ce sujet Juliette Sénéchal, « Lutte contre les clauses abusives et protection des données à caractère personnel », *AJ contrat*, 2019, p. 412.

[256] Alice Preminger, Matthew Kluger, « The Right of Publicity Cans Save Actors from Deepfake Armageddon », *Berkeley Technology Law Journal*, 2024, vol. 39, p. 781.

[257] Agathe Lepage, V° « Droits de la personnalité », *Encyclopédie de droit civil*, Dalloz, juillet 2022, paragr. 193.

[258] Voy. not. Cass. 1re Civ., 15 janvier 2005, pourvoi n° 13-25.634, Bull. civ. I, n° 12 :

[259] Ce qu'a admis la jurisprudence au regard des circonstances : Cass. 2e Civ., 4 novembre 2004, pourvoi n° 02-15.120, Bull., 2004, II, n° 487.

éléments de la personnalité par une IAG, ce consentement ne pouvant être sans limite en l'absence d'une information appropriée.

Pour les contrats nouveaux d'exploitation des droits de la personnalité, **la mission estime nécessaire que**, dans le cadre de la liberté contractuelle qui prévaut aujourd'hui, **des bonnes pratiques soient développées et promues**, en particulier dans les secteurs sur lesquels les autorités publiques peuvent avoir une influence. La rédaction de contrats types, protégeant mieux les droits des personnes, apparaît pertinente et la mission considère que la démarche engagée en matière de documentaires, conduisant à la rédaction de clauses types à intégrer dans les contrats d'auteurs[260], devrait être encouragée dans les autres secteurs.

b) La contractualisation collective

Deux séries de grands accords ont été conclus depuis le développement des systèmes d'IA génératives aux les Etats-Unis, dans deux secteurs moteurs – le cinéma et la musique – mais selon des perspectives différentes : les uns sont des accords relevant de la négociation collective ; les autres sont des accords entre entreprises ayant des intérêts initialement divergents.

Ainsi, à la suite de la longue grève des scénaristes de Hollywood en 2023, un accord a été conclu avec les producteurs[261], prévoyant notamment des principes d'utilisation de l'IA, afin d'encadrer son utilisation pour la rédaction de scénarii, outre une augmentation des rémunérations minimales, tenant compte notamment des services d'abonnement et de l'audience sur ces services. Quelques semaines plus tard, en février 2024, un accord a été conclu entre l'*American Federation of Musicians* et *Alliance of Motion Picture and Television Producers*[262], tenant mieux compte du développement des services de vidéos à la demande dans la rémunération des musiciens, précisant qu'un musicien au sens de l'accord ne peut qu'être humain et encadrant les conditions de recours à l'IA et aux répliques numériques (avec la reconnaissance explicite comme fondement de l'accord de certains usages, à titre d'assistance ou de compléments, qui préexistaient).

Ce type d'accord résultant de la négociation collective permet d'apporter des garanties adaptées aux acteurs concernés. Il suppose un engagement de chacune des parties pour créer un cadre sécurisant l'intervention d'acteurs humains, sans nier l'intérêt des outils d'intelligence artificielle. D'autres initiatives participent de la création de ce cadre, comme le choix récent de l'Académie des Oscars d'exclure de ses prix les acteurs et scénarios générés par IA[263].

Le second ensemble d'accords concerne l'exploitation du catalogue des détenteurs de droits par les entreprises développement des systèmes d'IA générative. Tel a été, en particulier, le cas dans le domaine de la musique avec la conclusion d'accords par les majors Universal Music Group, Warner Music Group ou Sony avec certains acteurs de l'IA générative comme Udio, Suno ou Klay, permettant dans certains cas d'éteindre des actions judiciaires en cours aux Etats-Unis à l'issue incertaine dans le contexte juridique américain[264]. De même, dans le domaine de la presse, depuis 2023, des accords de licence ont été conclus par de nombreux acteurs[265].

Ces accords sont de nature à lutter contre certains risques liés à l'utilisation non maîtrisée de l'IA générative et constituent une avancée notable. Toutefois, ils soulèvent des inquiétudes pour les ayants-droit qui doivent en subir les conséquences sans y avoir été parties, en particulier les OGC et les artistes-interprètes[266]. La mission relève qu'il est délicat de porter une appréciation sur le contenu de ces accords,

---

[260] Accord conclu entre La Boucle documentaire, la GARRD, LaScam d'une part et le SATEV, le SPECT, le SPI, l'USPA d'autre part, rendu public le 17 novembre 2015.

[261] https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba

[262] https://www.afm.org/wp-content/uploads/2024/04/2024-02-24-MPTVF-MOU.pdf

[263] « Oscars : les acteurs et les scénarios générés par IA ne sont pas éligibles, annonce l'Académie », *Le Monde*, 1er mai 2026.

[264] Voy. par ex. Susan Wang, « Play It Again, HAL : Evaluating Fair Use in Generative Music Artificial Intelligence Training », *UCLA Entertainment Law Review*, vol. 32, n° 1, 2026, p. 97.

[265] On se reportera au suivi effectué par le Tow Center for Digital Journalism de la Columbia Journalism School : https://tow.cjr.org/ai-deals-lawsuits/

[266] Par ex., voy. le communiqué du syndicat français des artistes interprètes, de l'ADAMI, de la SPEDIDAM, de la SNAM-CGT et de FO-SN3M du 3 novembre 2025 : https://sfa-cgt.fr/presse/2240

qui sont largement confidentiels, mais elle estime que **si la logique d'accords d'ensemble, permettant de préserver les droits des artistes, en particulier quant à leur consentement pour que leurs créations soient utilisées par des systèmes d'IA générative, ne peut qu'être encouragée, ces accords doivent assurer, d'une part, une rémunération équitable de l'ensemble des acteurs concernés et, d'autre part, préserver les intérêts de l'ensemble des acteurs de la création, y compris les artistes indépendants**. A cet égard, il apparaît qu'une transparence sur les données essentielles de ces accords apparaît indispensable, car elle ne concerne pas que leurs signataires directs, comme il est indispensable, afin que les droits de l'ensemble des acteurs concernés soient protégés, que **les fournisseurs de systèmes d'IA générative soient responsabilisés au-delà de la conclusion d'accords ponctuels**, aussi significatifs soient-ils en termes de part de marché.

3.2.2. Le cadre législatif : bref tour d'horizon de quelques initiatives extérieures

Face aux enjeux soulevés par les hypertrucages pour les personnes en particulier, et notamment les acteurs des secteurs culturels et créatifs, plusieurs États ont engagé des réflexions sur des évolutions législatives – l'instrument contractuel montrant ses limites à régir de manière systémique la situation. Quatre séries de textes et initiatives peuvent en particulier être mentionnées, bien que toute comparaison soit nécessairement limitée par les différences de législations préexistantes.

Aux Etats-Unis, outre les nombreuses initiatives législatives engagées par plusieurs Etats[267], il est à signaler que des textes ont été initiés au Congrès, en particulier le « *No Fakes Act* » (pour « *Nurture Originals, Foster Art, and Keep Entertainment Safe Act* ») présenté au Sénat en avril 2025[268], et le « *No AI Fraud Act* » présenté à la Chambre des représentants en octobre 2024[269]. Parmi l'ensemble de ces initiatives, il n'est pas évident de déterminer si elles aboutiront, bien que la stratégie sur l'intelligence artificielle publiée par la Maison Blanche en mars 2026[270] appelle expressément le Congrès, afin de respecter les droits de propriété intellectuelle et soutenir les créateurs, à établir un cadre fédéral pour protéger les individus contre l'usage non autorisé de répliques numériques de leur personne. Le *No Fakes Act*, dans sa récente version révisée, est, dans ce cadre, le plus susceptible d'aboutir au niveau fédéral, inspiré en grande partie du *copyright*, ce qui n'est pas sans susciter des réserves[271].

À cet égard, doit être signalé le rapport de l'*US Copyright Office* consacré aux répliques numériques[272] : relevant que le droit américain ne permet pas de protéger suffisamment les droits des personnes face à des répliques numériques non consenties et préjudiciables, il propose une nouvelle législation visant ces répliques, qu'elles aient été générées ou non par un système d'IA, de nature à préserver le nom, l'image et l'apparence de toutes les personnes, indépendamment de leur notoriété, et couvrant une période au-delà du décès. Il propose également que la responsabilité des distributeurs soit engagée en cas de non-respect de ces droits, en complément de celle des créateurs, ce qui implique une responsabilité des fournisseurs de services en ligne, sauf à ce qu'ils réagissent promptement en cas de signalement. En revanche, le rapport exprime des réserves quant à l'idée de protéger à cette occasion le style de création d'une personne.

En complément de ces réflexions, il est d'ailleurs intéressant de relever les nombreuses réflexions menées par la doctrine américaine sur la réponse pouvant être apportée face au développement des hypertrucages, qui s'inscrivent entre deux limites, qui se retrouvent en droit européen bien qu'avec une intensité moindre : la liberté d'expression du premier amendement à la Constitution américaine d'une

---

[267] On trouvera un état des lieux récent de ces initiatives en annexe du rapport d'Anya Schiffrin *et al, Deepfake Financial Fraud. The Global Regulation of AI-Driven Scams*, Data & Society, Policy Brief, mars 2026; Jennifer Rothman, « Reframing Deepfakes », *Columbia Journal of Law & the Arts*, 2026, vol. 49, n° 4, p. 101. Voy. également Marylou Le Roy, « L'encadrement des systèmes d'intelligence artificielle par les Etats-Unis – Bref panorama et convergence avec l'Europe », *Communication Commerce électronique*, n° 6, juin 2024, étude 8.

[268] Une nouvelle version a été présentée en septembre 2025.

[269] Sur ces deux propositions, voy. Dustin Marlan, « Generative Identity Theft: Criminalizing Deepfakes Using the Right of Publicity », *Akron Law Review*, 2025, vol. 58, p. 445.

[270] https://www.whitehouse.gov/wp-content/uploads/2026/03/03.20.26-National-Policy-Framework-for-Artificial-Intelligence-Legislative-Recommendations.pdf

[271] Daphne Keller, « The NO FAKES Bill : A Terrible Speech Law for Our Times », *Standford Public La Working Paper*, août 2025 ; Jennifer Rothman, « Reintroduced No FAKES Act Still Needs Revision », *The Regulatory Review*, 18 août 2025 (en ligne).

[272] *Copyright and Artificial Intelligence. Part 1 : Digital Replicas. A Report of the Register of Copyrights*, juillet 2024.

part[273] ; la section 230 du *Communications Decency Act*, qui exonère de responsabilité les fournisseurs de services intermédiaires pour les contenus diffusés[274]. Entre ces deux bornes, les auteurs américains se sont interrogés sur les outils les plus pertinents pour répondre aux hypertrucages. La piste du *copyright* a été explorée, bien qu'elle se heurte à la difficulté de connaître le fonctionnement précis des systèmes d'IA générative ; elle est toutefois susceptible d'être peu efficace au regard de la capacité à invoquer le *fair use* du fait de ce mode de fonctionnement[275]. Néanmoins, contrairement au *right of publicity*, le droit de la propriété intellectuelle n'est pas contraint par les deux bornes précédemment mentionnées du premier amendement et de la section 230 du *CDA*, ce qui offre des perspectives élargies, pour autant qu'on admette la protection de la fixation des éléments de la personne, à défaut de la personne elle-même[276]. Ce droit du *copyright* soulève cependant aussi des difficultés tenant à sa nature, la patrimonialisation des éléments de la personnalité ne permettant pas la prise en compte des aspects liés à l'intimité ou la dignité et présentant un risque de perte de contrôle de la personne sur ses attributs protégés en cas de cession ; il a alors pu être proposé d'envisager un droit de propriété intellectuelle adapté sur les éléments de la personne humaine, avec un droit d'autodétermination permanent (seule la personne pouvant être, tout au long de sa vie, titulaire de ses droits sur son nom, son apparence et sa voix, les droits n'étant attribués que pour un usage limitatif), une obligation de recueil du consentement pour toute utilisation nouvelle et le développement d'un contrôle *post mortem* encadré pour éviter des dérives mercantiles[277].

En parallèle, d'autres auteurs ont estimé pouvoir trouver une solution dans le cadre du *right of publicity*, notamment face au risque d'inopérance du *copyright* en cas d'invocation de l'exception de *fair use*. Compte tenu des différences substantielles entre la capacité de préjudice des hypertrucages au regard du contexte initial de développement de ce droit, c'est au bénéfice d'une invitation à une évolution jurisprudentielle que cette piste a été défendue, assurant un droit constitutionnel à défendre les éléments de sa personnalité et sa réputation comme élément majeur de la protection de la personne[278].

Au Danemark, un projet de loi a été présenté par le gouvernement afin de compléter la loi sur le droit d'auteur afin de protéger les personnes contre les répliques numériques. Deux dispositions sont principalement prévues : l'une impose le recueil du consentement de l'artiste ou de l'interprète pour rendre accessible au public une imitation réaliste générées numériquement d'une performance ou d'une présentation artistique ; l'autre introduit une protection contre ces imitations des caractéristiques personnelles, au bénéfice de l'ensemble des personnes, cette protection étant maintenue jusqu'à 50 ans après le décès de la personne imitée.

Ce projet a suscité un intérêt certain au sein du continent européen (et même au-delà), bien que les options retenues aient été critiquées, faute de couvrir la création de tels contenus, de responsabiliser les fournisseurs de systèmes d'IA générative[279] ou de prendre en compte la question du marquage et de l'étiquetage et, à l'inverse, d'exclure très largement les hypertrucages constituant une satire, une caricature, une parodie ou une critique sociale[280]. De manière plus substantielle, il a été relevé que ce texte présentait le risque de diluer dans une protection comparable à celle du droit d'auteur des enjeux relevant des droits de la personnalité, dont la protection ne peut être seulement patrimoniale, compte tenu des enjeux liés à

[273] Voy. en particulier, Shannon Reid, « The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections », *Journal of Constitutional Law*, 2021, vol. 23, p. 209 et Robert Post, Jennifer Rothman, « The First Amendment and the Right(s) of Publicity », *The Yale Law Journal*, vol. 130, novembre 2020, p. 86.
[274] Voy. not. Peter W. Johnston, « Section 203's Characterization Problem », *Harvard Journal of Law & Technology*, 2025, vol. 38, n° 4, p. 1197.
[275] Neeraja Seshadri, « Implications of Deepfakes on Copyright Law », WIPO, juillet 2020 ; Scott Lu, « Deepfakes Under Copyright Law – A Necessary Legal Innovation », *Southern Illinois University Law Journal*, 2024, vol. 48, p. 517.
[276] Jennifer Rothman, « Copyrighting People », *Journal of Copyright Society*, 2025, vol. 72, n° 1.
[277] Sur ces aspects, voy. l'art. préc. de Jennifer Rothman, ainsi que Michael Goodyear, « Dignity and Deepfakes », *Arizona State Law Journal*, 2025, vol. 57, p. 931, et, sur le dernier aspect, Anita Allen, Jennifer Rothman, « Postmortem Privacy », *Michigan Law Review*, 2024, vol. 123, p. 185.
[278] Shannon Reid, « The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections », *Journal of Constitutional Law*, 2021, vol. 23, p. 209 ; Alice Preminger, Matthew Kluger, « The Right of Publicity Cans Save Actors from Deepfake Armageddon », *Berkeley Technology Law Journal*, 2024, vol. 39, p. 781.
[279] Jakob Plesner Mathiasen, Johannes Stilhoff, Astrid Christine Bay, « Deepfakes: Can you "copyright" a person? », *The IPKat*, 8 décembre 2025 (en ligne).
[280] Catherine Régis, Emma Kondrup, Clare Mulrooney, « Protéger les personnes à l'ère des deepfakes », *Mila*, 16 janvier 2026 (en ligne).

l'honneur, l'intimité et la dignité[281]. Néanmoins, ce texte met en lumière la nécessité de consacrer expressément des droits de la personne afin d'assurer l'effectivité, face aux hypertrucages, des dispositions du DSA permettant le retrait des contenus illicites ; à cet égard, en rappelant que ces droits préexistaient sans toujours relever d'un corpus normatif uni et clair, il met en lumière les limites du DSA pour répondre, en l'état des textes, à la diffusion de contenus hypertruqués[282] – y compris les contenus à caractère sexuel non consentis ou pédopornographique, que la révision en cours du RIA par la proposition de règlement omnibus entend mieux prendre en compte. La Commission européenne a d'ailleurs émis des réserves sur le projet de texte à l'occasion de la notification qui lui a été faite en application de la directive 2015/1535[283], indiquant notamment que, de son point de vue, le projet de loi allait au-delà de ce que permettait le cadre européen s'agissant des droits des artistes-interprètes et pourrait interférer avec le droit de reproduction des exécutions régi par le droit de l'Union. S'agissant de la protection générale des caractéristiques personnelles, la Commission exprime des réserves sur l'application d'un droit de propriété intellectuelle, celui-ci visant en principe à récompenser son titulaire et à garantir le maintien et le développement ultérieur de la créativité et de l'innovation.

Au Royaume-Uni, un récent rapport parlementaire[284] invite le Gouvernement à renforcer la protection contre les répliques numériques non consenties et contre les contenus générés par IA créant un préjudice en imitant le style d'un créateur ou artiste. L'objectif affiché est de donner aux créateurs et interprètes un contrôle clair sur l'exploitation commerciale de leur identité, dans un contexte législatif qui ne comporte pas de reconnaissance forte des droits de la personnalité – en dépit d'évolutions jurisprudentielles récentes en faveur de la reconnaissance d'actions possibles face au détournement d'informations privées, mais sans qu'un droit absolu à la voix, à la ressemblance ou à la réputation n'existe[285]. Par ailleurs, ce rapport invite à imposer la transparence sur les données d'entrainement de l'IA et à imposer un marquage visible des contenus ayant eu recours à ces outils.

En Chine, une législation de 2023 consacrée aux hypertrucages impose que les fournisseurs de systèmes d'IA générative marquent les contenus, à la fois de manière visible et par des mentions dans les métadonnées[286]. La démarche chinoise de régulation se poursuit actuellement, avec un récent projet en cours de consultation pour régir les systèmes d'IA anthropomorphes[287], *Interim Measures for the Administration of Humanized Interactive Services Based on AI*.

À la lumière des réflexions précédentes de droit comparé, notamment américain, et au regard de la multiplicité – et des limites – des protections aujourd'hui applicables, un certain nombre de pistes d'évolutions peuvent apparaître dans le cadre européen et français que la mission recommande d'approfondir. La réflexion sur d'éventuelles évolutions du cadre normatif supposera, sans renoncer à faire preuve d'ambition, d'avancer avec une certaine prudence sur un terrain dans lequel la place des droits fondamentaux, et en particulier de la liberté d'expression, est importante.

---

[281] Valérie-Laure Benabou, « Doublage des voix : les professionnels face à l'IA », *Le club des juristes*, 11 février 2026 (en ligne) ; Giovanni Fleck, « Denmark Leads EU Push to Copyright Face in Fight Against Deepfakes », *Techpolicy*, 7 octobre 2025 (en ligne).

[282] Voy. à cet égard Sofia Karttunen, « The Danish approach to copyright and deepfakes : A model for the EU ? », European Parliamentary Research Service, janvier 2026, doc. PE 782.611.

[283] Procédure RIS 2025/0654/DK. Les observations reprises ici sont issues des observations adressées aux autorités danoises le 2 février 2026.

[284] House of Lords, Communications and Digital Committee, *AI, copyright and the creative industries*, 4th Report of Session 2024-26, HL Paper 267, mars 2026.

[285] John Zerilli, « Appropriation of Personality in the Era of Deepfakes », *in* Ernest Lim, Phillip Morgan (éd.), *Private Law and Artificial Intelligence*, Cambridge University Press, 2024, p. 227.

[286] « China's New AI Regulations », Lantham & Watkins Privacy & Cyber Practice, 16 août 2023 (en ligne).

[287] « IA : la Chine encadre strictement les « compagnons virtuels » pour préserver la stabilité sociale », *La Tribune*, 29 décembre 2025 ; Luiza Jarovsky, « China's Approach to AI Anthropomorphism », *Luiza's Newsletter*, 13 janvier 2026.

4. MECANISMES COGNITIFS ET ECONOMIE DE LA CONFIANCE

Les hypertrucages brouillent l'économie de la confiance dans les contenus culturels et médiatiques qui circulent non pas tant par leur capacité à tromper directement le public, mais surtout par divers mécanismes cognitifs convergents dont l'effet cumulatif est susceptible de transformer durablement leur réception et la valeur perçue de ces contenus. Les hypertrucages opèrent une érosion graduelle des heuristiques perceptives, des cadres mémoriels, des relations interpersonnelles, des asymétries probatoires et de la confiance institutionnelle. Considérés ensemble, ces mécanismes constituent ce que Naffi qualifie de « crise du savoir » : une déstabilisation des mécanismes par lesquels les sociétés construisent une compréhension partagée du réel, et non simplement une crise de la désinformation ponctuelle[288].

La question alimente des spéculations. Lorsque le virtuel permet de reconstituer ou de se rapprocher du réel, comment pourrais-je encore accorder ma confiance au réel, et comment pourrais-je ne pas accorder trop de confiance au virtuel ? Pour le dire autrement, si je peux écouter un discours du Général de Gaulle qu'il a vraiment prononcé mais jamais enregistré, ou voir des images de la vie parisienne à la Belle Epoque avec des couleurs supposées fidèles, comment pourrais-je conserver ma confiance dans les images réelles des camps de concentration filmés en 1945, et comment pourrais-je ne pas accorder trop de crédit à un duo de chanteurs qui, en réalité, ne se sont jamais rencontrés ?

Dans le présent chapitre nous abordons tout d'abord les mécanismes de perception et de réception des hypertrucages ainsi que leur rôle dans la construction de la mémoire au niveau individuel (4.1). Par ailleurs, dans l'ensemble de la société, de manière structurelle la simple existence des hypertrucages produit une dégradation généralisée de la confiance, y compris à l'égard des contenus authentiques, par effet de contamination (4.2).

Une précaution méthodologique doit guider la lecture des résultats présentés ci-après ; la littérature sur les hypertrucages mélange deux registres qu'il importe de distinguer. D'un côté, des résultats empiriques solides issus d'expérimentations contrôlées, de méta-analyses et d'enquêtes représentatives : ils établissent des faits sur la perception, la mémoire, l'incertitude et la confiance. De l'autre, des inquiétudes spéculatives (archives synthétiques, falsifications de l'histoire culturelle, érosion de l'imaginaire collectif) qui restent à ce jour peu étayées empiriquement mais dont le caractère prospectif est légitime au regard du rythme d'amélioration technologique. Le chapitre signale systématiquement le passage entre les deux registres, car confondre l'établi et le prospectif appauvrirait la rigueur de l'argument et la qualité du débat de politique publique qu'il vise à nourrir.

4.1. Perception, réception et construction de la mémoire individuelle

Les limites cognitives des humains face aux contenus synthétiques brouillent leur perception. De plus, à la question de la perception se superpose celle de la réception. Enfin, un mécanisme cognitif majeur engagé par les hypertrucages est celui qui relève de la mémoire.

4.1.1. Brouillage perceptif : les limites cognitives face aux contenus synthétiques

Les données empiriques convergentes des dernières années confirment que la grande majorité des auditeurs et des spectateurs sont incapables de distinguer les contenus générés par intelligence artificielle de ceux produits par des humains, à l'oreille comme à l'œil nu. Le constat est particulièrement net dans le domaine musical et vocal. L'enquête à grande échelle Deezer/Ipsos (2025), conduite auprès de neuf mille répondants dans huit pays (États-Unis, Canada, Brésil, Royaume-Uni, France, Pays-Bas, Allemagne, Japon), révèle que dans un test d'écoute à l'aveugle portant sur trois pistes, 97 % des participants ont échoué à identifier correctement les morceaux générés par IA[289]. 71 % se déclarent surpris par ce résultat et 52 % expriment un malaise face à leur propre incapacité de discernement. Une enquête du cabinet Deloitte

[288] Nadia Naffi, 'Deepfakes and the Crisis of Knowing', UNESCO, 1 October 2025, https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing.
[289] Deezer, 'Deezer and Ipsos Study: AI Fools 97% of Listeners', *Deezer Newsroom*, 12 November 2025, https://newsroom-deezer.com/2025/11/deezer-ipsos-survey-ai-music/.

corrobore cette tendance avec un constat plus large : 59 % des consommateurs déclarent avoir du mal à distinguer contenu humain et contenu IA, tous formats confondus[290].

Au-delà des enquêtes déclaratives, des études expérimentales viennent préciser la mécanique de cet échec perceptif. Une étude menée sur 57 participants exposés à 12 *extraits audio*, établit que les morceaux co-créés humain-IA sont les plus difficiles à identifier, révélant un brouillage, non pas binaire mais gradué, des frontières perceptives entre production humaine et artificielle[291]. Les différences entre experts (musiciens) et non-experts (non-musiciens) portent moins sur la performance d'identification (modeste dans les deux groupes) que sur la confiance subjective, les musiciens étant significativement plus assurés de leur réponse, sans que cette assurance ne se traduise par une meilleure précision. Ce décrochage entre confiance et performance constitue une vulnérabilité cognitive structurelle qui réapparaît dans toutes les modalités sensorielles étudiées.

Sur le terrain de la vidéo, une étude développe une comparaison directe entre la performance humaine et celle de modèles de détection automatisés[292]. Cent dix participants (55 anglophones natifs et 45 non natifs) ont visionné quarante vidéos issues du jeu de données FakeAVCeleb (20 authentiques et 20 hypertrucages), accédées par une plateforme web ludique. La précision humaine s'établit autour de 65 %, légèrement au-dessus du hasard, alors que cinq modèles d'IA atteignent près de 98 % sur les mêmes stimuli. Les participants surestiment leurs compétences de détection et s'appuient principalement sur des indices visuels, négligeant les indices auditifs. L'entraînement améliore la performance, mais l'écart entre confiance subjective et précision réelle persiste, confirmant la vulnérabilité cognitive structurelle déjà observée dans le domaine musical. La faible performance humaine à la détection d'hypertrucages vidéo est corroborée par une autre étude[293], portant sur 54 participants et 80 stimuli vocaux en espagnol et en japonais ; on observe une précision moyenne de détection d'environ 59 %. Les jugements reposent surtout sur l'intonation, le rythme, les pauses, la respiration et la fluidité du discours ; ils révèlent par ailleurs des différences interculturelles dans la notion de « naturalité » vocale, ce qui complique l'élaboration de critères universels de détection.

Le développement récent du benchmark DigiFakeAV[294] apporte un éclairage empirique sur les générations les plus récentes de modèles. Ce corpus multimodal d'environ 70 000 vidéos haute résolution générées par un modèle de diffusion sert à tester la reconnaissance humaine de contenus synthétiques. Le taux de non-reconnaissance chez des experts familiers de la génération de vidéo par IA atteint 68 %, ce qui suggère que les hypertrucages fondés sur les modèles de diffusion deviennent difficilement discernables même pour un public averti.

La détection des *voix clonées* présente un tableau similaire : des chercheurs ont montré sur plusieurs centaines de participants que la performance humaine se situe à peine au-dessus du hasard (précision d'environ 60 %) et que les auditeurs confondent des paires de voix réelle et clonée dans près de 80 % des cas[295]. Un Professeur de l'université de Laval qualifie dans un article publié sur le site de l'UNSECO ce point d'inflexion de « *synthetic reality threshold* » : un seuil au-delà duquel les humains ne peuvent plus distinguer authentique et synthétique sans assistance technique externe[296].

Le brouillage perceptif est donc désormais robustement établi : la grande majorité des publics échoue à distinguer les contenus synthétiques des contenus humains dans les conditions ordinaires de réception, et cette incapacité s'aggrave avec les modèles de diffusion les plus récents. La conséquence pratique pour l'industrie culturelle est que l'authenticité ne peut plus être garantie par l'œil

[290] 'Earning Trust as Gen AI Takes Hold: 2024 Connected Consumer Survey', Deloitte Insights, accessed 2 April 2026, https://www.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey/2024.html.

[291] Neil J. Rigole et al., 'Perceptions of AI-Generated and Co-Created Music among University Students and Social Media Users.', *Issues in Information Systems* 26, no. 4 (2025): 102–17.

[292] Ammarah Hashmi et al., 'Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes', arXiv:2405.04097, preprint, arXiv, 11 November 2024, https://doi.org/10.48550/arXiv.2405.04097.

[293] Eugenia San Segundo et al., 'Human Perception of Audio Deepfakes: The Role of Language and Speaking Style', arXiv.Org, 10 December 2025, https://arxiv.org/abs/2512.09221v1.

[294] Jiaxin Liu et al., 'Beyond Face Swapping: A Diffusion-Based Digital Human Benchmark for Multimodal Deepfake Detection', arXiv:2505.16512, preprint, arXiv, 28 January 2026, https://doi.org/10.48550/arXiv.2505.16512.

[295] Sarah Barrington et al., 'People Are Poorly Equipped to Detect AI-Powered Voice Clones', *Scientific Reports* 15, no. 1 (2025): 11004, https://doi.org/10.1038/s41598-025-94170-3.

[296] Naffi, 'Deepfakes and the Crisis of Knowing'.

ou l'oreille du spectateur : elle doit être construite institutionnellement, notamment par des labels, des certifications et des chaînes de provenance vérifiables.

4.1.2. Réception : entre demande de transparence et acceptation

L'enquête Deezer/Ipsos 2025[297] fournit la base la plus large pour cartographier les attentes du public dans le secteur musical. Sur les 9000 répondants, 80 % souhaitent un étiquetage clair des contenus entièrement générés par IA, 73 % veulent savoir si leur plateforme leur recommande de la musique synthétique. 51% craignent une production musicale de basse qualité et générique, et seulement 19 % font confiance aux capacités actuelles de l'IA musicale. 52% estiment que les pistes 100 % IA ne devraient pas figurer dans les classements principaux, contre 11 % seulement favorables à un traitement égal. Ces chiffres dessinent une demande normative claire : transparence, hiérarchisation, protection des œuvres humaines.

Une étude à grande échelle[298] sur 31 363 commentaires YouTube publiés sous des contenus musicaux générés ou manipulés par IA apporte un contrepoint important. Les sentiments exprimés sont majoritairement positifs ou neutres, signe d'une acceptation sociale réelle des productions musicales synthétiques dans les usages ordinaires. Les réactions négatives se concentrent sur les hypertrucages imitant des voix célèbres, qui activent des attentes d'authenticité et des normes morales liées à l'identité artistique. Cette étude met ainsi en évidence un décalage entre les inquiétudes normatives déclarées dans les enquêtes d'opinion et les formes effectives d'appropriation par les publics. Une étude examinant les déterminants de l'acceptation du chant généré par IA auprès de 310 participants, identifie les caractéristiques anthropomorphiques des voix synthétiques (animéité, ressemblance humaine) comme principaux prédicteurs de l'agréabilité, qui à son tour prédit l'intention d'usage[299]. La curiosité et le bouche-à-oreille apparaissent comme des moteurs importants de diffusion.

L'effet de l'étiquetage est, quant à lui, plus complexe que l'intuition ne le laisse penser. Les travaux récents convergent pour montrer que la révélation du caractère IA d'un contenu (« label IA ») n'est ni uniformément négative ni systématique. Certains auteurs[300], par une manipulation causale du label sur des morceaux identiques générés par IA, montrent que les morceaux présentés comme issus d'une IA suscitent des émotions plus positives (intérêt, énergie, émerveillement) — sans différence significative en termes de qualité perçue ou de préférence globale. À l'inverse, d'autres chercheurs[301] mettent en évidence un biais d'authenticité : les œuvres attribuées à l'IA sont alors jugées moins profondes émotionnellement ou moins authentiques, reflétant des attentes normatives associées à l'expression humaine. Le label opère ainsi comme un signal interprétatif qui peut moduler certaines dimensions du jugement (authenticité, valence émotionnelle) tout en laissant l'évaluation globale inchangée lorsque les propriétés perceptives sont saillantes.

Une étude expérimentale[302] prolonge ce raisonnement dans le champ politique. Sur deux expérimentations regroupant 2 808 participants américains exposés à des hypertrucages parodiques de politiciens prenant des positions opposées sur le changement climatique, l'inclusion d'une mention signalant l'usage de l'IA modifie significativement la compréhension de l'audience et sa capacité à reconnaître la parodie. Ces deux facteurs influencent ensuite l'enjouement, le rejet et la contre-argumentation, lesquels à leur tour structurent le soutien aux politiques publiques et l'intention de partage des hypertrucages.

L'étiquetage n'est donc pas un acte neutre d'information : il reconfigure l'interprétation et la mobilisation cognitive du contenu. Cela est d'autant plus important lorsque l'on relève que les hypertrucages immersifs

---

[297] Deezer, 'Deezer and Ipsos Study'.
[298] Francisco Tigre Moura and Visieu Lac, 'Deepfake Music and Listener Sentiments: A Large-Scale Analysis of YouTube Comments', 6 July 2025.
[299] Mikael Bagratuni et al., 'Innovation in Tune: An Empirical Investigation of User Acceptance of Artificial Intelligence-Generated Music', *Computers in Human Behavior Reports* 18 (May 2025): 100660, https://doi.org/10.1016/j.chbr.2025.100660.
[300] Suqi Chia et al., 'Do Listeners Devalue AI-Generated Pop Music? Exploring Negative Biases in Listeners' Responses to AI-Labelled vs Human-Labelled Pop Music', *Computers in Human Behavior: Artificial Humans* 6 (December 2025): 100217, https://doi.org/10.1016/j.chbah.2025.100217.
[301] Rigole et al., 'Perceptions of AI-Generated and Co-Created Music among University Students and Social Media Users.'
[302] Hang Lu and Shupei Yuan, '"I Know It'sa Deepfake": The Role of AI Disclaimers and Comprehension in the Processing of Deepfake Parodies', *Journal of Communication* 74, no. 5 (2024): 359–73.

(et sans label) sont plus partagés. C'est ce que suggère par exemple une étude[303] sur 1 031 participants américains exposés à dix hypertrucages répartis en cinq genres (célébrité, divertissement, publicité, politique, cheapfake). Les auteurs de l'étude trouvent que l'immersion dans le récit prédit positivement l'évaluation perceptive, l'enjouement et l'intention de partage, indépendamment du genre. **Les hypertrucages les mieux conçus narrativement sont aussi les plus susceptibles d'être partagés, indépendamment du fait que l'on sait qu'il s'agit d'une création artificielle.** Une étude (Oliullah et Murtuza (2025) révèle que le partage d'hypertrucages repose principalement sur l'alignement partisan et la signalisation identitaire plutôt que sur le jugement de crédibilité, ce qui invalide partiellement les stratégies de modération fondées sur l'étiquetage seul.

Ces travaux dessinent une **tension structurelle entre normes déclarées et usages effectifs**. **D'un côté**, **les enquêtes d'opinion révèlent des attentes fortes en matière de transparence, et une méfiance persistante vis-à-vis de la qualité et de la légitimité des productions entièrement synthétiques**. **De l'autre**, **les comportements observés en contexte et certaines études expérimentales suggèrent une acceptation pragmatique, voire une valorisation conditionnelle, notamment lorsque les contenus suscitent la curiosité ou activent le transport narratif.** Comme le résume le journaliste Guillaume Fraissard (2026)[304], « *savoir qu'une œuvre est fausse change la nature de l'émotion qu'elle nous procure* », mais ne supprime pas systématiquement cette émotion. Ces résultats invitent à concevoir des dispositifs d'information qui précisent le degré d'intervention de l'IA (assistée, co-créée, entièrement générée), tout en reconnaissant que **l'étiquetage ne suffit pas à garantir une réception critique.**

Une méta-analyse[305] portant sur 24 études expérimentales et 20 685 participants issus de dix pays, montre empiriquement qu'**en moyenne, l'exposition aux hypertrucages n'entraîne ni baisse robuste de la crédibilité perçue ni diminution de l'intention de partage. Cette absence d'effet global masque en réalité des effets opposés selon les sous-groupes** ; **l'étude met en avant le rôle modérateur de l'éducation aux médias et à l'information** ; les individus formés détectent davantage les hypertrucages, les jugent moins crédibles et sont moins enclins à les diffuser, tandis que ces effets disparaissent, voire s'inversent, en l'absence de formation. Parallèlement, les hypertrucages produisent de forts effets émotionnels positifs, notamment dans des contextes prosociaux (éducation, santé), où ils peuvent accroître la compassion ou l'engagement, effet encore plus marqué sans éducation au média et à l'information.

Une enquête récente sur 7 083 répondants dans sept pays européens (Royaume-Uni, Allemagne, Pays-Bas, France, Italie, Suède, Tchéquie) ajoute une dimension comparative[306]. En moyenne, les jeunes cohortes apparaissent moins alarmées dans seulement trois pays (Suède, France, Tchéquie), tandis que les publics néerlandais, allemand, britannique et italien combinent enthousiasme pour les usages prosociaux (simulations de crise, campagnes d'intérêt public, avatars thérapeutiques) et vigilance face aux risques. Le Royaume-Uni se distingue par la coexistence d'un fort niveau d'adoption et d'un fort niveau de vigilance.

4.1.3. La mécanique des faux souvenirs

Les recherches sur la malléabilité de la mémoire[307], jusque-là appuyées sur des photographies truquées ou des descriptions textuelles, peuvent désormais être prolongées par des stimuli audiovisuels d'un réalisme inédit, ce qui démultiplie le potentiel d'implantation de faux souvenirs. Le mécanisme cognitif sous-jacent est celui de la « parcimonie cognitive » : nous favorisons spontanément la récupération de souvenirs cohérents avec notre vision du monde, et les hypertrucages exploitent ce biais en fabriquant des contenus qui s'inscrivent dans nos attentes préexistantes. Comme un cheval de Troie, ces

[303] María T. Soto-Sanfiel et al., 'Deepfakes as Narratives: Psychological Processes Explaining Their Reception', *Computers in Human Behavior* 165 (2025): 108518.

[304] Guillaume Fraissard, 'Reprise de « Papaoutai » Avec l'IA : « Savoir Qu'une Œuvre Est Fausse Change La Nature de l'émotion Qu'elle Nous Procure »', *Le Monde*, 30 January 2026, https://www.lemonde.fr/idees/article/2026/01/30/reprise-de-papaoutai-avec-l-ia-savoir-qu-une-uvre-est-fausse-change-la-nature-de-l-emotion-qu-elle-nous-procure_6664754_3232.html.

[305] Seok Kang and Kayla Valadez, 'Deepfakes' Cognitive, Emotional, and Behavioral Impact: A Systematic Review and Meta-Analysis of Individual Responses', *Journalism & Mass Communication Quarterly* 102, no. 4 (2025): 958–92, https://doi.org/10.1177/10776990251357294.

[306] Nik Hynek et al., 'Risks and Benefits of Artificial Intelligence Deepfakes: Systematic Review and Comparison of Public Attitudes in Seven European Countries', *Journal of Innovation & Knowledge* 10, no. 5 (2025): 100782, https://doi.org/10.1016/j.jik.2025.100782.

[307] Nadine Liv and Dov Greenbaum, 'Deep Fakes and Memory Malleability: False Memories in the Service of Fake News', *AJOB Neuroscience* 11, no. 2 (2020): 96–104, https://doi.org/10.1080/21507740.2020.1740351.

faux souvenirs peuvent ensuite être réactivés à des moments stratégiques pour confirmer des préjugés ou orienter un comportement.

Dans une étude empirique[308], des chercheurs exposent 436 participants à des vidéos d'hypertrucages présentant de fausses suites ou remakes de films connus : Will Smith en Neo dans The Matrix, Charlize Theron dans Indiana Jones, Brad Pitt et Angelina Jolie dans The Shining, Chris Pratt dans Captain Marvel. Le taux moyen de faux souvenirs atteint 49 %, beaucoup de participants se souvenant même du « *remake* » comme étant meilleur que l'original. Le résultat contre-intuitif est le suivant : les hypertrucages audiovisuels ne sont pas plus efficaces que de simples descriptions textuelles pour induire ces faux souvenirs. Cette équivalence, retrouvée également par la même équipe puis confirmée par d'autres[309] montre que **c'est l'exposition à un récit cohérent qui crée le faux souvenir, et le format vidéo n'amplifie pas significativement cet effet par rapport à un format texte. Les hypertrucages audiovisuels sont une nouvelle variation, pas une rupture de nature avec d'autres mécanismes de manipulation de la mémoire.**

Une revue synthétisant la littérature académique formalise ce constat[310]. Cinq bases de données interrogées, 2 004 articles initialement identifiés, et 22 études empiriques ont été retenues par les auteurs. Ils concluent que les inquiétudes concernant la puissance unique de persuasion des hypertrucages sont rarement étayées par des données. La plupart des effets attendus (distorsion de la mémoire, altération de la crédibilité, modification du comportement) apparaissent non conclusifs, modérés, ou non significativement différents de ceux produits par d'autres formats (texte, photo truquée).

L'influence du texte est corroborée par une autre étude[311] visant à déterminer si des **récits historiques sont mémorisés et jugés vrais différemment lorsqu'ils sont présentés par une figure historique ressuscitée en hypertrucage vidéo ou sous forme de texte écrit**. Dans cette étude, 332 participants américains ont lu ou visionné des récits autobiographiques fictifs d'Albert Einstein ou Marie Curie, présentés en deux sessions séparées d'une semaine. Les participants étaient assignés aléatoirement à lire ou regarder le récit. Les versions texte et vidéo sont basés sur le même texte, et les récits sont fondés sur des informations biographiques vérifiées, même s'ils étaient formulés comme des récits fictionnalisés à la première personne. Les récits écrits produisent un rappel mnésique plus élevé et une vérité perçue plus forte que les vidéos, à la fois immédiatement et une semaine plus tard. Le « Need for Cognition », trait individuel mesurant l'engagement analytique du sujet, prédit positivement la vérité perçue dans toutes les conditions : ce sont les sujets cognitivement les plus engagés qui acceptent davantage la vraisemblance des récits. Bien que la vérité perçue décline au fil du temps, cette décroissance est plus rapide pour les vidéos que pour les textes, ce qui suggère que l'effet immédiat des hypertrucages vidéos peut être moins durable que le récit purement textuel.

Ainsi, les hypertrucages audiovisuels peuvent induire des faux souvenirs mais avec une efficacité non supérieure à celle de simples descriptions textuelles. Cette nuance déplace l'enjeu, de la spécificité technologique vers la spécificité narrative.

4.2. Contamination et dégradation généralisée de la confiance

En matière de confiance, la contamination est probablement l'effet le plus structurel pour l'industrie culturelle. Divers mécanismes se cumulent pour produire une dégradation généralisée de la valeur perçue des contenus.

4.2.1. Vérité illusoire et biais de l'imposteur

Le danger principal des hypertrucages ne serait pas de tromper massivement les individus, mais de produire de l'incertitude et, par ce biais, d'éroder la confiance générale du public. La dégradation de la confiance et la transformation durable des heuristiques de crédibilité a été théorisée sous le nom

[308] Gillian Murphy and Emma Flynn, 'Deepfake False Memories', *Memory* 30, no. 4 (2022): 480–92, https://doi.org/10.1080/09658211.2021.1919715.
[309] Murphy and Flynn, 'Deepfake False Memories'.
[310] Didier Ching et al., 'Can Deepfakes Manipulate Us? Assessing the Evidence via a Critical Scoping Review', *PLOS ONE* 20, no. 5 (2025): e0320124, https://doi.org/10.1371/journal.pone.0320124.
[311] María T. Soto-Sanfiel and Gina Junhan Fu, 'Deepfaking the Past: Memory and Perceived Truth of Resurrected Historical Figures', *Computers in Human Behavior*, 2026, 109008.

d' « *Impostor Bias* » ou biais de l'imposteur[312]. **L'enjeu cognitif central n'est pas tant le contenu particulier de chaque hypertrucage que l'effet systémique d'incertitude que sa simple existence introduit dans l'écosystème informationnel.** Cette intuition se trouve confirmée et précisée par diverses études.

La conclusion centrale d'un article détaillé en encadré est que les hypertrucages ne doivent pas être compris comme des outils de persuasion ou de mensonge direct, mais comme des technologies capables de produire une indétermination généralisée ; même quand les gens ne croient pas explicitement au contenu synthétique, ils peuvent devenir moins sûrs de ce qui est vrai, ce qui augmente les coûts cognitifs de vérification et fragilise la confiance dans l'écosystème informationnel.

Encadré : l'exposition à un hypertrucage crée une incertitude généralisée

Des chercheurs se sont intéressés à un hypertrucage politique produit par BuzzFeed, dans lequel on « voit » Barack Obama s'adresser à la caméra et insulter Donald Trump, alors qu'il s'agit en réalité d'une performance d'un humoriste (Jordan Peele) : son imitation vocale est combinée à un visage synthétique d'Obama[313]. Le but de l'étude est de comparer les effets de versions trompeuses de cette vidéo, où la révélation finale est supprimée, avec la version éducative complète, où l'on découvre que la vidéo est truquée. L'expérience est menée en ligne auprès d'un large échantillon britannique représentatif de 2 005 participants, assignés aléatoirement à trois conditions : un clip trompeur très court de 4 secondes, où le faux Obama insulte Trump ; un clip trompeur plus long de 26 secondes, qui contient un peu plus de contexte mais pas la révélation que la vidéo est trompeuse ; et la vidéo complète de 1 minute 10 secondes, qui révèle le trucage et explique le danger des hypertrucages.

Après exposition, les chercheurs demandent aux participants si Barack Obama a déjà insulté Donald Trump comme dans la vidéo. Une réponse "oui" est interprétée comme de la tromperie, une réponse "je ne sais pas" comme de l'incertitude, et une réponse "non" comme une absence de tromperie. Les auteurs mesurent aussi la confiance des participants dans les informations politiques vues sur les réseaux sociaux, afin de tester si l'incertitude produite par l'hypertrucage réduit cette confiance. Les résultats montrent d'abord que les versions trompeuses de l'hypertrucage ne trompent pas significativement plus que la version éducative : environ 14,9 % des participants sont trompés dans la condition 4 secondes, 16,4 % dans la condition 26 secondes, et 16,9 % dans la condition vidéo complète avec révélation. **L'hypothèse selon laquelle l'hypertrucage trompeur augmenterait directement la croyance fausse est donc rejetée. En revanche, les versions trompeuses augmentent nettement l'incertitude** : 35,1 % des participants répondent "je ne sais pas" après le clip trompeur de 4 secondes, 36,9 % après le clip trompeur de 26 secondes, contre 27,5 % après la vidéo complète avec révélation. Les analyses statistiques confirment que cette hausse de l'incertitude est significative. Enfin, les auteurs montrent que cette incertitude a un effet politique important : elle réduit indirectement la confiance dans les informations vues sur les réseaux sociaux. Autrement dit, **l'hypertrucage ne diminue pas directement la confiance ; il produit d'abord une incapacité à trancher entre vrai et faux, et c'est cette incertitude qui affaiblit ensuite la confiance.**

Une étude transnationale[314] dans huit pays montre que l'exposition antérieure à des hypertucages augmente la croyance dans la désinformation lorsque celle-ci est rencontrée ultérieurement, indépendamment de la capacité cognitive du sujet et ce, de manière transversale aux pays étudiés. Les consommateurs d'actualités sur les réseaux sociaux apparaissent particulièrement vulnérables. Ce résultat établit un mode d'action des hypertrucages qui ne passe pas par la persuasion immédiate (souvent absente ou modeste, comme l'a montré la revue de littérature présentée ci-dessus[315] mais par un effet d'imprégnation à long terme ; **c'est la familiarité accumulée avec des contenus synthétiques qui rend**

[312] Mirko Casu et al., 'GenAI Mirage: The Impostor Bias and the Deepfake Detection Challenge in the Era of Artificial Illusions', *Forensic Science International: Digital Investigation* 50 (September 2024): 301795, https://doi.org/10.1016/j.fsidi.2024.301795.

[313] Cristian Vaccari and Andrew Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society* 6, no. 1 (2020): 2056305120903408, https://doi.org/10.1177/2056305120903408.

[314] Saifuddin Ahmed et al., 'Social Media News Use Amplifies the Illusory Truth Effects of Viral Deepfakes: A Cross-National Study of Eight Countries', *Journal of Broadcasting & Electronic Media* 68, no. 5 (2024): 778–805, https://doi.org/10.1080/08838151.2024.2410783.

[315] Ching et al., 'Can Deepfakes Manipulate Us?'

le sujet ultérieurement perméable à la désinformation, et non pas la persuasion ponctuelle de tel ou tel hypertrucage.

Des auteurs ont étudié la « contamination » de la perte de confiance lors de la révélation de la nature synthétique d'un contenu[316]. Cette étude porte sur 951 participants exposés à des hypertrucages dans trois formats (audio, vidéo, vidéo immersive) avec mesure avant et après révélation du caractère réel ou IA du contenu. Lorsque le sujet apprend que ce qu'il a vu était truqué, la crédibilité accordée aux médias en général diminue de manière généralisée, y compris pour des contenus authentiques visionnés dans la même session. Le sentiment subjectif de discernement, lui aussi, baisse durablement : les participants se déclarent moins capables qu'avant de distinguer le vrai du faux, même quand ils n'ont pas été eux-mêmes trompés. Cette contamination affecte donc à la fois le jugement externe (sur la fiabilité des médias) et le jugement métacognitif (sur sa propre capacité d'évaluation), et elle suggère que la simple existence reconnue des hypertrucages reconfigure durablement la relation aux preuves audiovisuelles.

Cette transformation correspond au biais de l'imposteur[317]. **La simple existence des hypertrucages modifie les heuristiques de crédibilité que les sujets mobilisent face aux contenus synthétiques, instaurant un scepticisme généralisé y compris à l'égard des contenus humains.** Ce biais n'est pas une erreur ponctuelle de jugement mais une transformation structurelle de la confiance informationnelle qui affecte l'évaluation des preuves, la réception des messages et les dynamiques de persuasion. La menace centrale des hypertrucages réside ainsi moins dans leur capacité à nous faire croire au virtuel que dans leur capacité à nous faire douter du réel, basculement épistémique majeur. Le biais de l'imposteur dégrade donc la valeur perçue de l'ensemble des contenus audiovisuels (films, documentaires, reportages, captations de concerts) indépendamment de leur authenticité.

4.2.2. Le dividende du menteur

Là où les craintes initiales portaient sur la capacité des hypertrucages à faire passer le faux pour du vrai, le synthétique pour l'authentique, la littérature empirique récente documente un risque symétrique, largement aussi important : la capacité offerte à des acteurs économiques, politiques ou culturels de qualifier le vrai de faux pour échapper aux conséquences d'un comportement réel.

Des chercheurs ont nommé ce mécanisme « *liar's dividend* », ou dividende du menteur[318]. Ce mécanisme s'apparente à une stratégie de « *brouillage cognitif* » dont l'objectif n'est pas la persuasion mais la paralysie épistémique comme utilisé dans le cas des guerres pour décrédibiliser les informations fournies par la partie adverse[319].

La confirmation empirique la plus complète à ce jour porte sur la sphère politique[320]. Cinq expériences de sondage administrées à plus de 15 000 adultes américains testent deux stratégies de déni : invoquer l'incertitude informationnelle (« cette vidéo est probablement un hypertrucage ») ou mobiliser une rhétorique de ralliement partisan (« mes adversaires fabriquent des fake news contre moi »). Les deux stratégies augmentent le soutien aux politiciens accusés de scandale, et ce à travers tous les sous-groupes partisans. De plus, ces dénis produisent un soutien supérieur par rapport à des réponses alternatives au scandale (silence, excuses), ce qui en fait une stratégie économiquement efficace pour un acteur rationnel.

---

[316] Teresa Weikmann et al., 'After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media', *The International Journal of Press/Politics* 30, no. 1 (2025): 187–210, https://doi.org/10.1177/19401612241233539.

[317] Casu et al., 'GenAI Mirage'.

[318] Chesney and Citron, 'Deep Fakes'.

[319] John Twomey et al., 'Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets That Discuss Deepfakes in the Russian Invasion of Ukraine', *PLOS ONE* 18, no. 10 (2023): e0291668, https://doi.org/10.1371/journal.pone.0291668; 'Information Manipulation in the Age of Generative Artificial Intelligence | Think Tank | European Parliament', accessed 5 June 2026, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)779259.

[320] Kaylyn Jackson Schiff et al., 'The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?', *American Political Science Review* 119, no. 1 (2025): 71–90, https://doi.org/10.1017/S0003055423001454.

Une revue systématique des effets psychologiques et comportementaux des hypertrucages politiques, fondé sur 25 études empiriques sélectionnées selon les lignes directrices PRISMA 2020, confirme cette analyse[321]. Les hypertucages politiques ne produisent pas d'effet persuasif intrinsèquement supérieur à celui d'autres formes de désinformation, mais ils génèrent des effets distinctifs par « déstabilisation épistémique » : la confiance se trouve érodée non pas parce que les sujets ont été persuadés, mais parce que l'écosystème informationnel devient peu fiable. Les auteurs soulignent que les avertissements officiels, pourtant intuitifs comme remède, peuvent produire des effets paradoxaux, en accroissant la défiance généralisée plutôt qu'en restaurant la discrimination entre synthétique et authentique.

Transposé au secteur culturel, ce mécanisme suggère que le dividende du menteur ne produit pas nécessairement une défiance globale envers tous les contenus culturels ; il peut plutôt introduire une fragilisation locale, stratégique et sélective de l'authenticité. Autrement dit, **le public ne cesse pas forcément de croire aux documentaires, aux archives, aux captations ou aux témoignages audiovisuels en général. En revanche, dès qu'un contenu devient conflictuel, par exemple parce qu'il menace une réputation, engage une responsabilité juridique, affecte la valeur commerciale d'une œuvre ou contredit un récit institutionnel, la possibilité de le qualifier de "fake" devient une ressource rhétorique disponible.** Le risque n'est pas seulement l'effondrement généralisé de la confiance, mais l'apparition d'un doute opportuniste qui se déclenche dans les situations où l'authenticité a des conséquences économiques, symboliques ou morales.

4.2.3. Le documentaire, cas emblématique de confusion entre réel et virtuel

Les documentaires occupent un statut épistémique particulier dans la réception audiovisuelle. **Le pacte de lecture qui lie le spectateur au documentaire diffère fondamentalement de celui qui le lie à une fiction.** Dans la fiction, le spectateur accepte par convention que tout soit inventé (décors, costumes, dialogues, jeu des acteurs) au profit d'une histoire pour laquelle il « *suspend volontairement son incrédulité* ». Dans le documentaire, à l'inverse, il s'attend à ce que les images témoignent du réel, et c'est précisément cette attente d'indexicalité qui fait la force persuasive du genre. L'introduction de contenus synthétiques dans ce cadre rompt donc avec un dispositif de confiance spécifique, et non avec une simple convention de représentation.

Un travail analyse la manière dont les documentaires et docudramas façonnent les croyances perceptives du public, montrant que l'usage d'images réelles, de sons authentiques et de codes narratifs associés à la factualité renforce la confiance dans le contenu même lorsque la représentation reste en partie construite et interprétative[322]. Le pouvoir persuasif du documentaire ne tiendrait donc pas seulement à la qualité indexicale de ses images mais à un ensemble plus large de signaux conventionnels (voix hors champ, structure argumentative, mention de sources, traitement visuel sobre) qui activeraient chez le spectateur un mode de réception spécifique. L'introduction de contenus synthétiques pourrait affecter chacun de ces signaux séparément ou simultanément, ce qui rend la question empirique de leur effet sur la croyance particulièrement complexe à isoler.

Les effets comparés des fictions et des documentaires sont analysés empiriquement dans des contextes thématiques précis. Des chercheurs ont mesuré, sur un design avant/après exposition, l'impact d'un film de fiction ou d'un documentaire portant sur la question environnementale ; les modifications observées de croyances, de perceptions et de motivations comportementales confirment que le documentaire agit comme un vecteur de transformation des représentations sociales[323]. Un autre travail portant sur la communication scientifique a établi que le statut de la source dans un documentaire (scientifique, politique ou anonyme) influence directement les intentions comportementales du public, notamment la volonté de s'informer ou de modifier ses comportements. Les sources incarnant une autorité légitime amplifient ces

[321] Md Oliullah and H. M. Murtuza, 'Exploring How Political Deepfakes Trigger Cognitive Processes and Downstream Psychological and Behavioral Outcomes: A Systematic Review', *Computers in Human Behavior Reports* 22 (May 2026): 101066, https://doi.org/10.1016/j.chbr.2026.101066.

[322] Enrico Terrone, 'Documentaries, Docudramas, and Perceptual Beliefs', *The Journal of Aesthetics and Art Criticism* 78, no. 1 (2020): 43–56, https://doi.org/10.1111/jaac.12703.

[323] Nayla Majestya and Irwanto Irwanto, 'Beyond Behavioral Change: Evaluating the Impact of Environmental Films on Audience Perceptions and Beliefs about Food System Issues', *Frontiers in Communication* 10 (June 2025), https://doi.org/10.3389/fcomm.2025.1519348.

changements[324]. **Ces résultats confirment la capacité du format documentaire à agir comme dispositif persuasif puissant, ce qui rend d'autant plus sensible l'introduction d'images synthétiques.**

Or, le champ documentaire a, paradoxalement, été l'un des premiers à adopter les hypertrucages, et l'examen de cette adoption précoce est particulièrement instructif. Lees (2024) analyse deux cas emblématiques de documentaires produits entre 2019 et 2022 ayant recouru à différentes formes d'hypertrucages à partir d'entretiens avec les cinéastes[325]. Le cas le plus discuté est celui de *Welcome to Chechnya* (David France, 2020), film consacré aux persécutions homophobes en Tchétchénie[326] : le réalisateur a utilisé des hypertrucages pour masquer les visages des sujets persécutés, substituant à leurs traits ceux de militants consentants ailleurs dans le monde. L'hypertrucage devient ici un dispositif de protection, paradoxalement mis au service de la vérité documentaire — comme le formule un producteur cité par Lees : « *I'm looking at deepfake as a way of telling the truth* » (Benjamin Field – *Je vois l'hypertrucage comme un moyen de dire la vérité*). Cette inversion éclaire la complexité du jugement éthique : le caractère synthétique d'une image ne dit rien en lui-même de sa fonction documentaire, et la question pertinente devient celle de la transparence du dispositif et du consentement des sujets.

Une étude récente prolonge cette analyse en distinguant trois usages principaux des technologies de synthèse faciale en contexte documentaire[327] : la restauration d'archives dégradées, la reconstitution de personnages disparus ou impossibles à filmer, et la génération d'images plausibles pour illustrer des récits factuels. Chacun de ces usages soulève des enjeux distincts d'authenticité et d'acceptabilité sociale, et appelle des dispositifs de transparence différenciés. La restauration s'inscrit dans une longue tradition d'intervention technique sur l'image (colorisation, étalonnage, restauration sonore) qui n'a jamais été perçue comme problématique ; la reconstitution mobilise des conventions documentaires comme la « dramatisation reconstituée » qui ont longtemps coexisté avec le genre ; la génération d'images plausibles, en revanche, est une rupture nouvelle dont la légitimité documentaire reste largement à construire.

Une analyse qualitative d'usages réels et de discours d'experts confirme que les technologies de synthèse faciale ne constituent pas une catégorie homogène sur le plan de leurs effets[328] : leur valeur —pédagogique, documentaire, patrimoniale — dépend de la configuration dans laquelle elles s'insèrent, c'est-à-dire de la combinaison entre la finalité déclarée, la transparence vis-à-vis du public et la présence de mécanismes de contrôle sur les usages dérivés. Lorsque ces conditions sont réunies, les auteurs de l'analyse recensent des usages légitimes et difficilement substituables : reconstitution de situations impossibles à filmer, animation de figures historiques dans des contextes éducatifs, enrichissement de récits dont les protagonistes ne peuvent être montrés. Lorsqu'elles ne le sont pas, les mêmes technologies fragilisent la véracité documentaire et désorientent des publics dont les repères perceptifs ne suffisent plus à distinguer l'authentique du fabriqué. Ce résultat rejoint celui de Lees : la question pertinente n'est pas de savoir si une image est synthétique, mais dans quel cadre elle a été produite, avec quel degré de transparence, et avec quelles garanties contre le détournement.

Au-delà de ces quelques articles, peu d'études académiques ont directement mesuré l'effet spécifique des images synthétiques (par rapport aux images authentiques ou aux reconstitutions classiques) sur la confiance, la mémoire ou les croyances du public dans un contexte documentaire. **Les travaux disponibles sur les hypertrucages et la désinformation suggèrent que des images réalistes mais fabriquées peuvent influencer la perception de la réalité et la confiance dans les médias, mais leur transposition au contexte documentaire reste à investiguer systématiquement.** De même, le risque d'accorder sa confiance à des « archives synthétiques » qui combine l'effet d'authenticité narrative[329] et la

---

[324] Sara K. Yeo et al., 'An Inconvenient Source? Attributes of Science Documentaries and Their Effects on Information-Related Behavioral Intentions', *Journal of Science Communication* 17, no. 2 (2018): A07, https://doi.org/10.22323/2.17020207.

[325] Dominic Lees, 'Deepfakes in Documentary Film Production: Images of Deception in the Representation of the Real', *Studies in Documentary Film* 18, no. 2 (2024): 108–29, https://doi.org/10.1080/17503280.2023.2284680.

[326] 'David France | Identity Protection with Deepfakes'.

[327] Bangjian You, 'Explore the Impact of the Application of Artificial Intelligence Facial Image Processing and Synthesis Technology in Documentaries on Social Acceptance', *Proceedings of the 2nd International Conference on Applied Psychology and Marketing Management*, 2025, https://www.scitepress.org/Papers/2025/141122/141122.pdf.

[328] Ebba Lundberg and Peter Mozelius, 'The Potential Effects of Deepfakes on News Media and Entertainment', *AI & SOCIETY* 40, no. 4 (2025): 2159–70, https://doi.org/10.1007/s00146-024-02072-1.

[329] Gillian Murphy et al., 'Face/Off: Changing the Face of Movies with Deepfakes', *PLOS ONE* 18, no. 7 (2023): e0287503, https://doi.org/10.1371/journal.pone.0287503.

dépendance des publics à des marqueurs perceptifs qui ne suffisent plus[330] est souvent évoqué ; la force normative de ces hypothèses est cependant encore largement spéculative sur le plan empirique même si leur cohérence avec les mécanismes établis justifie un investissement de recherche prioritaire. Crépel, Cardon et al. (2025) rappellent que **la crédibilité d'un contenu ne relève pas uniquement de sa fidélité au réel, mais de son inscription dans un écosystème médiatique structuré[331].** Ces contenus héritent alors d'un capital de crédibilité institutionnelle qui excède leurs propriétés propres. Pour Cardon (2026) ce qui assure la crédibilité d'une image n'est plus l'image en soi mais la personne ou l'institution qui nous l'adresse et avec laquelle nous passons un « contrat de lecture » de l'image[332]. Autrement dit, l'autorité épistémique se déplace du contenu vers le médiateur, déplacement qui pourrait être bénéfique pour les organisations culturelles déjà établies. Les risques associés de concentration et de blocage de nouveaux entrants éventuels, méritent cependant l'attention du décideur public.

Conclusion : érosion de la confiance et externalités négatives

À l'ensemble des mécanismes cognitifs examinés dans ce chapitre s'ajoute une dimension économique souvent négligée pour comprendre l'impact des hypertrucages. Des chercheurs ont établi empiriquement que la confiance dont une société hérite des générations précédentes prédit son développement économique de long terme, en créant les conditions d'une coopération soutenue, d'investissements de long terme et d'institutions stables[333]. Dans les sociétés où la confiance institutionnelle est déjà précaire, la déstabilisation peut s'opérer de manière d'autant plus rapide et aiguë. Cette analyse macroéconomique trouve un équivalent au niveau sectoriel : la confiance dans l'industrie culturelle d'un pays (sa réputation, l'autorité de ses créateurs, la crédibilité de ses institutions) est un stock accumulé sur plusieurs générations, dont l'érosion peut être rapide mais la reconstitution lente.

Or la production d'un hypertrucage engendre des coûts pour l'ensemble de la société en érodant ce stock de confiance. La production d'hypertucages non détectés dégrade en effet, par contamination, la valeur perçue de l'ensemble des contenus des secteurs médiatiques et culturels, y compris ceux produits par des humains. Ainsi, les hypertrucages qui sont produits par des acteurs identifiables (cf. Chap.2), produisent des effets négatifs sur l'ensemble des acteurs. Cette configuration est précisément celle d'une externalité négative ; le coût social de la production d'un hypertrucage excède son coût privé, parce qu'une partie du coût est répercutée sur des tiers sans compensation par celui qui est à l'origine du problème. Autrement dit les conséquences des hypertrucages ne sont pas internalisées par leurs producteurs qui en captent les bénéfices sans en supporter les coûts sociaux.

La question est d'autant plus sensible que le marché tend à produire de plus en plus d'hypertrucages, engendrant de plus en plus d'externalités négatives. En effet, comme le montre le rapport Europol le coût de production d'un hypertrucage de qualité décroît rapidement (logiciels open source, services de génération d'hypertrucages de type « deepfake-as-a-service », GPU accessibles) ; le coût de sa détection, au contraire, croît avec la sophistication des modèles génératifs (cf. Chap.6 pour les aspects techniques)[334]. Cette course asymétrique signifie que l'équilibre de marché tend vers une production massive non en raison de l'accroissement du nombre d'acteurs malveillants mais parce que la structure des coûts favorise la multiplication des hypertrucages.

Ainsi, sans signal fort qui internalise les coûts des effets négatifs des hypertrucages pour la société et réduise le coût asymétrique de la production face à la détection, les acteurs privés n'ont pas d'intérêt à agir. Les chapitres suivants examinent les dispositifs techniques et réglementaires qui peuvent contribuer à limiter ces effets.

---

[330] Hashmi et al., 'Unmasking Illusions'.

[331] Maxime Crépel et al., *Une Cartographie Web de l'écosystème Médiatique Français*, 2025, https://sciencespo.hal.science/hal-05094289/.

[332] Dominique Cardon, 'Les Vidéos Générées Par IA Ne Font Pas Disparaître l'intérêt Que Nous Portons Au Réel', *Le Monde*, 11 February 2026, https://www.lemonde.fr/idees/article/2026/02/11/dominique-cardon-sociologue-les-videos-generees-par-ia-ne-font-pas-disparaitre-l-interet-que-nous-portons-au-reel_6666388_3232.html.

[333] Yann Algan and Pierre Cahuc, 'Inherited Trust and Growth', *American Economic Review* 100, no. 5 (2010): 2060–92, https://doi.org/10.1257/aer.100.5.2060.

[334] 'Facing Reality? Law Enforcement and the Challenge of Deepfakes', Europol, accessed 1 June 2026, https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

5. LA TRANSPARENCE, UNE REPONSE PARTIELLE A LA DEMANDE DU PUBLIC ET DES AYANTS-DROIT, DEVANT ACQUERIR SA PLEINE PORTEE

La définition de l'hypertrucage, telle qu'elle figure à l'article 3, point 60, du RIA, n'a, pour les auteurs de ce règlement, d'autre portée que de déterminer les contenus auxquels s'appliquent les obligations spéciales prévues par l'article 50. Cela ressort clairement de l'articulation entre les paragraphes 2 et 4 de cet article, et des considérants 133 et 134 du règlement : compte tenu des risques spécifiques d'usurpation d'identité, de tromperie et de manipulation de l'information, le législateur a entendu rendre visible le fait qu'un contenu a été généré ou modifié par un système d'IA et constitue un hypertrucage.

La transparence est essentielle et le législateur français a lui-même souhaité agir sur ce plan face aux comportements d'influenceurs de nature à porter atteinte à l'intérêt général, en complément de certaines interdictions[335]. Pour autant, face aux enjeux soulevés, dont les facettes sont nombreuses et vont au-delà de ceux identifiés au considérant 132, **la seule obligation de transparence est insuffisante** : la circonstance que la réponse apportée aux hypertrucages méconnaissant les droits des personnes requière de mobiliser de nombreux corpus de normes manifeste le fait que la transparence ne suffit pas à répondre aux questions et difficultés que ces contenus soulèvent. Outre les points soulignés dans la partie 3, il en résulte également un enjeu de renforcement de la responsabilité, qui sera évoqué dans le chapitre 7. **En sens inverse, la transparence ne suffit pas à écarter le risque de manipulation ou celui de confusion par le public** : non seulement cela dépend de la connaissance du degré de manipulation, mais de nombreux éléments contextuels, ainsi que cela a été souligné au chapitre 4, sont de nature à limiter la portée de la transparence.

S'agissant, en outre, de la transparence elle-même, on ne peut limiter les obligations relatives à ces contenus hypertruqués à ces deux seuls paragraphes relatifs à la transparence : si la transparence est au cœur de l'appréhension des hypertrucages par le RIA (**5.1**), ceux-ci s'inscrivent dans un ensemble de règles plus vastes, au sein même de ce règlement (**5.2**) et il apparait nécessaire donner sa pleine portée au principe de transparence.

5.1. La gestion des hypertrucages par la transparence : une démarche asymétrique et incomplète au sein du RIA

Le RIA, comme les récentes législations européennes portant sur le secteur numérique, repose sur une identification et une gestion des risques par les opérateurs afin de protéger les droits fondamentaux[336] et permettre une meilleure adaptabilité aux évolutions techniques et technologiques, en association avec les acteurs[337]. Or, pour répondre aux risques identifiés de manière prioritaire – et sans préjudice, on le verra, d'autres risques plus importants – liés à « *certains* » systèmes d'IA, auxquels le chapitre IV du règlement est consacré, et qui sont, en particulier, ceux qui génèrent du contenu (autrement dit, les IA génératives), le règlement s'appuie sur **une obligation de transparence, qui se décline selon deux cercles concentriques : une obligation, applicable aux fournisseurs de systèmes d'IA générative, portant sur l'ensemble des contenus de synthèse générés par IA (art. 50, paragr. 2) et une obligation, pesant sur les déployeurs de ces SIA, spécialement applicable aux hypertrucages (art. 50, paragr. 4)**.

Il faut relever ici ce qui s'inscrit dans la logique d'ensemble du RIA : **celui-ci s'intéresse aux systèmes et en fixe les principes de fonctionnement ; il ne s'intéresse pas directement aux contenus**, si ce n'est comme élément de mise en perspective de ces systèmes[338]. Or, une des difficultés tient, précisément, au risque de génération de contenus illicites ou préjudiciant aux droits des tiers. Compte tenu de son approche, le RIA appréhende ces risques de manière objective, ce qui conduit à une objectivisation

---

[335] Loi n° 2023-451 du 9 juin 2023 visant à encadrer l'influence commerciale et à lutter contre les dérives des influenceurs sur les réseaux sociaux, modifiée par l'ordonnance n° 2024-978 du 6 novembre 2024.

[336] Arnaud Latil, « Le Règlement relatif à l'intelligence artificielle et l'approche par les risques : application d'une méthode législative structurante », *RTD eur.*, 2024, p. 563.

[337] Voy. sur la mise en avant des avantages de cette approche, par ex. Giovanni de Gregorio, Oreste Pollicino, « The European Constitutional Way to Address Desinformation in the Age of Artificial Intelligence », *German Law Review*, 2025, vol. 26, p. 449.

[338] Approche qui a pu être critiquée (Théo Antunes, « Prévenir l'inacceptable : le rôle du règlement européen dans la gouvernance des pratiques prohibées liées à l'intelligence artificielle générative », *Rev. de l'Union européenne*, 2025, p. 609) mais qui vient compléter, ainsi qu'on le verra, celle du règlement sur les services numériques.

des droits fondamentaux en jeu[339] – et l'absence de prise en compte de la dimension subjective de ces droits, qui est également confrontée aux conséquences de la génération et de la diffusion d'hypertrucages. Cette approche présente l'avantage de permettre la prise en compte, en amont, de la potentialité de certains préjudices, mais elle est restreinte pas la circonstance que le RIA s'abstient d'évoquer, au-delà de réserves de principes, les questions qui intéressent directement les acteurs de la création et de la culture. Il s'agit ainsi, pour la mission, de **s'efforcer de resituer les responsabilités des différents acteurs à l'égard des hypertrucages au regard de l'ensemble des textes applicables et enjeux en cause, à commencer évidemment par ce qui concerne la transparence**.

Cette notion de transparence renvoie, selon le considérant 27, « *au fait que les systèmes d'IA sont développés et utilisés de manière à permettre une traçabilité et une explicabilité appropriées, faisant en sorte que les personnes réalisent qu'elles communiquent ou interagissent avec un système d'IA, que les déployeurs soient dûment informés des capacités et des limites de ce système d'IA et que les personnes concernées soient informées de leurs droits* ». **Il ne s'agit néanmoins que d'une transparence portant sur la nature du contenu**, et non sur d'autres aspects plus larges, en particulier les contenus utilisés à titre de données d'apprentissage, dimension que le rapport Voss[340] suggère de prendre en compte et que les obligations de l'article 50 ne sont vraisemblablement pas susceptibles de contribuer à assurer[341]. Elle constitue toutefois une nécessité intrinsèque car, non seulement elle évite la manipulation (ou, à tout le moins, elle tend à en réduire le risque), mais elle crée une distance par rapport au contenu généré[342].

S'agissant des hypertrucages, **ce principe de transparence renvoie à deux déclinaisons prévues à l'article 50 : celle de marquage, applicable à l'ensemble des contenus synthétiques, et celle d'étiquetage, propre aux hypertrucages**, complétée par une déclinaison relative à certains textes.

5.1.1. L'obligation de transparence applicable à tous les contenus synthétiques : le marquage

> Selon l'article 50, paragraphe 2 :
> « 2. Les fournisseurs de systèmes d'IA, y compris de systèmes d'IA à usage général, qui génèrent des contenus de synthèse de type audio, image, vidéo ou texte, veillent à ce que les sorties des systèmes d'IA soient marquées dans un format lisible par machine et identifiables comme ayant été générées ou manipulées par une IA. Les fournisseurs veillent à ce que leurs solutions techniques soient aussi efficaces, interopérables, solides et fiables que la technologie le permet, compte tenu des spécificités et des limites des différents types de contenus, des coûts de mise en œuvre et de l'état de la technique généralement reconnu, comme cela peut ressortir des normes techniques pertinentes. Cette obligation ne s'applique pas dans la mesure où les systèmes d'IA remplissent une fonction d'assistance pour la mise en forme standard ou ne modifient pas de manière substantielle les données d'entrée fournies par le déployeur ou leur sémantique, ou lorsque leur utilisation est autorisée par la loi à des fins de prévention ou de détection des infractions pénales, d'enquêtes ou de poursuites en la matière. »

L'obligation de marquage qui pèse ainsi sur les fournisseurs d'IAG, c'est-à-dire ceux qui développent ou font développer un tel système d'IA ou un modèle d'IA à usage général permettant la génération de contenus et le mettent sur le marché ou mettent le système d'IA en service sous leur propre nom ou marque, à titre onéreux ou gratuit[343], se veut techniquement contraignante. Elle n'en demeure pas

---

[339] Alexandra Bensamoun, Grégoire Loiseau, *Droit de l'intelligence artificielle*, Paris, LGDJ-Lextenso, 2024, 2e éd., p. 44.
[340] *Le droit d'auteur et l'intelligence artificielle générative – opportunités et défis*, rapport de la Commission des affaires juridiques adopté par le Parlement européen lors de sa séance du 10 mars 2026, doc. A10-0019/2026 – P10_TA(2026)0066.
[341] Kateryna Militsyna, « Can Copyright Law Benefit from the Marking Requirement of the AI Act? », *International Review of Intellectual Property and Competition Law*, vol. 56n 2025, p. 1734.
[342] Voy. la chronique de Guillaume Fraissard, *Le Monde*, 20 janvier 2026, « Reprise de « Papaoutai » avec l'IA : « Savoir qu'une œuvre est fausse change la nature de l'émotion qu'elle nous procure » », ainsi que l'interview de Dominique Cardon, « Les vidéos générées par IA ne font pas disparaître l'intérêt que nous portons au réel », *Le Monde*, 11 février 2026.
[343] Définition du « fournisseur » de l'article 3, point 3, du RIA.

moins sujette à interprétation et les travaux sur le code de bonnes pratiques ont été l'occasion pour les participants et observateurs[344] d'exprimer des positions contradictoires[345].

On soulignera à cet égard que les travaux sur ce **code de bonnes pratiques de l'article 50** se sont déroulés parallèlement à la présente mission, de sorte que celle-ci a mené ses réflexions au regard des deux premières versions de ce document, telle que présentées par la Commission aux fins de consultation ; elle s'est toutefois efforcée de tenir compte de **la version finale, publiée quelques jours avant la finalisation du présent rapport**[346]**.** Il n'en demeure pas moins que, s'agissant d'un document de droit souple portant sur une matière particulièrement évolutive, et comme le souligne le rapport Voss pour le code relatif à l'IA à usage général, de tels documents « *doivent être révisés et considérés comme des documents évolutifs nécessitant des mises à jour régulières* » (point 14). La pratique doit pouvoir, dans ce cadre, faire l'objet d'une étude attentive afin de faire évoluer les recommandations et s'interroger, le cas échéant, sur la manière de leur donner une portée normative. **Des évaluations régulières apparaissent donc nécessaires.**

L'obligation de transparence résultant de l'article 50, paragraphe 2, appelle plusieurs commentaires sur le plan juridique.

a) **Les débiteurs de l'obligation sont les fournisseurs de SIA**, selon la définition du RIA, sans autre restriction.

Le législateur européen a donc entendu retenir, en ce qui concerne tant le champ matériel que territorial, une approche large, en cohérence avec les objectifs poursuivis. Elle permet de responsabiliser l'ensemble de ces fournisseurs, qui ne peuvent, ce faisant, se désintéresser de l'utilisation d'outils dont les risques sont susceptibles d'être supérieurs aux avancées offertes, tant au regard des enjeux démocratiques[347] que pour les droits des personnes[348]. La dissémination d'hypertrucages contrevenant à une législation est certes de la responsabilité des déployeurs, mais les outils d'IA générative ont précisément la particularité de permettre cette dissémination de manière très aisée, au risque de remettre en cause la confiance minimale nécessaire au fonctionnement d'une société démocratique et libérale. C'est aussi une question de réalisme technique et économique : seuls les fournisseurs disposent de la surface critique permettant de mettre en œuvre une telle obligation technique répondant à des contraintes fortes.

Force est de constater que l'UE n'a pas été la seule dans cette démarche, qui était aussi celle retenue par l'administration Biden aux Etats-Unis[349]. En outre, bien que sa portée fût limitée aux élections américaines de 2024, la « déclaration de Munich[350] » retient cette approche, démontrant la prise de conscience par les signataires de leur responsabilité.

La mission estime donc que, malgré les demandes de certains acteurs, la portée conférée à l'article 50, paragraphe 2, doit être aussi étendue que possible. Même si l'utilisateur (au sens large, y compris les déployeurs) a la responsabilité de l'instruction générative (« *prompt* »), le fournisseur du système ne peut se désintéresser de ce qui est généré et doit, à ce titre, participer

---

[344] Dont la liste a été publiée par la Commission :
https://ec.europa.eu/newsroom/dae/redirection/document/122890

[345] https://digital-strategy.ec.europa.eu/en/news/working-groups-advance-discussions-transparency-obligations-under-article-50-ai-act

[346] https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content#1720699867912-0

[347] Voy. par ex. Kassandra Goni, « La police de la désinformation à l'épreuve des deepfakes », *AJDA*, 2025, p. 689 ; Thomas Hochmann, « Lutter contre les fausses informations : le problème préliminaire de la définition », *Rev. des droits et libertés fondamentaux*, 2018, n° 6 (en ligne).

[348] On pense notamment à la génération d'hypertrucages à caractère pornographique, qui constituent une part importante de l'utilisation de ces outils, mais les hypothèses d'illicéité sont plus nombreuses, ainsi que cela sera évoqué par la suite.

[349] « *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* », *Executive Order* 14110 du 30 octobre 2023. Sur la similitude avec l'approche du RIA, voy. Alistair Knott *et al.*, « AI content detection in the emerging information ecosystem: new obligations for media and tech companies », *Ethics and Information Technology*, vol. 26, 2024, article n° 63.

[350] *Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, signé lors de la conference de Munich sur la sécurité de février 2024 par Adobe, Amazon, Anthropic, ARM, ElevenLabs, Gen, GitHub, Google, IBM, Inflection AI, Intuit, LG AI Research, LinkedIn, McAfee, Meta, Microsoft, NetApp, Nota, OpenAI, Snap, Stability AI, TikTok, TrendMicro, TrueMedia, TruePic, X et ZEFR.

activement à la capacité de détection de ces contenus. Il est d'ailleurs notable que, s'agissant d'une des catégories les plus sensibles d'hypertrucages – ceux à caractère sexuel – les négociations en cours sur le règlement omnibus envisagent de responsabiliser les fournisseurs de SIA en en faisant expressément une pratique interdite.

Par ailleurs, la question a été posée de savoir si, en aval, l'obligation s'étendait *de facto* aux déployeurs ou si ceux-ci étaient libres de supprimer ce marquage[351]. Le règlement est en effet silencieux sur ce point et son article 50 semble distinguer deux champs distincts. Toutefois, l'effet utile du règlement appelle une réponse claire, résultant des termes mêmes de l'article 50, paragraphe 2 : le marquage doit être solide, ce qui supposerait qu'il ne puisse être supprimé. **La mission estime donc particulièrement pertinent que les fournisseurs de systèmes d'IA générative incluent dans leurs clauses contractuelles une interdiction de supprimer les dispositifs de marquage qu'ils auront insérés**.

b) Sans évoquer dès à présent les aspects techniques (point 5.2. ci-après), **cette obligation de marquage est également définie de manière large et les dérogations sont limitées**.

Il faut d'abord relever que, hors champ des hypertrucages définis par le RIA, **cette disposition inclut les textes**, malgré les difficultés techniques importantes qu'ils posent pour assurer un marquage pérenne. Pour ces contenus, la détection paraît plus opérationnelle que le marquage, ce qui montre d'ores et déjà l'imbrication très forte entre ces deux exigences. La notion de « texte » demeure toutefois imprécise, puisqu'il n'est pas posé d'exigence de signification de ce texte (les lignes directrices de la Commission sur l'article 50 précisant néanmoins (paragr. 60) qu'un texte au sens de ce paragraphe doit pouvoir être lu et interprété sur le plan sémantique par un humain), tandis que les enjeux sont, notamment au stade de leur diffusion, très différents entre la génération, par ces systèmes d'IA, d'un texte prenant la forme d'une nouvelle dans un genre littéraire (ce qui n'est pas sans poser des questions quant aux inspirations) ou d'un faux article de presse ou scientifique.

Ensuite, la définition de cette obligation permet de couvrir tous les contenus générés <u>*ou*</u> manipulés par ces systèmes d'IA. L'introduction de la référence aux contenus manipulés par IA répond à une interrogation soulevée au cours de la consultation : l'obligation de marquage porte sur tous les contenus pour lesquels l'IA est intervenue, même s'il ne s'agit pas d'une création entièrement réalisée par IA. La seule circonstance qu'un contenu puisse être modifié partiellement par un tel système suffit à rendre applicable cette exigence[352] et cela est de nature à donner son plein effet à l'obligation, les contenus manipulés étant les plus susceptibles de voir la trace de la modification disparaître[353].

La limite de cette obligation figure dans le même alinéa de l'article 50, paragraphe 2 : il s'agit, outre l'exception pénale qui n'est pas au cœur de la présente mission, des systèmes d'IA qui ne remplissent qu'une fonction d'assistance pour la mise en forme ou ne modifient pas de manière substantielle les données d'entrée fournies par le déployeur ou leur sémantique. Il faut relever que cette exception, même si des interrogations ont pu être soulevées notamment sur le caractère substantiel de la modification, repose sur une appréciation <u>en amont de la génération</u> du contenu, qui tient compte des seules capacités du système d'IA : **l'obligation est définie en effet par rapport au fournisseur et à l'égard de son système, et non au regard du contenu généré**. Par sa rédaction, cette exception confirme en outre que les systèmes d'IA permettant la seule manipulation de contenus sont soumis à cette obligation. Néanmoins, si l'on admet que cette exception couvre les traductions[354], cela pose une question de traçabilité des traductions opérées par un système d'IA, mais il ne s'agit alors pas d'un hypertrucage au sens du règlement.

---

[351] Thomas Gils, « A Detailed Analysis of Article 50 of the EU's Artificial Intelligence Act », *in* C. N. Pehlivan, N. Forgó and P. Valcke (eds.), *The EU Artificial Intelligence (AI) Act: A Commentary*, Kluwer Law International, 2025, p. 776-823.

[352] Voy. également en ce sens Kateryna Militsyna, « Can Copyright Law Benefit from the Marking Requirement of the AI Act? », préc. ou Thomas Gils, préc.

[353] Alistair Knott *et al.*, préc.

[354] En ce sens : Martina Block, « A Critical Evaluation of Deepfake Regulation through the AI Act in the European Union », *EuCML*, 2024, vol. 4, p. 184.

Quant à la nature du marquage lui-même, les principes d'efficacité, d'interopérabilité, de solidité et de fiabilité sont essentiels à l'effectivité de l'obligation. **La mission estime que sa mise en œuvre doit veiller à lui conférer sa pleine portée à plusieurs égards** :

- Il s'agit d'une obligation technique, qui, contrairement à celle de l'article 50, paragraphe 4, n'a pas nécessairement vocation à altérer la perception du contenu généré ou manipulé. Le paragraphe 5 du même article impose seulement que les informations soient fournies de manière claire et reconnaissable au plus tard au moment de la première interaction ou exposition, ce qui implique en réalité un système aval de détection, évoquée au point *c* ci-dessous. Il n'y a donc pas d'obstacle à ce qu'elle soit appliquée de manière aussi large que possible ;
- Il s'agit d'une obligation évolutive au regard des technologies existantes et le marquage nécessite une mobilisation de l'ensemble des acteurs. A cet égard, le règlement prévoit expressément que cette obligation est appréciée au regard de l'état de la technique généralement reconnu. En revanche, la mission estime que l'exception liée aux coûts de cette technique doit être lue avec prudence, d'autant qu'elle s'appliquera essentiellement à des opérateurs disposant de capacités financières étendues : celles-ci ne doivent pas conduire à des obligations à géométrie variable selon l'opérateur fournissant le système d'IA, sous réserve des aménagements prévus pour les petites et moyennes entreprises, et, de ce fait, incitent à relativiser cette dimension, qui ne peut conduire à une vision minimaliste de l'exigence de transparence. En outre, la prise en compte de la dimension technique et financière ne peut, en tout état de cause, prévaloir sur l'intérêt démocratique qui a guidé ces dispositions : en effet, quelle que soit la volonté des acteurs des systèmes d'IA de s'inscrire dans une démarche régulatrice, leurs perspectives sont, du fait d'un biais d'analyse inévitable, distinctes des enjeux généraux de la société[355].
- Compte tenu des enjeux, la mission approuve la démarche du code de bonnes pratiques consistant à ce que marquage soit « multi-couches », seule manière d'éviter un effacement trop aisé du marquage. La charge pèse certes sur les fournisseurs de systèmes d'IA, mais il s'agit de prévenir les utilisations attentatoires aux droits des tiers du fait de contenus générés par IA, et la responsabilité du fournisseur ne peut être dégagée dès lors qu'il lui appartient de mettre en place un marquage solide et fiable ;
- Le règlement ne prévoit aucune obligation de marquage de l'origine du contenu généré ou manipulé par IA, ce qui est une limite évidente pour les questions de propriété littéraire et artistique ou d'atteinte aux droits de la personnalité faute de pouvoir identifier l'auteur de l'illicéité. Les lignes directrices de la Commission sur l'article 50 en tirent les conséquences en précisant que, bien que le considérant 133 évoque les techniques « *permettant de prouver la provenance et l'authenticité du contenu* », les fournisseurs ne sont pas tenus d'enregistrer ou conserver l'ensemble des informations sur l'origine des contenus (paragr. 73). Pour autant, la mission ne peut qu'encourager de telles démarches. L'efficacité du dispositif ne pouvant reposer que sur l'émergence d'un écosystème d'acteurs responsables et les sanctions prévues par le RIA, pour lourdes qu'elles soient, supposant l'identification des auteurs des manquements ainsi qu'une réelle capacité à les sanctionner, la mission estime également qu'au-delà du cadre répressif posé par l'article 99 du règlement, des incitations doivent compléter la mise en œuvre de ces obligations, afin que les obligations pesant sur ces acteurs ne soient pas supérieures à celles pesant sur les opérateurs non respectueux de la réglementation européenne.

c) La **portée de l'obligation** définie par l'article 50, paragraphe 2, pose une question centrale qui est celle de **la détection et** de **l'exigence de mise à disposition d'outils à cette fin**. La détection est indispensable au fonctionnement des mécanismes prévus par l'article 50[356]. Lors des travaux sur le code de bonnes pratiques, un débat est apparu sur l'obligation conjointe, pour les fournisseurs de systèmes d'IA, de fournir des outils de détection, débat qui était déjà apparu en doctrine[357].

---

[355] Voy. à cet égard Maria Pawelec, « Decent deepfakes? Professional deepfake developers' ethical considerations and their governance potential », *AI and Ethics*, 2025, vol. 5, p. 2641.

[356] Mateusz Łabuz, « Deep fakes and the Artificial Intelligence Act—An important signal or a missed opportunity? », *Policy Internet*, 2024, vol. 16, p. 783.

[357] Par ex. Alistair Knott *et al.*, préc. ; Felipe Romero-Moreno, « Deepfake detection in generative AI : A legal framework proposal to protection human rights », *Computer Law & Security Review: The International Journal of Technology Law and Practice*, 2025, vol. 58, 106162.

L'article 50 ne comporte pas d'obligation directe sur ce point, puisqu'il se borne, au paragraphe 2, à poser une obligation de résultat : le contenu généré doit être identifiable (« *detectable* » dans la version en anglais)**, ce qui suppose qu'il soit possible de procéder à cette détection.** A cet égard, le paragraphe 5 rappelé ci-dessus impose que l'information du paragraphe2, c'est-à-dire l'existence d'un marquage, soit fournie aux personnes physiques de manière claire et reconnaissable au plus tard à la première interaction ou exposition. Si l'on fait abstraction de la responsabilité qui pèse sur le déployeur au titre du paragraphe 5, **l'obligation qui en résulte pour le fournisseur est bien de rendre possible la détection du contenu généré par l'outil qu'il met à disposition : le marquage, *a priori* purement technique car devant être lisible par machine, doit *in fine* pouvoir devenir visible et compréhensible pour l'être humain**. Si la lettre des dispositions n'est pas d'une clarté immédiate, l'effort d'interprétation est faible dès lors qu'il s'agit de la seule manière de donner un effet utile à l'inclusion du paragraphe 2 dans l'obligation du paragraphe 5 et que l'intention du législateur est claire : le considérant 135 du RIA pose qu' « *il convient d'exiger que les fournisseurs de ces systèmes intègrent des solutions techniques permettant le marquage dans un format lisible par machine et la détection du fait que les sorties ont été générées ou manipulées par un système d'IA, et non par un être humain* ».

Au demeurant, la mission estime qu'au-delà de la fourniture d'un système de détection, le RIA implique que, pour satisfaire à l'obligation du paragraphe 5, **les fournisseurs participent à la mise à disposition d'outils de détection**, qu'ils leurs soient propres ou mutualisés. C'est d'ailleurs le sens du considérant 135, qui appelle à une coopération sur ce point :

> « Sans préjudice de la nature obligatoire et de la pleine applicabilité des obligations de transparence, la Commission peut également encourager et faciliter l'élaboration de codes de bonne pratique au niveau de l'Union afin de faciliter la mise en œuvre effective des obligations relatives à la détection et à l'étiquetage des contenus générés ou manipulés par une IA, y compris pour favoriser des modalités pratiques visant, selon qu'il convient, à rendre les mécanismes de détection accessibles et à faciliter la coopération avec d'autres acteurs tout au long de la chaîne de valeur, à diffuser les contenus ou à vérifier leur authenticité et leur provenance pour permettre au public de reconnaître efficacement les contenus générés par l'IA. »

C'est au demeurant également l'interprétation retenue par la Commission européenne dans ses lignes directrices sur l'article 50 du RIA (paragr. 75) et les signataires du code de bonne pratique s'engageront en ce sens.

Dans ce cadre, **la mission partage l'interprétation de l'article 50 qui retient que les fournisseurs de SIA doivent participer activement à la détection des contenus générés, soit par la mise à disposition d'un outil, soit par la participation à un outil mutualisé.** L'interopérabilité exigée par le RIA devrait à cet égard inciter à suivre cette seconde voie, en complément d'une standardisation des méthodes de marquage, par ailleurs susceptible de réduire les coûts associés. En tout état de cause, et également au regard des considérations qui suivent sur la notion de déployeur, **la mission est encline à estimer que le RIA mériterait d'être précisé sur ce point.**

Néanmoins, si la mission observe que de nombreuses entreprises engagées dans la conception de systèmes d'IA investissent également dans les outils de détection, la capacité à détecter les contenus manipulés ou générés par IA constitue également un enjeu démocratique. Il serait opportun qu'à l'occasion d'évolutions futures du RIA, une obligation de transparence sur la manière de détecter ces contenus soit mise en place, permettant le développement aisé d'outils de détection, qui ne soient pas seulement liés à l'engagement des acteurs du secteur.

5.1.2. L'obligation de transparence spécifique aux hypertrucages : l'étiquetage

L'article 50, paragraphe 4, donne sa portée à la définition de l'hypertrucage par l'article 3 du RIA ; c'est en effet la seule disposition qui lui est dédiée :

> « 4. **Les déployeurs d'un système d'IA** qui génère ou manipule des images ou des contenus audio ou vidéo constituant un hypertrucage indiquent que les contenus ont été générés ou manipulés par une IA. Cette obligation ne s'applique pas lorsque l'utilisation est autorisée par la loi à des fins de

prévention ou de détection des infractions pénales, d'enquêtes ou de poursuites en la matière. Lorsque le contenu fait partie d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue, les obligations de transparence énoncées au présent paragraphe se limitent à la divulgation de l'existence de tels contenus générés ou manipulés d'une manière appropriée qui n'entrave pas l'affichage ou la jouissance de l'œuvre.
Les déployeurs d'un système d'IA qui génère ou manipule des textes publiés dans le but d'informer le public sur des questions d'intérêt public indiquent que le texte a été généré ou manipulé par une IA. Cette obligation ne s'applique pas lorsque l'utilisation est autorisée par la loi à des fins de prévention ou de détection des infractions pénales, d'enquêtes ou de poursuites en la matière, ou lorsque le contenu généré par l'IA a fait l'objet d'un processus d'examen humain ou de contrôle éditorial et lorsqu'une personne physique ou morale assume la responsabilité éditoriale de la publication du contenu. »

L'obligation d'étiquetage (selon le terme utilisé par le considérant 134 du RIA) ainsi instaurée est à destination du public : **il s'agit de l'informer** qu'il est face à un hypertrucage, ce qui suppose, ainsi que le prévoit le projet de code de bonne pratiques, un signe, dont il est souhaitable qu'il soit harmonisé afin d'en faciliter la compréhension par le grand public.

a) Les débiteurs de l'obligation d'étiquetage

Cette **obligation d'étiquetage pèse sur les déployeurs** , tels que définis par l'article 3, point 4 : « *une personne physique ou morale, une autorité publique, une agence ou un autre organisme utilisant sous sa propre autorité un système d'IA sauf lorsque ce système est utilisé dans le cadre d'une activité personnelle à caractère non professionnel* ».

Cette définition du déployeur appelle en elle-même un commentaire dans la mesure où, si elle se veut la plus englobante possible quant aux personnes utilisant un système d'IA, **elle comporte une exclusion importante** : l'utilisation de ce système dans le cadre d'une activité personnelle à caractère non professionnel, en cohérence avec l'article 2, paragraphe 10, du RIA. **Il s'agit, autrement dit, de ne pas mettre d'obligation sur la personne physique qui utilise ce système à titre exclusivement privé**, ce qui semble correspondre à l'exception de l'activité strictement personnelle ou domestique que comporte le RGPD (art. 2, paragr. 2, sous c). Le changement de terme retenu par rapport à la proposition initiale de la Commission, qui se référait à l'utilisateur, manifeste clairement cette exclusion.

Or, d'une part, face à la capacité de dissémination des publications en ligne faites par des particuliers, notamment à travers les réseaux sociaux, les risques ne sont pas moindres que lorsque le système d'IA est utilisé à des fins professionnelles, s'il n'est d'ailleurs amplifié. D'autre part, la référence au caractère « non professionnel » de l'activité, bien connue et précisée dans d'autres législations européennes (dont il paraît peu vraisemblable de se détacher totalement nonobstant l'autonomie que les définitions peuvent avoir d'un texte à l'autre), paraît peu compatible avec une interprétation étroite de cette exclusion, qui se limiterait au seul usage privé dans un cercle restreint[358]. Il s'agit néanmoins d'une exception particulière, tenant à la fois à la vie privée et à la liberté d'expression et il faut être attentif à la double condition du caractère personnel et non professionnel, ce qui implique d'apprécier la finalité de l'utilisation et son cadre, en particulier au regard du critère usuel de l'audience de la diffusion[359]. La mission a à cet égard relevé que, dans les lignes directrices relatives à l'article 50 du RIA, la Commission européenne n'a finalement pas limité, comme le projet publié en mai 2026 l'envisageait, la portée de l'usage à titre purement personnel autant que possible, en estimant notamment que des activités criminelles étaient nécessairement exclues de cette qualification, de même que tout hypertrucage rendu public ou ayant un impact public (paragr. 19-20). La version finale des lignes directrices de l'article 50 semble plus réservée sans pour autant écarter explicitement les cas d'usages limites (un influenceur professionnel qui pourrait agir comme déployeur par exemple).

La régulation des hypertrucages « privés » est renvoyée en aval, vers les fournisseurs de services intermédiaires, dans les conditions qui seront précisées par la suite et qui sont déterminées par

358 Interprétation proposée par ex. par Martina Block, art. préc.
359 Par ex. CEDH, 25 mars 2021, *Matalas c. Grèce*, n° 1864/18, paragr. 55 ; 18 janvier 2024, *Allée c. France*, n° 20725/20, paragr. 48

les autres législations applicables, en particulier le règlement sur les service numériques[360] (DSA). Mais, à l'égard d'un usage non professionnel, le principal point d'entrée d'une action efficace sera l'article 35 de ce règlement, qui met à la charge des fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne (ce qui constitue une cible restreinte, même si elle touche l'essentiel de la population) l'obligation de mettre en place des mesures d'atténuation des risques systémiques, parmi lesquels l'article 34, paragraphe 1, sous *a* identifie « *la diffusion de contenus illicites par l'intermédiaire de leurs services* ». Pour cela, ils sont invités à déployer différents outils, parmi lesquels le paramétrage des algorithmes, des systèmes de modération ou la définition de conditions générales adaptées (art. 34, paragr. 2).

Mais **cela manifeste aussi l'importance de donner sa pleine portée à l'obligation pesant sur les fournisseurs de ces systèmes en matière de marquage et de détection : ce sont les seuls acteurs qui interviennent en toute hypothèse** et seuls les outils de détection sont, au stade de la diffusion, en mesure de prévenir les risques liés aux hypertrucages.

b) La notion de contenu généré ou manipulé constituant un hypertrucage

Les travaux sur le code de bonne conduite ont montré que le champ de l'obligation d'étiquetage pouvait prêter à interprétation afin de tenter d'en réduire la portée au bénéfice d'une uniformisation de la signification du terme « *manipulé* » dans les différents articles du RIA, en particulier aux articles 5 et 50.

Or, du point de vue de la mission, le RIA est clair dès lors qu'on garde à l'esprit que le verbe « *manipuler* » a deux significations[361] :

- Prendre dans ses mains, tenir entre ses mains certains produits, appareils, objets, pour effectuer une opération déterminée ;
- Modifier ou falsifier quelque chose pour masquer ou déformer une réalité ; truquer.

Si la référence à la manipulation à l'article 5 renvoie de manière évidente à cette seconde définition, péjorative, tel n'est pas le cas des mentions figurant à l'article 50, qui sont largement objectives, renvoyant à la première définition (lue à la lumière des outils numériques), ainsi qu'en témoigne la juxtaposition des termes « *généré* » et « *manipulé* ». Comme indiqué précédemment, ce dernier terme renvoie aux contenus qui ont été modifiés par un système d'IA, contrairement à ceux générés qui ont été créés par un tel système. Ainsi qu'en témoigne l'historique de la discussion sur la proposition de RIA et la mention initiale de l'utilisateur plutôt que du déployeur quant à cette obligation, c'est bien l'usage d'un contenu généré ou manipulé par IA qui est visé. Toute autre interprétation est manifestement en contradiction avec la prohibition de l'article 5.

En revanche, il est notable que, dans ses lignes directrices sur l'article 50 du RIA, , **la Commission européenne estime qu'il n'y a pas lieu de considérer comme étant un contenu manipulé par IA un contenu qui n'aurait fait l'objet que de manipulations mineures, d'ordre technique** (paragr. 116). **Cette position *de minimis* est conforme tant à l'effet utile du règlement qu'à l'équilibre de la définition de l'hypertrucage** : il faut, pour retenir cette qualification, qu'existe un risque de perception erronée par le public du contenu comme authentique ou véridique. Une manipulation mineure, d'ordre technique, n'est pas de nature à caractériser ce critère, puisque le contenu demeure alors authentique. Ainsi, un film qui ferait l'objet de simples améliorations de lumière par IA n'est pas qualifiable d'hypertrucage.

c) Le contenu de l'obligation d'étiquetage

Le principe de cette obligation a pu faire débat car sa finalité n'apparaît pas, à la réflexion, toujours évidente[362]. La seule circonstance qu'un contenu ait été généré ou manipulé par un système d'IA n'implique pas en effet qu'il soit faux ou frauduleux, objectif qui, spontanément, semble devoir être immédiatement poursuivi au regard des risques identifiés[363]. En se référant à la définition même de l'hypertrucage, qui

---

[360] 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).
[361] Nous reprenons celles de la 9e édition du dictionnaire de l'Académie française.
[362] Michael Veale, Frederik Zuiderveen Borgesius, « Demystifying the Draft EU Artificial Intelligence Act », *Computer Law Review International*, 2021, vol. 22, n° 4, p. 97-112.
[363] Mariëtte van Huijstee *et al.*, *Tackling deepfakes in European policy*, European Parliamentary Research Service – Study, juillet 2021, doc. PE 690.039.

s'intéresse au risque de perception comme authentique ou véridique, il s'en infère que l'objectif est de faire la part entre le contenu réel et celui qui est généré par un système d'IA, laissant toutefois de côté le contenu généré par d'autres outils susceptibles de produire les mêmes effets, même si les moyens nécessaires sont plus importants. La lutte contre les faux relève néanmoins du champ du DSA (voy. chapitre 7) et le législateur a fait le choix d'inscrire l'étiquetage dans le seul cadre du RIA. Ce choix paraît opportun, car une identification qui porterait sur la fiabilité du contenu soulèverait des questions pratiques de mise en œuvre bien trop complexes pour assurer l'effectivité de sa mise en œuvre, alors même que les déployeurs de **contenus synthétiques manipulatoires** risquent d'être peu enclins à mettre en œuvre une information objective sur la génération ou manipulation par IA.

L'étiquetage ainsi prévu devrait faire l'objet, dans le cadre de la mise en œuvre du code de bonnes pratiques, d'un modèle commun. Pour permettre l'appropriation de cette étiquette, et après une phase nécessaire d'expérimentation et d'évaluation du signe retenu par le code de bonnes pratiques, **la mission estime qu'il serait pertinent qu'un modèle uniformisé obligatoire soit retenu à brève échéance** dans le cadre d'un acte exécution, ainsi que le prévoient les articles 50, paragraphe 7, et 56, paragraphe 6, du RIA. Elle regrette à cet égard que la deuxième version du projet se traduise par un engagement moindre des parties prenantes sur ce point, pourtant essentiel.

Cependant, compte tenu de la diversité des pratiques en matière d'utilisation de l'IA, des types de contenus susceptibles d'être qualifiés d'hypertrucages au sens du RIA et des enjeux liés aux fausses informations, **un tel étiquetage devrait permettre de connaître réellement ce qui a été manipulé par IA**. La lecture des dispositions en cause va en ce sens, puisqu'il est prévu, à titre principiel une indication « *que les contenus ont été générés ou manipulés par IA* » et à titre d'exception pour les créations, la divulgation de la seule « *existence* » de tels contenus : la comparaison des deux rédactions montre qu'en principe, il doit être possible de connaître ce qui a été manipulé. L'information du public ne peut en effet être réelle si elle se limite à la mention « IA », tant un hypertrucage qui se limite à coloriser un film d'archives ne présente pas les mêmes enjeux de connaissance que l'utilisation sans l'accord de l'intéressé d'une voix pour un doublage ou la génération d'une fausse vidéo de reddition d'un chef d'Etat en guerre. A cet égard, **la mission estime que les solutions**, prévues par le premier projet de code de bonnes pratiques, **permettant, à partir de l'étiquette, d'obtenir plus d'informations sur le niveau d'utilisation de l'IA apparaissent particulièrement pertinentes et devraient être encouragées à titre complémentaire.** Si, compte tenu de la diversité des acteurs, une pluralité de solutions peuvent être laissées, **il apparaît indispensable que la portée réelle de l'obligation soit claire pour tous.** Sur ce point également, la mission ne peut que regretter le recul opéré dans le deuxième projet de code de bonnes pratiques, mais elle constate que la version finale du code prévoit trois logos uniformes : un générique « IA » et deux autres qui permettent de préciser si le contenu est généré ou seulement modifié par IA, ce qui constitue une première étape notable.

d) Exceptions

L'article 50, paragraphe 4, prévoit deux exceptions, dont l'une sort du champ de la mission, concernant les hypertrucages utilisés à des fins pénales. En revanche, la seconde exception, qui limite l'ampleur de l'obligation d'étiquetage intéresse directement les domaines culturels et créatifs : il s'agit des contenus générés ou manipulés par IA s'insérant dans « *un œuvre ou [un] programme manifestement artistique, créatif, satirique, fictif ou analogue* ».

Cette exception, dont il faut relever qu'elle ne consiste en réalité qu'en un aménagement de l'obligation d'étiquetage, a suscité de nombreux commentaires car la détermination de son champ d'application est essentielle pour l'effectivité de l'article 50, paragraphe 4, dans son ensemble, les risques d'une interprétation extensive, vidant de sa substance cette disposition, ayant été soulignés[364], en particulier s'agissant de l'humour ou de la fiction. Le RIA présente à cet égard des éléments de prévention de ce risque, qui tiennent, précisément, à la circonstance que cette dimension artistique ou humoristique n'induit qu'un aménagement de l'obligation et à l'ajout de l'adverbe « *manifestement* », qui renvoie à une idée d'*évidence*[365] : ce caractère doit être assumé par le contenu diffusé pour bénéficier de l'aménagement. Pour autant, comme le soulignent les lignes directrices de la Commission sur l'article 50, cette évidence est,

[364] Not. Mariëtte van Huijstee *et al*., préc. ou Mateusz Łabuz, préc.
[365] Terme qui apparaît dans la version anglaise avec « *evidently* », et non « *obviously* », en cohérence avec la version allemande qui emploie le terme « *offensichtlich* ».

notamment pour l'humour, dépendante du contexte[366] (la même vidéo aura ce caractère de manière évidente si elle est diffusée sur une chaîne satirique, mais ne l'aura plus lorsqu'elle sera relayée, sans mention du contexte de sa diffusion initiale, par une chaîne conspirationniste) et s'avère relative d'une personne à l'autre.

Cette exception doit être interprétée à la lumière des libertés qu'elle entend protéger, ainsi que le précise le considérant 134 : « *Le respect de cette obligation de transparence ne devrait pas être interprété comme indiquant que l'utilisation du système d'IA ou des sorties qu'il génère entrave le droit à la liberté d'expression et le droit à la liberté des arts et des sciences garantis par la Charte, en particulier lorsque le contenu fait partie d'un travail ou d'un programme manifestement créatif, satirique, artistique, de fiction ou analogue, sous réserve de garanties appropriées pour les droits et libertés de tiers* ». Ce cadre appelle deux observations.

En premier lieu, contrairement à ce qui avait été initialement envisagé par la proposition de RIA, cette exception n'emporte pas disparition de toute obligation : il s'agit d'un aménagement, dont la portée apparaît très relative si l'étiquetage proposé par le code de bonne conduite est appliqué à titre principiel, tant celui-ci a été conçu, d'emblée, pour ne pas entraver l'affichage et la jouissance d'un contenu. C'est le deuxième volet de l'étiquetage, permettant de déterminer plus précisément ce qui a été généré par IA, qui n'est pas applicable pour ces contenus et **il semble important que les potentiels bénéficiaires de cet aménagement soient bien informés que l'utilisation de l'IA pour générer un contenu utilisé dans leur création doit être portée à la connaissance du public**, bien qu'une information minimale soit seulement requise.

En second lieu, dès lors qu'elle vise à préserver les droits et libertés, cette exception a vocation à faire l'objet d'une interprétation ouverte. Néanmoins, ces droits et libertés sont encadrés et l'exception ne peut être invoquée que pour autant que, d'une part, le cadre du déployeur soit *manifestement* lié à ces libertés : diffuser une vidéo mettant en scène un responsable politique dans une scène de film connue peut être considéré comme relevant de la liberté d'expression (qui plus est celle d'un homme politique[367]) et d'une forme de satire ou, à tout le moins de création, pour autant que la mise en scène se prête à cette appréciation et qu'il n'y ait pas détournement ou utilisation abusive[368] (il ne peut s'agir, par exemple, de procéder à une mise en scène compromettante, auquel cas il y aurait diffamation). L'analyse est nécessairement contextuelle, comme l'est déjà l'appréciation des abus de la liberté d'expression[369]. On peut d'ailleurs également trouver une source d'inspiration à cet égard dans l'appréhension de l'exception de parodie en droit de la propriété intellectuelle : bien que celle-ci s'apprécie plutôt *in abstracto*, elle tient compte du risque de confusion et de la recherche d'appropriation, ainsi que de l'intention de dénigrement[370]. De même, le délit d'usurpation d'identité de l'article 226-4-1 s'entend sous réserve de la liberté d'expression[371].

D'autre part, bien que ce soit le seul considérant 134 qui l'explicite, **cette exception est strictement cantonnée à son objet : un aménagement de l'obligation de transparence posée par l'article 50, paragraphe 4. Elle ne saurait en aucune manière être interprétée comme permettant à un déployeur de diffuser tout contenu créatif, au seul motif qu'il porterait la mention « généré par IA », notamment s'il méconnaît d'autres dispositions applicables ou s'il porte atteinte aux droits des tiers, y compris en matière de propriété littéraire et artistique**, celle-ci constituant une limite légitime

[366] Voy. sur ces aspects, Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1.

[367] « *Par nature, le discours politique est source de polémiques et est souvent virulent. Il n'en demeure pas moins d'intérêt public, sauf s'il dégénère en un appel à la violence, à la haine ou à l'intolérance. Là se trouve la limite à ne pas dépasser* » - CEDH, gr. ch., 15 octobre 2015, *Perinçek c. Suisse*, n° 27510/08, paragr. 231 ; CEDH, 11 juin 2020, *Baldassi et autres c. France*, nos 15271/16 et autres, paragr. 79.

[368] CEDH, 23 juillet 2009, *Hachette Filipacchi Associés (Ici Paris) c. France*, n° 12268/03, paragr. 46.

[369] Voy. par ex. CEDH, 13 décembre 2005, *Wirtschafts-Trend Zeitschriften-Verlagsgesellschaft mbH c. Autriche (n° 3),* nos 66298/01 et 15653/02, paragr. 47

[370] Sur ces points, voy. Pierre-Yves Gautier, Nathalie Blanc, *Droit de la propriété littéraire et artistique*, Paris, LGDJ-Lextenso, 2003, 2de éd., p. 317.

[371] Cass. Crim, 16 novembre 2016, pourvoi n° 16-80.207, inédit.

à la liberté d'expression[372], les autorités compétentes disposant d'une marge d'appréciation particulièrement importante pour faire la balance entre ces deux intérêts.

En tout état de cause, **il est de la responsabilité du déployeur de décider s'il relève de cette exception, au risque, le cas échéant, d'encourir les sanctions prévues par l'article 99 du RIA**. Toutefois, ce régime n'est adapté que pour les professionnels, en cohérence avec la notion de déployeur, et c'est vraisemblablement dans le droit général, abordé ci-après, qu'il faudra identifier une voie permettant de responsabiliser les particuliers diffusant des hypertrucages dépassant les limites des libertés garanties, tout en étant dispensés de l'obligation d'étiquetage.

Par ailleurs, **le second alinéa de l'article 50, paragraphe 4, prévoit un régime spécifique pour les textes qui, bien que ne relevant pas de la définition de l'hypertrucage retenue par le RIA, sont considérés par le considérant 134 comme présentant des enjeux proches** : il s'agit des **textes publiés dans le but d'informer le public sur les questions d'intérêt public**. Une même obligation d'étiquetage est prévue, mais **avec une exception concernant le contenu qui « *a fait l'objet d'un processus d'examen humain ou de contrôle éditorial* *et lorsqu'une personne physique ou morale assume la responsabilité éditoriale de la publication du contenu* »**. L'effet de cette exception est plus absolu que pour les hypertrucages au sens du RIA : le législateur européen dispense alors de tout étiquetage.

La lecture de cette exception relève également d'une **démarche de responsabilisation des acteurs et celle-ci devrait guider son interprétation**. C'est en effet l'existence d'une supervision humaine, avec la possibilité d'identifier une personne assumant la responsabilité de la publication – et le cas échéant les poursuites afférentes. Cette exception couvre évidemment, en premier lieu, la presse. Cela apparaît particulièrement pertinent au regard des enjeux de désinformation, à l'heure où les faux sites se multiplient[373] et alors que même les acteurs de la presse peuvent s'avérer victimes de faux contenus[374]. Il faut souligner que cette dérogation ne s'applique qu'aux textes, à l'exclusion des autres contenus, quand bien même ceux-ci bénéficient également de la liberté de la presse, avec un enjeu de responsabilité qui est même considéré comme accru par rapport à la presse écrite[375] : l'application dans ce cas de l'alinéa 1er de l'article 50, paragraphe 4, s'avère cohérente.

Les éditeurs de presse bénéficient à cet égard d'un régime de responsabilité spécifique, contrepartie de leur rôle dans les sociétés démocratiques : la Cour européenne des droits de l'homme leur assigne, en contrepartie, des devoirs et responsabilités, tenant notamment à la nécessité d'agir de bonne foi sur la base de faits exacts, en fournissant des informations fiables et précises, conformes aux règles déontologiques[376]. La circonstance qu'un éditeur de presse publie un contenu généré par IA ne le dispense pas de s'assurer du respect de ces principes et il engage sa responsabilité en cas de méconnaissance, sans qu'une comparaison puisse être raisonnablement faite avec un forum : dans ce cas, la CEDH a en effet admis qu'on ne pouvait exiger du site de presse qu'il exerce un contrôle éditorial sur tous les commentaires[377], mais une telle plateforme ne relèverait pas, selon l'analyse de la mission, du champ de l'exception applicable ici, faute, précisément, d'un tel contrôle. La contrepartie de cette responsabilité – et de l'identification d'un responsable résultant des dispositions de la loi du 29 juillet 1881 – permet d'alléger l'exigence de

[372] CEDH, déc., 19 février 2013, *Neif et Sunde Kolmisoppi c. Suède*, n° 40397/12. Rappr. CEDH, 10 janvier 2013, *Ashby Donald et autres c. France*, n° 36769/08, paragr. 40-41. Egalement CJUE, gr. ch., 3 septembre 2014, *Deckmyn et Vrijheidsfonds*, C-201/13.

[373] « Les faux sites d'information générés par IA récoltent déjà une audience significative, selon une étude Médiamétrie », *Le Monde*, 19 décembre 2025. Voy. également « NewsGuard évalue un réseau de 139 faux sites français liés au Kremlin », *NewsGuard*, 16 octobre 2025, ainsi que les bilans de l'ARCOM sur la lutte contre la manipulation de l'information sur les plateformes en ligne.

[374] « Manipulierte Fotos in Berichten zu Iran entdeckt », *Der Spiegel*, 11 mars 2026.

[375] Par ex. CEDH, 21 février 2017, *Orlocskaya Iskra c. Russie*, n° 42911/08, paragr. 109.

[376] CEDH, gr. ch., 21 janvier 1999, *Fressoz et Roire c. France*, n° 29183/95, paragr. 54 ; gr. ch., 7 février 2012, *Axel Springer AG c. Allemagne*, n° 39954/08, paragr. 93 ; gr. ch., *Bladet Tromsø et Stensaas c. Norvège*, n° 21980/93, paragr. 65 ; 20 juin 2023, *Karaca c. Turquie*, n° 25285/15, paragr. 157.

[377] CEDH, gr. ch., 16 juin 2015, *Delfi AS c. Estonie*, n° 64569/09, paragr. 112-113.

transparence. Le RIA est, sur ce point, en cohérence avec le règlement européen sur la liberté des médias adopté la même année[378], et qui définit expressément la notion de responsabilité éditoriale[379].

Dans ce cadre, **la mission partage l'idée discutée lors de la préparation du code de bonnes pratiques selon laquelle ces éditeurs, dès lors qu'ils satisfont aux conditions d'une telle qualification et pour les seules publications relevant de la presse, devraient bénéficier d'une présomption de conformité au RIA**, évitant une analyse de chaque article au regard de l'apport des systèmes d'IA.

En outre, les auditions menées par la mission ont souligné combien la transparence constituait un enjeu fondamental pour l'utilisation de l'IA dans le domaine de l'information, mais, compte tenu du développement des fausses informations, la crédibilité des organes de presse repose largement sur leur capacité à garantir l'authenticité des contenus. **Plus que la transparence sur l'utilisation de l'IA, c'est alors la vérification de l'information diffusée et la preuve de cette vérification qui constitue un enjeu**. Des outils ont été mis en place – ce dont certains acteurs auditionnés par la mission ont témoigné en évoquant les solutions auxquelles elles font appel comme Blue efficence – et, entre autres, Reporters sans frontières opère une certification des organes de presse.

Cette démarche ne peut qu'être encouragée mais la mission est toutefois consciente de l'équilibre délicat que ce type d'appréciation suppose dans le contexte du développement de l'intelligence artificielle[380] ; la réflexion sur ce point devrait être poursuivie dans la continuité des états généraux de l'information et du rapport de mission des députés Arthur Delaporte et Stéphane Vojetta[381].

5.2. Hypertrucages, pratiques interdites, systèmes d'IA à haut risque et modèles d'IA à usage général présentant un risque systémique dans le RIA

L'article 50, en étant consacré à certains systèmes d'IA, notamment ceux générant des contenus qualifiés d'hypertrucage, pourrait laisser penser qu'il s'agit d'une disposition spéciale, seule applicable à ces systèmes. Une telle interprétation serait néanmoins préjudiciable à la régulation des contenus générés par IA[382] et tel n'est pas le cas puisque, d'une part, l'architecture du règlement traite les risques par ordre décroissant et, d'autre part, le paragraphe 6 de l'article 50 réserve l'application du chapitre III. En outre, cela est précisé par les considérants du règlement : le considérant 132 indique en effet que ces systèmes générant du contenu « *devrai[ent] donc être soumi[s] à des obligations de transparence spécifiques sans préjudice des exigences et obligations relatives aux systèmes d'IA à haut risque*[383] ». Le règlement est toutefois, en lui-même, dans sa rédaction initiale, silencieux s'agissant des pratiques interdites par le chapitre II, alors même que les hypertrucages peuvent être un outil à travers lequel certaines de ces pratiques prohibées sont déployées. Le risque de manipulation est, au demeurant, ce qui justifie l'attention particulière portée aux hypertrucages et **il semble que ce soit bien ce caractère plurifactoriel des risques, dont l'ampleur dépend moins de la technique elle-même que de l'usage qui en est fait**[384]**, qui ait conduit à cette catégorisation propre et distincte des hypertrucages dans l'approche du RIA, sans que cela exclut la prise en compte des autres risques** que le règlement prend en compte au titre de chacune des catégories[385]. On observera au demeurant que, dans le cadre de la proposition de règlement dit « omnibus » en cours de négociation entre le Parlement européen et le Conseil, il est envisagé[386] de prohiber de manière

---

[378] Règlement (UE) 2024/1083 du Parlement européen et du Conseil du 11 avril 2024 établissant un cadre commun pour les services de médias dans le marché intérieur et modifiant la directive 2010/13/UE.

[379] Art. 2 : « *8) «responsabilité éditoriale»: l'exercice d'un contrôle effectif tant sur la sélection des programmes ou des publications de presse que sur leur organisation, aux fins de la fourniture d'un service de médias, indépendamment de l'existence d'une responsabilité en vertu du droit national à l'égard du service fourni* ».

[380] Voy. par ex. récemment le communiqué du Syndicat de la presse indépendante d'information en ligne du 3 décembre 2025, « Agrément des sites d'information générés par IA : le Spiil sonne l'alarme ».

[381] *Influence et réseaux sociaux. Face aux nouveaux défis, structurer la filière de la création, outiller l'Etat et mieux protéger*, rapport au Premier ministre, janvier 2026, recommandations n° 47, p. 87, et n° 49, p. 89.

[382] Théo Antunes, « Prévenir l'inacceptable : le rôle du règlement européen dans la gouvernance des pratiques prohibées liées à l'intelligence artificielle générative », *Rev. de l'Union européenne*, 2025, p. 609.

[383] Soulignement ajouté.

[384] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, Yang Liu, « Countering Malicious DeepFakes: Survey, Battleground, and Horizon », *International Journal of Computer Vision*, 2022, Vol. 130, pages 1678.

[385] Sur l'historique et les enjeux du choix fait à cet égard par le législateur européen, voy. Mateusz Łabuz, « Regulating Deep Fakes in the Artificial Intelligence Act », *Applied Cybersecurity & Internet Governance*, vol. 2, n° 1, 2023, p. 1.

[386] A la date de la rédaction du présent rapport, il est fait référence à la version du texte telle que présente dans le mandat du Conseil adopté par le COREPER Le 13 mars 2026 en vue du trilogue (doc. 7322/26 du 13 mars 2026).

plus explicite des pratiques qui se traduisent, dans les faits, par la génération et la diffusion d'hypertrucages (notamment les hypertrucages à caractère sexuel représentant une personne sans son consentement et hypertrucage à caractère pédopornographique).

Les lignes directrices de la Commission européenne sur les pratiques interdites en matière d'intelligence artificielle[387] considèrent qu'**un seul et même comportement interdit puisse constituer une violation de plusieurs dispositions du RIA**, en prenant pour exemple le non-étiquetage des hypertrucages et la technique trompeuse (paragr. 56). Toutefois, la Commission s'en tient largement à considérer que le marquage visible réduit le risque de tromperie (paragr. 71), mais que l'interdiction de l'article 5 s'applique aux contenus présentant des informations fausses ou trompeuses d'une manière qui vise ou a pour effet de tromper des personnes et d'altérer leur comportement (paragr. 72). On ne peut toutefois que regretter que ces lignes directrices se limitent à identifier la question, sans pour autant apporter de réelle réponse à cette question complexe et essentielle pour les acteurs, y compris dans les champs culturels et créatifs qui bénéficient assez largement de l'aménagement de l'obligation d'étiquetage au titre de l'article 50, paragraphe 4.

Or, la mission estime qu'une réponse claire s'impose et qu'il ne peut y avoir de doute : l'intention du législateur européen a été d'interdire l'ensemble des pratiques mentionnées à l'article 5, y compris les hypertrucages susceptibles de relever des cas qui y sont énumérés. Au demeurant, les termes de l'article 5 ont été renforcés au cours des débats sur le projet de texte, afin de renforcer la lutte contre les utilisations à des fins de manipulation[388]. Dès lors, si l'article 50, paragraphe 6, se borne à réserver le chapitre III, c'est qu'il ne s'intéresse qu'aux pratiques autorisées et prévoit que se cumulent les obligations de ce chapitre et du chapitre IV lorsqu'un système d'IA relève de ces deux champs. Cela ne condamne pas les systèmes d'IA permettant de générer des hypertrucages, puisque, outre la mise sur le marché et la mise en service, qui visent le principe même de la mise à disposition de ce système, la simple *utilisation* d'un tel système pour une des fins prohibées est interdite. C'est pour cela que les lignes directrices soulignent d'ailleurs que « *L'interdiction s'applique par conséquent à la fois aux fournisseurs et aux déployeurs de systèmes d'IA, chacun dans le cadre de leurs responsabilités respectives* » (paragr. 62). Bien que les utilisateurs à des fins privées ne soient pas concernés, il faut relever que, dans ce cadre, **il appartient tout autant aux fournisseurs qu'aux déployeurs de veiller aux usages interdits de l'IA**, ce qui est particulièrement sensible pour les hypertrucages. **Il appartient à ce titre aux fournisseurs de mettre en place des procédures permettent de prévenir la génération par les systèmes qu'ils mettent à disposition de contenus relevant de pratiques interdites**[389]. Concernant les hypertrucages à caractère sexuel non consentis, la proposition de règlement omnibus précitée précise d'ailleurs les obligations des fournisseurs à cet égard.

S'agissant des systèmes d'IA à haut risque, qui sont expressément réservés, la mission constate que la qualification d'un haut risque relève de la prise en compte d'hypothèses listées en particulier à l'annexe III qui, pour l'essentiel, n'ont pas d'incidence sur les secteurs créatifs et culturels. Elle n'estime donc pas utile de formuler d'observations sur ce terrain[390].

En revanche, la question de l'articulation entre les obligations de l'article 50 et les dispositions du chapitre V sur les modèles d'IA à usage général présente des enjeux particuliers pour ces secteurs[391]. **La mission considère en effet que les chapitres IV et V sont susceptibles de s'appliquer cumulativement**, nonobstant le silence des dispositions du règlement (en dehors de l'article 50, paragraphe 2), dès lors que le système d'IA permettant de générer un hypertrucage répond, par ailleurs, à la définition du système d'IA

[387] Doc. C(2025) 5052 final du 29 juillet 2025.

[388] En réponse notamment aux critiques formulées par certains auteurs, dont Nathalie A. Smuha, Emma Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli, Karen Yeung, *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act* (5 août 2021).

[389] Théo Antunes, art. préc.

[390] Bien qu'il ait pu être proposé de considérer comme étant à haut risque les systèmes d'AI utilisés pour des hypertrucages malveillants (Felipe Romero Moreno, « Generative AI and deepfakes: a human rights approach to tackling harmful content », *International Review of Law, Computers & Technology*, vol. 38, 2024, p. 297), mais en l'état de l'annexe III une telle lecture paraît difficile à suivre.

[391] Sur le régume de ces modèles et systèmes d'IA à usage général, voy. Cécile Pellegrini, « Modèles et systèmes d'IA à usage général (GPAI) dans l'AI Act », *RTD eur.*, 2024, p. 587.

à usage général résultant des points 63[392] et 66[393] de l'article 3. Au demeurant, le considérant 137 est plus explicite sur ce point que l'article 50 : alors que ce dernier se borne à indiquer que « *les paragraphes 1 à 4 n'ont pas d'incidence sur les exigences et obligations énoncées au chapitre III et sont sans préjudice des autres obligations de transparence prévues par le droit de l'Union ou le droit national pour les déployeurs de systèmes d'IA* », les motifs énoncent que le respect des obligations de transparence applicables aux systèmes d'IA relevant du RIA « *ne devrait pas être interprété comme indiquant que l'utilisation du système d'IA ou de ses sorties est licite en vertu du présent règlement ou d'autres actes législatifs de l'Union et des Etats membres et devrait être sans préjudice d'autres obligations de transparence pour les déployeurs de systèmes d'IA prévues par le droit de l'Union ou le droit national* ».

Or, en application de l'article 53, les fournisseurs de modèles d'IA à usage général sont tenus à plusieurs obligations et, à ce titre, ils sont tenus d'autres obligations de transparence, telles que définies aux *a*, *b* et *d* du paragraphe 1 (que le considérant 101 qualifient d'obligations de transparence), qui peuvent se cumuler avec celles de l'article 50. A ce titre, ils doivent notamment mettre en place « *une politique visant à se conformer au droit de l'Union en matière de droit d'auteur et droits voisins, et notamment à identifier et à respecter (…) une réservation de droits exprimée conformément à l'article 4, paragraphe 3, de la directive (UE) 2019/790* » (c du paragraphe 1).

Cette disposition est importante quant à la prise en compte des enjeux des secteurs culturel et créatifs au regard du développement des hypertrucages. Un nombre substantiel des systèmes d'IA permettant de générer de tels contenus relèvent en effet également de la qualification de système d'IA à usage général et sont soumis à ces obligations de l'article 53, dont le respect des droits réservés dans la fouille de textes et de données prévue par la directive 2019/790. Pour autant, le champ d'application de l'article 53 ne recouvre pas entièrement celui de l'article 50, alors qu'il aurait été pertinent que les dispositions de l'article 53 soient également applicables aux systèmes d'IA générative, de sorte qu'un fournisseur d'un tel système d'IA devrait en principe s'assurer qu'en phase d'apprentissage, les réservations de droit ont été respectées. Le code de bonnes pratiques qui a été adopté à cet égard contient un chapitre dédié à la protection des droits de propriété intellectuelle[394] qui va au-delà de la seule protection des réservations de droit mais vise aussi à limiter les risques d'atteinte à ces droits. A défaut d'extension des obligations de l'article 53 à l'ensemble des systèmes d'IAG, **la mission constate que la Commission lui a, à ce jour, conféré un statut de code pouvant faire l'objet d'une adhésion volontaire**[395] **et estime opportun qu'il puisse être rendu obligatoire dans les conditions prévues à l'article 56, paragraphe 6, alinéa 2**. Elle relève toutefois que la base légale pourrait s'avérer insuffisante pour rendre obligatoire les éléments relatifs à la prévention des atteintes aux droits, de sorte qu'une évolution législative pourrait s'avérer nécessaire pour prendre pleinement en compte ces enjeux.

Néanmoins, il convient de rappeler que le chapitre V du RIA identifie, au sein des modèles d'IA à usage général ceux qui présentent un « *risque systémique* », entendu 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* » (art. 3, point 65). Or, dès lors que les risques d'atteinte aux droits fondamentaux entrent dans le champ de ces risques systémiques, **la mission considère qu'il y a lieu de tenir spécialement compte dans l'évaluation des risques systémiques du risque d'atteinte aux droits de propriété intellectuelle**, qui sont une composante du droit fondamental de propriété[396]. Il

---

[392] « 63) «modèle d'IA à usage général», 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 ».

[393] « 66) «système d'IA à usage général», 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 ».

[394] https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[395] Doc. C(2025) 5361 final du 1er août 2025.

[396] Art. 17, paragr. 2, de la Charte des droits fondamentaux de l'UE ; CEDH, gr. ch., 11 janvier 2007, *Anheuser-Busch Inc. c. Portugal*, n° 73049/01, paragr. 72. En droit interne, en revanche, après avoir retenu le même principe d'inclusion dans le droit de propriété (déc. n° 2006-540 DC du 27 juillet 2006), le Conseil constitutionnel pose désormais un objectif

s'en infère alors une responsabilisation accrue des fournisseurs de modèles d'IA à usage général, en cohérence avec le code de bonnes pratiques relatif à l'article 53, afin d'atténuer ce risque.

Ce raisonnement devrait, au surplus, être étendu également à deux autres droits fondamentaux qui sont directement questionnés par les hypertrucages générés par ces modèles d'IA à usage générale : **le droit au respect de la vie privée et le droit à la protection des données personnelles**[397]. **Il apparaît ainsi indispensable à la mission que la Commission européenne se saisisse de ces enjeux et les intègre**, contrairement à la situation actuelle[398], **dans les lignes directrices et codes de bonne conduite** relatifs à la mise en œuvre du RIA, à défaut d'une clarification des enjeux dans les dispositions du règlement lui-même. Cela serait parfaitement cohérent avec l'approche du RIA, qui vise à protéger les droits fondamentaux (art. 1er, paragr. 1er, du RIA).

autonome de valeur constitutionnelle de sauvegarde de la propriété intellectuelle (déc. n° 2020-841 QPC du 20 mai 2020).

[397] Art. 7 et 8 de la Charte des droits fondamentaux de l'UE.

[398] Concernant les modèles d'IA à usage général : doc. C(2025) 5045 final.

6. MODALITES TECHNIQUES D'IDENTIFICATION ET DE DETECTION

L'esprit général du RIA, on l'a vu, est d'encourager la transparence essentiellement grâce à des solutions techniques. La législation européenne impose en effet aux fournisseurs d'IA un ensemble d'obligations techniques de marquage pour tous les contenus synthétiques (y compris les écrits), ainsi que des obligations d'étiquetage, d'information du public, pour les seuls hypertrucages tels que définis par le texte. L'article 50 fait peser l'essentiel de ces obligations de transparence sur les acteurs de l'IA, fournisseurs ou déployeurs, au moment de la génération des contenus.

Pour les diffuseurs en ligne tels que Spotify, Deezer ou, plus largement, les acteurs traditionnels de la diffusion (éditeurs de télévision notamment), ce ne sont pas des obligations légales mais la survie de leur modèle économique qui est en jeu et qui les pousse à mobiliser des techniques de détection.

Sans attendre l'entrée en vigueur du RIA, des groupes notamment de presse ou audiovisuels se sont dotés de Chartes pour définir leurs propres conditions et modalités de recours à l'IA Générative dans les contenus d'information et le degré de transparence nécessaire auprès du public. D'autre part, lorsque ces groupes sont eux-mêmes victimes d'hypertrucages, ils se sont équipés d'outils pour détecter automatiquement et tenter de faire supprimer ceux problématiques sur les réseaux sociaux (infox, fausses publicités destinées à de l'escroquerie…). La résistance des réseaux sociaux à faire supprimer, une fois détectés, ces hypertrucages, dont la viralité peut être très importante, au nom notamment de la liberté d'expression est cependant propice à leur propagation au détriment des personnes physiques et morales. L'anonymat des comptes et des auteurs de publications sur les réseaux sociaux est un élément qui favorise ces comportements et les difficultés à lutter efficacement pour protéger les victimes de ces agissements.

De nombreuses entreprises hébergeuses ou fournisseuses de contenus culturels doivent donc investir de manière croissante dans la détection, la suppression et la gestion des contenus hypertruqués. You Tube a déployé en 2023 un outil qui exige des créateurs qu'ils indiquent lorsque le contenu qu'ils mettent en ligne est modifié de manière significative ou généré synthétiquement et qu'il semble réaliste [399]. Ce nouvel outil est intégré au flux de mise en ligne existant, ce qui permet aux créateurs d'ajouter facilement cette mention. Si l'utilisateur ne le fait pas, la vidéo est tout de même analysée et une détection automatique conduit à l'intégration d'un label. En principe, si l'utilisateur omet de déclarer l'utilisation d'IA de multiples fois, l'entreprise peut supprimer la vidéo, voire, supprimer les revenus du créateur. De même sur Gemini, depuis août 2024, la génération d'images bloque les contenus nuisibles, trompeurs ou sexuellement explicites[400]. Elle restreint la création d'images photoréalistes de personnes identifiables et applique systématiquement des métadonnées et le filigrane numérique SynthID.

Certaines entreprises ont choisi de prendre des mesures et de communiquer sur leur démarche auprès du public. Dans la musique les stratégies des acteurs de la diffusion en flux (*streaming*) sont diverses. Deezer a développé et breveté sa propre technologie interne capable de détecter les contenus issus des modèles Suno et Udio et autres générateurs majeurs. L'outil ne se contente pas de de comparer l'audio à une base de données (« *fingerprinting* » classique) mais analyse la structure intrinsèque du signal pour détermine une probabilité d'origine synthétique. Un étiquetage systématique des morceaux détectés est effectué (mention « IA » sur certains morceaux)[401]. Deezer commercialise de plus son outil « radar » auprès d'autres organisations comme la SACEM. Le contexte quantitatif est significatif : Deezer reçoit environ 50 000 pistes 100 % IA par jour (soit 34 % des nouveaux chargements, contre 10 000 par jour en janvier 2025), bien que ces contenus ne représentent que moins de 1 % des écoutes totales, dont 70 % sont frauduleuses (voir supra). Deezer a d'ores et déjà exclu les pistes IA de ses playlists éditoriales et algorithmiques. L'outil vise donc à les exclure des recommandations et de la monétisation, et non à les supprimer. Apple qui se revendique « marque premium » se concentre sur la curation humaine pour se différencier des contenus de basse qualité qui prolifèrent ailleurs avec l'IA. Spotify joue la carte de la segmentation des marchés ; elle a industrialisé des filtres de spams spécifiques aux pistes générées par IA en rachetant des *starts-ups* spécialisées et parallèlement propose en 2026 une offre premium de grande qualité technique et garantie authentiquement humaine. L'alliance « *Music Fights Fraud* » regroupe des acteurs de l'industrie pour

[399] The YouTube team, 'How We're Helping Creators Disclose Altered or Synthetic Content', Blog.Youtube, 2024, https://blog.youtube/news-and-events/disclosing-ai-generated-content/.
[400] Google, 'Règlement Sur Les Utilisations Interdites de l'IA Générative', 2024, https://policies.google.com/terms/generative-ai/use-policy.
[401] Afchar et al., 'Detecting Music Deepfakes Is Easy but Actually Hard'.

mutualiser les efforts de détection. Ces coûts incluent le développement de technologies de détection, le tatouage numérique (*watermarking*), les procédures juridiques et la modération humaine et technique, représentant des dépenses récurrentes qui réduisent la rentabilité des acteurs. Par ailleurs des faux positifs se multiplient alourdissant encore le coût de la détection. Des sorties légitimes d'artistes humains sont en effet bloquées par la détection d'anomalies. Or chaque faux positif génère un ticket de support client nécessitant une intervention humaine pour vérification.

Pour de nombreuses plateformes qui mettent des contenus en ligne à la disposition du public, les contenus synthétiques peuvent apparaître, tout du moins à court terme, positivement en produisant de l'engagement et des revenus publicitaires alors que la détection est un centre de coût important. Cependant la séparation du bon grain de l'ivraie apparaît à plus long terme comme la seule voie pour restaurer une économie de la confiance aussi bien au niveau du public, des partenaires éditeurs ou producteurs ou encore des annonceurs publicitaires qui exigent des garanties d'audience humaine.

C'est pourquoi cette partie propose d'expertiser les outils de détection que peuvent mettre en œuvre les acteurs, qu'ils agissent pour des raisons de conformité à la législation ou pour des raisons de survie de leur modèle économique.

La recherche de reconnaissance de contenus altérés ne commence pas avec l'IA générative ; elle s'inscrit dans un continuum d'outils techniques mobilisés pour authentifier les images et vidéos comme l'ont montré les travaux de Koopman *et al.* (2018) sur l'analyse PRNU (non uniformité de la réponse photo). La détection des contenus synthétiques et des hypertrucages est devenue plus récemment un enjeu majeur ; les domaines de recherche et de développement sont encore relativement nouveaux, et l'éventail des solutions techniques disponibles évolue rapidement. Les propriétés et limites des techniques de création d'hypertrucages, qu'ils soient visuels, audio ou textuels déterminent en retour les possibilités de détection des contenus qu'elles produisent. Si les humains ne sont pas capables de distinguer les contenus synthétiques (cf. Chap.4), les machines semblent bien meilleures. De plus, une étape complémentaire consiste à vérifier si des contenus synthétiques imitent une célébrité ou un artiste, ce qui relève de techniques spécifiques de reconnaissance d'identité.

6.1. Architectures techniques de génération d'hypertrucages

La fabrication des hypertrucages repose sur un ensemble de techniques fondées sur des modèles d'IA générative. En pratique, les créateurs s'appuient sur des familles de modèles et de chaînes de traitement plutôt que sur une architecture unique. Dans tous les cas, le processus repose sur un socle commun : un modèle génératif de grande taille, pré-entraîné sur de très nombreuses données (images, vidéos, voix, musique). Ce pré-entraînement, extrêmement coûteux en calcul, n'est réalisé qu'une seule fois (ou un nombre de fois très limité) ; les modèles résultants sont ensuite diffusés, certains en libre accès (Stable Diffusion, Flux, etc.)[402] et servent de fondation à l'ensemble des usages en aval.

C'est l'étape suivante, la spécialisation du modèle vers une identité cible, qui a connu une évolution rapide, abaissant considérablement la barrière technique à la production d'hypertrucages.

- Affinage sur données cibles (*fine-tuning*). Historiquement, la reproduction de l'apparence, du visage ou de la voix d'une personne exigeait d'affiner le modèle sur un jeu de données de la cible — plusieurs centaines d'images pour les premiers autoencodeurs adversariaux (DeepFaceLab), puis une dizaine seulement grâce à des techniques d'adaptation légère comme DreamBooth ou LoRA (Low-Rank Adaptation). Ces adaptateurs, de petite taille, laissent le modèle de base fixe et n'apprennent que les paramètres nécessaires à la capture d'une identité ou d'un style. Ce schéma rend le clonage de nombreux artistes différents peu coûteux : une fois le modèle de base entraîné, l'ajout d'une nouvelle identité ne nécessite qu'un faible investissement computationnel supplémentaire.

- Conditionnement par image de référence (*one-shot*). Plus récemment, des techniques dites « *tuning-free* » (sans affinage) ont supprimé toute nécessité d'affinage sur la cible. Des modules

[402] Il existe au moins trois cas de libre accès : les modèles à poids téléchargeables, donc en accès libre ; des modèles ouvert sous licence (utilisation commerciale interdite) ; des modèles fermés mais accessibles via une plateforme ou une API, avec potentiellement un abonnement.

comme IP-Adapter (Tencent), InstantID (Xiaohongshu) ou les outils de substitution de visage (« *face-swap* ») (SimSwap, FaceFusion) utilisent des réseaux de neurones pré-entraînés une fois pour toutes sur de larges bases de visages, pour extraire une incrustation (*embedding*) d'identité à partir d'une seule photographie de référence et conditionner la génération en temps réel. Il n'y a alors plus aucun entraînement spécifique à la cible : **une seule image suffit, et le résultat est produit en quelques secondes. Le coût marginal par hypertrucage tend ainsi vers zéro.**

Cette évolution, de l'affinage coûteux vers le conditionnement instantané, a des implications directes sur l'échelle de la menace : la production d'hypertrucages n'exige plus ni compétences techniques avancées, ni ressources de calcul significatives, ni accès prolongé à des données de la personne visée. **Trois grandes familles de modèles sous-tendent les hypertrucages contemporains**.

Les **réseaux antagonistes génératifs** (GANs, *Generative Adversarial Networks*) sont composés de deux réseaux de neurones en compétition : un générateur, qui crée de fausses données (images, sons) à partir de bruit aléatoire, et un discriminateur, qui détermine si les données sont réelles ou synthétiques. L'entraînement procède par itérations successives : le générateur produit des données, le discriminateur les évalue, et les deux réseaux s'ajustent en fonction de leurs erreurs respectives. Le processus se répète jusqu'à ce que le générateur devienne suffisamment performant pour que le discriminateur ne puisse plus distinguer le vrai du faux. L'analogie classique est celle d'un faussaire (générateur) et d'un expert en art (discriminateur) dont les compétences progressent mutuellement. Les GANs ont dominé la génération d'hypertrucages entre 2017 et 2022.

Les **modèles de diffusion constituent la technologie dominante depuis 2022–2023.** Au lieu de générer directement une image ou un son, ils partent d'un bruit aléatoire et le « débruitent » progressivement en guidant le processus avec des conditionnements spécifiques (identité cible, pose, expression, texte). Les implémentations les plus récentes, comme Stable Diffusion 3 ou Flux, reposent sur des architectures internes plus puissantes et plus aisément déployables à grande échelle que les premières générations, ce qui explique en partie l'amélioration rapide de la qualité des contenus générés. Ils produisent des résultats plus réalistes et plus cohérents que les GANs (en particulier pour l'éclairage et les textures de peau), mais sont plus lents à la génération.

Les **transformeurs**, architecture initialement développée pour le traitement du langage, se distinguent des deux précédents par leur capacité à traiter simultanément plusieurs types de données (texte, image, son, vidéo) et à apprendre les relations entre elles. Là où un modèle de diffusion part d'un bruit pour reconstruire une image, un transformeur traite l'ensemble des informations disponibles comme un tout cohérent. C'est ce qui permet aujourd'hui de produire des hypertrucages audiovisuels complets (visage, voix et mouvements de lèvres synthétisés simultanément) à partir d'une simple description textuelle ou d'une image de référence. Concrètement, un tel modèle peut prendre une séquence « mélangée » d'unités textuelles et visuelles, aligner des régions d'image sur des expressions linguistiques, et produire en sortie du texte, des images ou des sons, en fonction de la tâche.

Les méthodes diffèrent par ailleurs selon le support visé, qu'il s'agisse de vidéo, d'image, de son ou de texte.

6.1.1. Vidéos et visages

Les hypertrucages visuels sont essentiellement utilisés pour l'échange de visage (*face swapping*) c'est-à-dire le remplacement du visage d'une personne par celui d'une autre et la reconstitution de mouvement via le transfert de l'animation motrice (*reenactment*). Ils servent également à manipuler des attributs (comme l'âge afin de redonner par exemple à un acteur l'apparence de sa jeunesse, le genre, la coiffure ou les émotions d'un visage donné), à animer une image fixe, ou à restaurer et compléter une image ou une vidéo. Dans tous les cas, ces modèles sont entraînés sur de nombreuses identités, puis adaptés ou conditionnés sur une identité cible spécifique à partir d'un nombre relativement limité d'images.

Échange de visage (*face swapping*)

L'échange de visage consiste à sélectionner un visage source (la personne qui joue la scène) et à le remplacer par le visage cible (la célébrité, l'artiste), tout en conservant les expressions, la pose et l'éclairage de la scène originale. L'idée centrale est de « coder » le visage puis de le « réinjecter » sur un autre corps, de sorte que l'expression et les mouvements du visage restent ceux de la vidéo source. Historiquement, cette technique

a été démocratisée par des outils fondés sur les GANs (DeepFaceLab, FaceSwap, Reface, FaceApp). L'essor des modèles à un coup (*one-shot*) comme InsightFace/inswapper_128 a marqué un tournant en permettant un échange de visage convaincant à partir d'une seule image de référence, sans entraînement spécifique. Des outils comme Roop, ReActor et FaceFusion ont ensuite rendu l'échange de visage accessible à un très large public via des interfaces simples intégrées à Stable Diffusion (open source, gratuit, accessible depuis un ordinateur récent ou depuis un simple navigateur web).

Avec les modèles de diffusion, le processus a gagné en réalisme. Au lieu de générer directement l'image finale, on part d'un bruit aléatoire et on le débruite progressivement en guidant le processus avec l'identité cible (*embeddings* faciaux), la pose et l'expression de la source, et parfois un masque indiquant la région du visage à modifier. Le résultat est un visage synthétisé qui correspond à la cible mais épouse très finement les contraintes visuelles de la scène de départ (éclairage, angle de vue, texture de peau), ce qui rend les hypertrucages plus difficiles à détecter par simple inspection visuelle. L'échange de visage par diffusion s'appuie aujourd'hui sur des modules comme InstantID, IP-Adapter ou Layered LoRA, intégrés à Stable Diffusion XL ou Stable Diffusion 3.

Reconstitution de mouvements (*reenactment*)

La reconstitution de mouvements suit la même logique architecturale que l'échange de visage, mais poursuit un objectif différent : transférer les mouvements (expressions, articulation des lèvres, regard) d'un visage source vers un visage cible, qui peut être réel ou déjà synthétique. Techniquement, on extrait d'abord les paramètres de mouvement (codes de pose ou d'expression) à partir de la vidéo source, puis on utilise un modèle générateur (GAN ou diffusion) pour « animer » le visage cible avec ces paramètres. Cette technique permet de transférer les expressions et les mouvements, mais pas le visage ou le corps eux-mêmes. Elle est illustrée historiquement par des systèmes comme Face2Face (Thies et al., 2016, Université d'Erlangen-Nuremberg / Max Planck Institute), Neural Talking Heads (NVIDIA, 2021), et Avatarify, un outil open source développé par Ali Aliev dès 2020, qui permettait déjà à l'époque de substituer un visage en temps réel lors d'appels vidéo à partir d'une seule photo de référence. Des plateformes commerciales comme HeyGen, Synthesia et Veo 3 (Google DeepMind, mai 2025) combinent désormais ces techniques avec la synchronisation labiale à partir d'un texte ou d'un signal audio.

La tendance actuelle est aux architectures hybrides qui combinent les atouts des GANs (rapidité, possibilité de traitement en temps réel) et des modèles de diffusion (réalisme, cohérence), comme VividFace (Shao et al., 2024). Parallèlement, des modèles vidéo natifs ont émergé, qui ne se contentent plus d'un simple échange de visage mais génèrent directement des séquences vidéo cohérentes conditionnées sur un texte, un plan de caméra ou une référence visuelle. OpenAI Sora, Runway Gen-3, Pika, Kling (Kuaishou) et Luma Dream Machine représentent cette nouvelle génération. Ces modèles transforment le paysage des hypertrucages dans l'industrie culturelle en permettant la génération de vidéos entières (et non plus seulement la manipulation d'un visage au sein d'une vidéo existante), ce qui rend la détection considérablement plus difficile. La montée du vidéo-clonage multimodal — synchronisation labiale et mouvement à partir d'un signal audio ou textuel (HeyGen, Synthesia, Veo 3 (Google DeepMind) — ajoute une couche supplémentaire de réalisme en intégrant plusieurs modalités sensorielles dans un flux vidéo cohérent.

En résumé, les GANs permettent des traitements rapides, y compris en temps réel, mais génèrent des artefacts repérables. Les modèles de diffusion sont plus réalistes et plus cohérents (notamment pour l'éclairage et la peau) mais plus lents. Les architectures hybrides cherchent à combiner les atouts des deux approches, et les modèles vidéo natifs repoussent les limites de ce qui peut être généré entièrement par synthèse.

La reconstitution de l'apparence complète d'une personne réelle repose quant à elle, sur la combinaison de plusieurs couches de techniques. Le visage est capturé via une représentation numérique faciale extrait de photographies disponibles (réseaux sociaux, photos de presse, extraits vidéo), suffisamment nombreuses pour couvrir différentes orientations et expressions. La voix est clonée séparément à partir d'échantillons audio, même brefs — des systèmes comme ElevenLabs ou Tortoise-TTS permettent aujourd'hui un clonage vocal convaincant à partir de quelques dizaines de secondes d'enregistrement. Le corps et les mouvements peuvent être transférés depuis une vidéo source d'un acteur différent via les techniques de reconstitution de mouvements décrites ci-dessus. L'ensemble est ensuite assemblé dans un pipeline de synchronisation labiale (HeyGen, Synthesia) qui aligne les mouvements de la bouche sur le signal audio synthétisé. Le

résultat est une vidéo où l'apparence, la voix, les expressions et les mouvements d'une personne réelle sont entièrement synthétisés sans sa participation — à partir de données publiquement accessibles.

6.1.2. Audios et voix

Les hypertrucages audio et musicaux sont essentiellement utilisés pour cloner des voix existantes (le plus souvent de célébrités) et pour imiter le style d'un artiste. Ils peuvent également servir à restaurer ou compléter des œuvres.

Évolution des architectures : des GANs audio aux modèles Codec

Le clonage vocal s'est historiquement appuyé sur des GANs audio, puis sur des modèles de diffusion audio, suivant la même évolution que l'hypertrucage visuel. Néanmoins, ces deux techniques présentent des limites spécifiques pour la voix. Les GANs apprennent surtout à produire une texture sonore réaliste, mais peinent à imiter de façon fiable car ils séparent mal l'identité, la prosodie et le contenu linguistique. Les modèles de diffusion sont plus lents et peuvent manquer de précision sur les détails prosodiques.

De ce fait, l'architecture dominante repose aujourd'hui sur deux alternatives de synthèse vocale neuronale : les modèles de type Codec et les modèles de type VITS (*Variational Inference with adversarial learning for end-to-end Text-to-Speech* – Inférence variationnelle avec apprentissage adverse pour la synthèse vocale de bout en bout). Ces deux approches ont en commun de distinguer explicitement le contenu linguistique, le timbre, la prosodie et la dynamique acoustique. Elles peuvent ainsi reproduire non seulement les caractéristiques acoustiques d'une voix (timbre, formants), mais aussi sa vitesse, son émotion et son accent.

Modèles Codec et grands modèles audio fondamentaux (fondationnels)

Les modèles de type Codec sont pour l'audio ce que sont les grands modèles de langage (LLM) pour le texte. Ils reposent sur la compression de l'audio en unités acoustiques discrètes (les « *tokens* »), ce qui permet de séparer le contenu, le timbre et la prosodie. Cette séparation permet de reproduire fidèlement ces trois dimensions dans le cadre du clonage. Les modèles Codec sont les plus performants pour le clonage vocal ; ils peuvent cloner une voix à partir d'un extrait de quelques secondes seulement.

En pratique, la génération vocale repose aujourd'hui sur de grands modèles d'IA entraînés sur des quantités massives d'enregistrements audio. Des systèmes comme Meta Voicebox, OpenAI Voice Engine, ElevenLabs v3 ou Moshi (2025) sont capables de synthétiser une voix convaincante à partir d'un texte, en reproduisant le timbre, le rythme et les intonations d'une personne cible. Les avancées les plus récentes, comme GLM-TTS (décembre 2025), permettent d'évaluer et d'optimiser la qualité de la voix générée selon plusieurs critères simultanément (naturel, intelligibilité, ressemblance avec la cible) ce qui améliore considérablement le réalisme des voix synthétisées.

Mécanisme de clonage : représentation numérique de locuteur et adaptation

Le mécanisme de clonage vocal repose sur une chaîne technique précise. Un grand modèle de parole est d'abord pré-entraîné sur des centaines ou milliers d'heures de voix variées pour apprendre la correspondance entre texte et sons (phonèmes, prosodie, rythme) et la structure acoustique d'un signal de parole (spectrogrammes, forme des formants, intonation). À ce stade, le modèle sait « parler » de manière naturelle, mais avec une voix moyenne ou quelques voix génériques, pas encore la voix précise d'un artiste. On introduit ensuite une représentation numérique qui correspond à l'identité vocale d'une personne. En pratique, on passe des extraits audios de plusieurs locuteurs dans un réseau qui produit, pour chacun, une représentation numérique spécifique. Le modèle de conversion apprend ensuite à utiliser cette représentation pour contrôler le timbre, les formants, l'accent et une partie de la prosodie.

Pour cloner un artiste, il suffit alors d'estimer sa représentation numérique spécifique à partir de quelques secondes ou minutes de sa voix. Pour rendre le clonage très fidèle (respiration, grain, tics prosodiques), on procède à une adaptation légère du modèle : soit on « gèle » la plupart des paramètres du modèle pré-entraîné et on n'ajuste qu'une petite couche ou un module (par exemple un adaptateur), soit on effectue un affinage très limité avec les enregistrements de l'artiste. C'est cette architecture qui permet, avec peu de données, d'obtenir un clone vocal convaincant pour du discours ou du chant. Quelques secondes

d'enregistrement suffisent pour cloner une voix et créer un avatar audio avec des mouvements faciaux naturels et des lèvres qui suivent automatiquement la diction du propos.

Clonage express et temps réel

Le clonage vocal à partir de très peu d'enregistrements s'est considérablement simplifié. Il suffit désormais de quelques secondes d'audio (un extrait d'interview, une vidéo publiée sur les réseaux sociaux, une émission de radio) pour produire un clone vocal fonctionnel d'une personne. Techniquement, aucun réentraînement n'est nécessaire : le modèle, entraîné une fois pour toutes sur des milliers de voix différentes, extrait une empreinte vocale à partir de cet échantillon minimal (timbre, intonations, rythme) et s'en sert comme référence pour générer de nouveaux contenus dans ce style vocal, sans aucune intervention supplémentaire. Ni la coopération de la personne concernée, ni un enregistrement en studio ne sont requis. Cette capacité s'étend au chant : des outils récents permettent de cloner non seulement la parole mais aussi la voix chantée d'un artiste à partir de ses enregistrements existants. Pour les cas nécessitant une ressemblance très fine (notamment pour le chant) un affinage léger reste possible, à partir de quelques minutes d'enregistrements, sans modifier le modèle de base dans son ensemble. Pour l'industrie culturelle, cela signifie que le catalogue vocal de n'importe quel artiste ayant une présence publique constitue en pratique une matière première librement exploitable pour produire des contenus synthétiques à son image sonore.

Le clonage vocal en temps réel va plus loin encore : il ne s'agit plus de générer un fichier audio après coup, mais de transformer une voix en direct, instantanément. Des outils comme Seed-VC reposent sur le même principe de conditionnement (l'empreinte vocale de la cible est extraite de quelques secondes d'enregistrement et sert de référence permanente) mais appliqué en continu sur le flux audio entrant. Un individu parle avec sa propre voix et est entendu, en temps réel par ses interlocuteurs, avec la voix d'une autre personne (un artiste, une célébrité, un dirigeant, etc.). Le délai introduit par la transformation est de l'ordre de 300 millisecondes, imperceptible dans une conversation normale. Cela ouvre des possibilités d'usurpation d'identité vocale dans des contextes jusqu'ici difficiles à manipuler : réunions professionnelles en visioconférence, streaming en direct, ou communications téléphoniques.

Pour résumer, la reconstitution de la voix chantée d'un artiste repose sur une combinaison analogue de couches techniques. L'empreinte vocale est extraite à partir d'enregistrements existants (albums, clips, interviews) qui fournissent au modèle les caractéristiques distinctives de la voix : timbre, vibrato, tessiture, style d'articulation. Aucun réentraînement spécifique à l'artiste n'est nécessaire : le modèle utilise ces enregistrements comme simple référence de conditionnement, de la même façon qu'une photo de visage guide la génération d'une image. Pour les cas nécessitant une ressemblance très fine (par ex. reproduction précise d'un style vocal particulier ou d'un registre spécifique) un affinage léger reste possible à partir de quelques minutes d'enregistrements, sans modifier le modèle de base. Des paroles ou une mélodie cibles sont ensuite fournies en entrée, et le modèle génère un enregistrement chanté dans la voix de l'artiste. Des outils comme Seed-VC permettent également de convertir une voix chantée source (celle d'un autre chanteur) vers la voix cible, transférant ainsi l'interprétation tout en substituant l'identité vocale. Le résultat est un enregistrement où la voix, le style interprétatif et les nuances expressives d'un artiste réel sont entièrement synthétisés sans sa participation, à partir de son catalogue publiquement accessible.

6.1.3. Textes et style « cognitif »

L'hypertrucage textuel consiste à reproduire un texte avec le style d'une personne ou d'une institution précise dès lors que ce style est identifiable. Les applications culturelles incluent l'imitation d'un auteur ou d'un style rédactionnel, le prolongement de l'univers d'une fiction, la simulation de la pensée d'un auteur et la co-écriture. D'un point de vue technique, les grands modèles de langage (GML/*LLM*) les plus récents (GPT-4, Claude, Gemini, DeepSeek-R1) peuvent imiter un style selon trois niveaux d'effort croissant : une simple instruction textuelle suffit souvent ("écris dans le style de") ; quelques extraits de l'auteur fournis directement dans la requête permettent une imitation plus fidèle sans aucun réentraînement ; et un affinage sur un corpus plus large produit les résultats les plus précis. Leur capacité à exécuter finement ces instructions doit également aux techniques d'alignement par retour humain, c'est-à-dire des évaluateurs humains ayant noté des milliers de réponses pour apprendre au modèle à respecter des consignes stylistiques précises.

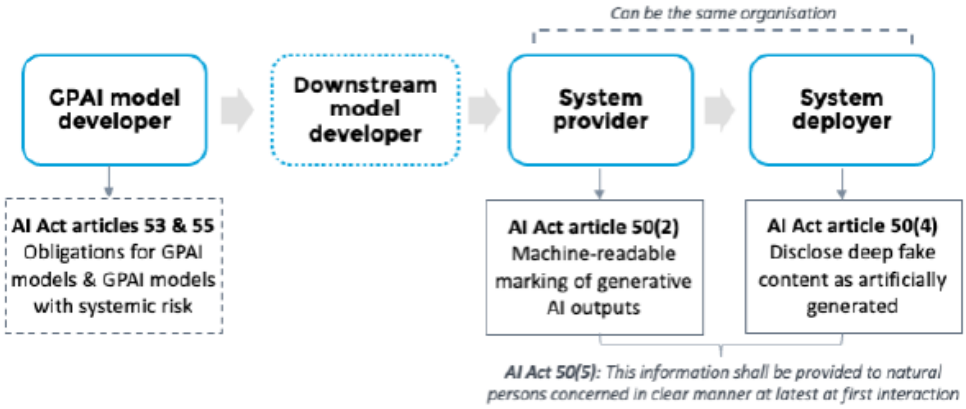
Contrairement à l'image ou à la voix, l'identité textuelle est plus cognitive et moins biologique qu'une voix ou qu'un corps. Le style est abstrait et peut être imité assez facilement, parfois même par hasard. Il est de ce fait plus difficile à détecter : l'absence de signal biométrique (visage, voix) rend les méthodes de détection fondées sur la physique (comme les caractéristiques visuelles et cinétiques d'un visage, l'acoustique d'une voix) inopérantes. Le « clonage mental » demeure néanmoins imparfait : les LLM peuvent reproduire des tics stylistiques et des structures syntaxiques, mais capturent mal les ruptures créatives, les références biographiques implicites et la cohérence thématique profonde d'un auteur.

6.2. Les solutions techniques de détection des hypertrucages : cadre d'évaluation

Les solutions se distinguent selon la modalité concernée (visuelle, audio, textuelle) et selon le moment d'intervention le plus précoce : la détection de « signatures » des modèles d'IA en aval de la génération de contenu en l'absence de marquage (dite détection forensique), et la détection d'un marquage placé en amont de la génération d'un contenu. La différence fondamentale entre ces modalités réside dans la nature du signal, l'existence d'une contrainte physique sous-jacente et la possibilité d'une identité perceptive stable.

Le moment d'intervention varie selon les acteurs. Les fournisseurs d'IA peuvent marquer les contenus au moment de la génération ; les plateformes, éditeurs, services de diffusion en flux (*streaming*) et diffuseurs peuvent préserver les marques et contrôler les contenus avant mise à disposition ; les titulaires de droits, autorités et tiers de confiance peuvent procéder à des vérifications en aval. Pour l'audio, il faut enfin distinguer les signaux audio consommés comme tels des formats symboliques comme le MIDI : ces derniers relèvent davantage de méthodes proches du texte et des métadonnées que des techniques appliquées au signal sonore.

L'applicabilité du RIA pour le marquage et les règles de transparence dans une chaine de production simplifiée de l'IA générative (d'après [403])



Avant d'examiner les différentes familles techniques de détection et de marquage, il importe d'introduire **le cadre d'évaluation dérivé directement de l'article 50(2) du RIA, qui structure les rapports techniques commandés par la Commission européenne [404]**. Celui-ci comprend cinq critères :

1. **L'effectivité** (*effectiveness*) renvoie à la capacité réelle d'identifier les contenus synthétiques par rapport aux contenus humains.
2. **La robustesse** (*robustness*) désigne la résilience aux transformations subies en aval, en particulier l'édition et la recompression.

[403] Bram Rijsbosch et al., 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act', arXiv:2503.18156, preprint, arXiv, 8 October 2025, https://doi.org/10.48550/arXiv.2503.18156.
[404] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act* (Publications Office of the European Union, 2026), https://data.europa.eu/doi/10.2759/9763153; Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act* (Publications Office of the European Union, 2026), https://data.europa.eu/doi/10.2759/7579127; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act* (Publications Office of the European Union, 2026), https://data.europa.eu/doi/10.2759/0784462.

3. **La fiabilité** (*reliability*) recouvre la cohérence des performances à travers les domaines d'usage et la difficulté pour un tiers d'apprendre puis de falsifier la signature ou de tromper le détecteur.
4. **L'accessibilité** (*accessibility*) impose que les parties prenantes (utilisateurs, plateformes, régulateurs) puissent vérifier la marque ou les sorties du détecteur et en interpréter correctement les résultats.
5. **L'interopérabilité** (*interoperability*), enfin, désigne la faisabilité d'un dispositif unifié de marquage ou de détection, indépendant du fournisseur et de la technologie sous-jacente.

Le diagnostic transversal des rapports est sans appel : aucune méthode existante, ni de détection ni de marquage, ne satisfait simultanément les cinq exigences.

Trois opérations sont par ailleurs utilement distinguées, qui sont souvent confondues dans le débat public [405] :

- **Le marquage désigne l'insertion proactive d'une information signalant l'origine IA d'un contenu (métadonnées ou filigrane) par le système qui le produit.**
- **La détection (en aval lorsqu'un marquage a été réalisé) consiste à vérifier la présence ou l'intégrité de cette « marque » dans un contenu donné.**
- **L'identification (ou détection en aval sans marquage) renvoie à l'analyse « forensique » qui cherche à déterminer l'origine (généralement le modèle génératif utilisé) en l'absence de toute marque préalable, en s'appuyant uniquement sur les artefacts caractéristiques laissés par le générateur.**

Ces trois opérations peuvent intervenir à différents moments du cycle de vie d'un contenu : à l'entraînement du modèle (marquage embarqué dans les données), à la génération (marquage *post-hoc* ou intégré au décodeur), à la diffusion (vérification par les plateformes), ou en aval lors du contrôle *ex post* par les régulateurs, ayants droit ou journalistes. Cette pluralité de points d'intervention explique pourquoi aucune solution technique unique ne suffit et pourquoi un dispositif crédible doit articuler plusieurs mécanismes opérant à différents stades.

La grille des cinq critères que la Commission applique dans ses rapports techniques [406] n'est pas la seule en circulation. La littérature académique sur le tatouage numérique (*watermarking*) propose au moins deux cadres concurrents, partiellement convergents et partiellement différents, dont la comparaison éclaire le périmètre exact du dispositif réglementaire et suggère plusieurs pistes d'amélioration.

Encadré — Cadres alternatifs d'évaluation de la littérature académique et articulation avec les cinq critères du RIA

Le premier cadre est celui retenu d'une revue récente des techniques de marquage pour les contenus d'IA générative [407]. Hérité de la littérature classique sur la stéganographie numérique, l'auteur identifie quatre propriétés en tension.

L'imperceptibilité est la condition première : le contenu marqué doit rester indiscernable par les sens humains —invisible à l'œil pour une image, inaudible à l'oreille pour un enregistrement, naturel à la lecture pour un texte. Dans sa version la plus forte, c'est-à-dire l'absence de distorsion, cette exigence va plus loin : non seulement le contenu marqué ne doit pas paraître dégradé, mais sa distribution statistique doit être rigoureusement identique à celle d'un contenu non marqué, de sorte qu'aucun test algorithmique ne puisse distinguer les deux. Atteindre cette propriété forte est techniquement possible pour certaines modalités (des schémas récents l'ont démontré pour le texte et l'image) mais au prix de contraintes sévères sur les autres propriétés.

[405] Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

[406] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*; Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

[407] Lele Cao, 'Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities', arXiv:2504.03765, preprint, arXiv, 2 April 2025, https://doi.org/10.48550/arXiv.2504.03765; Xuandong Zhao et al., 'Sok: Watermarking for Ai-Generated Content', *2025 IEEE Symposium on Security and Privacy (SP)*, 2025, 2621–39, https://ieeexplore.ieee.org/abstract/document/11023391/.

La robustesse désigne la capacité du filigrane à survivre aux transformations que subit le contenu après sa création. Une image peut être compressée, recadrée, filtrée ou repartagée en basse résolution ; un enregistrement peut être converti, compressé par codec neuronal ou rejoué dans une pièce ; un texte peut être traduit, résumé ou légèrement reformulé. Un filigrane robuste reste détectable avec une forte confiance statistique après ces altérations, dans le cadre d'un test d'hypothèse avec des seuils de faux positifs et de faux négatifs acceptables. L'exigence de robustesse n'est pas absolue : elle se définit toujours par rapport à un ensemble d'attaques considérées comme raisonnables — les attaques les plus sophistiquées, comme la régénération par un modèle de diffusion ou la paraphrase par un grand modèle de langue, sont dans une catégorie à part.

La sécurité couvre deux menaces distinctes. La première est l'effacement (*scrubbing*) : un adversaire cherche à supprimer le filigrane tout en conservant la qualité du contenu, pour que ses productions ne soient pas détectées comme IA-générées. La seconde, plus grave, est l'usurpation (*spoofing*) : l'adversaire cherche à reconstituer la règle secrète de marquage (par exemple en interrogeant massivement le système ou en analysant un corpus de contenus marqués) pour apposer ensuite une fausse signature sur des contenus qu'il n'a pas produits, ou pour accuser à tort un contenu humain d'être artificiel. Un schéma sécurisé doit résister aux deux types d'attaque, ce qui implique que la règle d'insertion et la clé secrète ne puissent pas être reconstituées même par un adversaire disposant d'un accès important au système.

La capacité désigne la quantité d'information pouvant être encodée dans la signature comme un élément (*bit*) unique (présent ou absent), ou un code binaire de plusieurs dizaines de *bits* permettant d'identifier le modèle source, l'utilisateur, la date de génération ou d'autres métadonnées de traçabilité. Une capacité plus élevée permet un usage plus riche de la signature, mais elle impose généralement de modifier davantage le contenu original, ce qui dégrade l'imperceptibilité et fragilise la robustesse.

Ces quatre propriétés sont structurellement en tension, et aucun schéma existant ne les satisfait simultanément à un niveau élevé. Rendre un filigrane plus robuste aux transformations le rend statistiquement plus saillant et donc plus facile à localiser et tromper. Augmenter sa capacité dégrade l'imperceptibilité. Viser l'absence de distorsion statistique contraint fortement la robustesse. Cette impossibilité partielle n'est pas seulement empirique : des résultats théoriques récents suggèrent qu'elle est fondamentale pour certaines combinaisons de propriétés, **ce qui impose aux concepteurs de choisir explicitement, selon l'usage visé, quels compromis ils acceptent**[408].

Un second cadre académique, plus granulaire, compte six axes[409]. Trois recoupent largement le cadre de Cao : la qualité (qui regroupe l'imperceptibilité, la propriété de faible distorsion et la propriété d'indistinguabilité statistique évoquée plus haut), la robustesse face aux attaques d'évasion (qui correspond à la composante anti-effacement de la sécurité chez Cao), et le support de messages encodés (équivalent de la capacité). Trois axes supplémentaires apparaissent qui ne figurent pas explicitement chez Cao. Le taux de faux positifs renvoie à la probabilité qu'un contenu authentique soit signalé à tort comme marqué. Le caractère infalsifiable (*unforgeability*) correspond à la composante anti-usurpation de la sécurité chez Cao, mais elle est traitée comme un axe distinct. **L'efficacité computationnelle, enfin, désigne le coût d'exécution du marquage et de la détection, qui conditionne la viabilité industrielle du déploiement à grande échelle.**

Comparée à ces deux cadres académiques, **la grille des cinq critères du RIA présente une parenté évidente mais aussi des spécificités notables.** L'effectivité du RIA correspond à un agrégat des notions de qualité, de taux de faux positifs et de taux de faux négatifs chez Zhao, ou plus largement à l'imperceptibilité couplée à la détectabilité chez Cao. La robustesse a le même contenu dans les trois cadres : résilience aux transformations subies en aval. La fiabilité du RIA recouvre essentiellement l' « unforgeability » chez Zhao et la composante anti-usurpation de la sécurité chez Cao, en y ajoutant la cohérence des performances à travers les domaines d'usage (un aspect que les deux cadres académiques ne thématisent pas explicitement). Les deux derniers critères du RIA, en revanche, sont absents des deux cadres académiques. L'accessibilité (capacité pour les parties prenantes de vérifier la marque et d'en interpréter les sorties) et l'interopérabilité (unification du dispositif de détection à travers les fournisseurs et les technologies) relèvent du déploiement sociotechnique du dispositif plutôt que de la conception interne du filigrane. Ce sont des critères ajoutés par les rapports de la Commission pour rendre opérationnelle, à l'échelle d'un marché unique, ce qui dans la littérature académique est traité l'état de spécifications de schémas individuels. Le tableau suivant résumé ces différences.

[408] Cao, 'Watermarking for AI Content Detection'.
[409] Zhao et al., 'Sok'.

Cadre académique (Cao 2025)	Cadre académique (Zhao 2025)	Critère RIA correspondant
Imperceptibilité	Qualité	Effectivité (composante qualité)
Robustesse	Taux de faux négatifs et robustesse	Robustesse
Sécurité, composante anti-usurpation	*Unforgeability*	Fiabilité
Sécurité, composante anti-effacement	Sous robustesse / attaques d'évasion	Robustesse
Capacité (absent chez Cao)	Support de messages Taux de faux positifs	(absent du RIA) (intégré à l'effectivité du RIA, non dissocié)
(absent chez Cao)	Efficacité computationnelle	(absent du RIA)
(absent chez Cao et Zhao)	(absent)	Accessibilité
(absent chez Cao et Zhao)	(absent)	Interopérabilité

Contrairement à l'apport sociotechnique de la Commission, **plusieurs propriétés mises en avant dans la littérature académique n'ont pas de contrepartie explicite dans la grille du RIA. Trois pistes méritent d'être considérées.**

Premièrement, l'ajout d'un critère explicite de capacité (quantité d'information encodée dans la signature) permettrait de distinguer les schémas se limitant à signaler la nature synthétique d'un contenu (un bit) de ceux capables d'identifier le modèle générateur, l'utilisateur émetteur ou la date de génération (plusieurs dizaines de bits). Cette distinction est cruciale pour les usages de traçabilité fine, en particulier pour la lutte contre la fraude et l'attribution en cas de litige.

Deuxièmement, la dissociation explicite du taux de faux positifs et du taux de faux négatifs, suivant Zhao, clarifierait un compromis aujourd'hui dissimulé dans le critère unique d'effectivité. Les conséquences d'un faux positif (accuser à tort un contenu humain d'être artificiel) et d'un faux négatif (laisser passer un contenu IA sans signalement) sont radicalement asymétriques, et un même schéma peut viser des points de fonctionnement très différents selon la pondération choisie. Rendre cette dissociation explicite faciliterait les choix politiques en aval.

Troisièmement, l'efficacité computationnelle proposée par Zhao et al., mériterait d'apparaître comme critère à part entière. Le coût marginal du marquage et de la détection à l'échelle de milliards de contenus quotidiens, sur les plateformes de diffusion par exemple, conditionne directement la viabilité de l'adoption, en particulier pour les petits acteurs. Faire de l'efficacité un critère évaluable inciterait à publier des parangonnages sur la consommation effective de ressources, alignés avec l'objectif d'accessibilité aux PME.

Les grilles d'évaluation que nous venons de présenter s'appliquent, avec quelques nuances selon les cas, à l'ensemble des familles de détection. On distingue les opérations de détection et d'identification en aval, sans marquage préalable (6.3) et les opérations, sur lesquelles portent les obligations voulues par le législateur européen, de marquage en amont et de détection des marques (6.4).

6.3. Détecter en aval les contenus synthétiques qui circulent, en l'absence de marquage (détection « forensique »)

Face à la multiplication des contenus synthétiques qui circulent, des demandes importantes de détection se font jour. L'objectif de la détection d'un hypertrucage se décompose en deux étapes : détecter qu'un contenu est synthétique et détecter qu'une source spécifique est utilisée.

6.3.1. Distinguer automatiquement contenus synthétiques et humains

En l'absence de marquage en amont, les techniques de détection vont comparer les propriétés statistiques des contenus humains et celles des contenus synthétiques avec pour objectif d'identifier pour ces derniers des artefacts détectables. Ces approches forensiques, qui sont des analyses probabilistes et non déterministes, ont des spécificités qui dépendent de la modalité considérée.

Images et vidéos

La détection en aval des images et vidéos non marquées repose sur l'analyse d'artefacts statistiques et d'incohérences. Il est possible d'utiliser des comparaisons statistiques entre des images synthétiques générées par des modèles GAN ou de diffusion et des images authentiques[410]. Pour les vidéos, des réseaux de neurones peuvent être entraînés à repérer des artefacts de bas niveau et des incohérences temporelles : frontières de fusion, artefacts fréquentiels liés au suréchantillonnage, anomalies de clignement des yeux, incohérences labiales et de pose de tête, ainsi que des indices physiologiques comme les micro-mouvements ou le pouls[411]. La détection peut également être multimodale, en comparant la cohérence entre le mouvement des lèvres, le son et la temporalité[412]. Toutefois, en s'améliorant, les modèles génératifs commencent à intégrer ces signaux et à produire une meilleure synchronisation, ce qui réduit progressivement l'efficacité de la détection par artefacts.

Pour l'image et la vidéo, la détection en aval en l'absence de marquage correspond à ce que le rapport de la Commission européenne consacré aux contenus visuels appelle la détection passive, ou analyse forensique[413]. Elle doit être distinguée des trois techniques qui supposent une intervention en amont ou au moment de la publication : les signatures cryptographiques et *Content Credentials* (métadonnées standardisée de provenance et d'authenticité attachées à un contenu numérique tel que l'image, la vidéo, l'audio, etc.), les métadonnées non cryptographiques, et les filigranes numériques intégrés aux pixels ou aux trames (cf. *supra*). Dans le cas de la détection passive, le fichier ne porte pas nécessairement de marque explicite. L'analyse cherche donc à repérer des traces laissées par le processus de génération lui-même : motifs de bruit, corrélations anormales entre pixels, signatures fréquentielles, incohérences d'éclairage, défauts de fusion du visage, clignements d'yeux atypiques, incohérences entre mouvements de lèvres et piste audio, ou discontinuités temporelles entre les images d'une vidéo. Le rapport indique que certains détecteurs peuvent même apprendre à reconnaître la famille de modèle ayant produit l'image ou la vidéo, par exemple en distinguant les empreintes statistiques associées à différents générateurs.

Cette approche reste fragile. Elle dépend fortement des contenus utilisés pour entraîner les détecteurs et des modèles connus au moment de cet entraînement. Lorsqu'un nouveau générateur apparaît, ou lorsqu'un modèle existant est amélioré, les artefacts précédemment utilisés peuvent disparaître. Les transformations ordinaires du fichier (compression, recadrage, redimensionnement, réencodage vidéo, filtres, capture d'écran, repost sur une plateforme) peuvent aussi affaiblir ou effacer les signaux exploités par le détecteur. À l'inverse, certains traitements peuvent produire de faux signaux et conduire à des erreurs. Pour cette

---

[410] Riccardo Corvi et al., 'On The Detection of Synthetic Images Generated by Diffusion Models', *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095167; Chandler Timm C. Doloriel et al., 'Towards Sustainable Universal Deepfake Detection with Frequency-Domain Masking', *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17 March 2026, 3797266, https://doi.org/10.1145/3797266.

[411] Javier Hernandez-Ortega et al., 'DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-Frame', in *Handbook of Digital Face Manipulation and Detection*, ed. Christian Rathgeb et al., Advances in Computer Vision and Pattern Recognition (Springer International Publishing, 2022), https://doi.org/10.1007/978-3-030-87664-7_12.

[412] Marcella Astrid et al., 'Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies', arXiv:2408.06753, preprint, arXiv, 14 October 2024, https://doi.org/10.48550/arXiv.2408.06753.

[413] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*.

raison, la détection passive est utile comme outil complémentaire, notamment pour analyser des contenus non marqués déjà en circulation, mais elle ne peut pas constituer à elle seule une infrastructure fiable de transparence au sens de l'article 50. Elle doit être articulée avec des mécanismes de marquage et de provenance, tels que les *Content Credentials*, les signatures cryptographiques, les métadonnées standardisées et les filigranes robustes. Le rapport de la Commission européenne sur l'image et la vidéo insiste précisément sur cette logique de combinaison plutôt que sur l'idée d'une solution unique.

Audio

Comme pour les images et les vidéos, il est possible d'utiliser des comparaisons statistiques entre des sons synthétiques et authentiques[414], y compris pour les modèles de type codec[415]. Les voix synthétiques souffrent pour l'heure de quelques anomalies détectables : des transitions trop lisses, une régularité excessive et de faibles micro-variabilités. Des incohérences physiologiques vocales sont également détectables dans la respiration, la temporalité et la prosodie[416]. Toutefois, les solutions de détection en aval déclinent rapidement quand les modèles d'IA progressent, la même course aux outils de détection que pour la vidéo s'appliquant au domaine audio.

Ainsi, comme pour l'image et la vidéo, des traces peuvent signaler la génération synthétique. Elles peuvent provenir des vocodeurs neuronaux, des modèles de synthèse vocale ou des décodeurs utilisés pour produire de la musique générée. Elles peuvent prendre la forme d'un lissage spectral excessif, de discontinuités temporelles, d'irrégularités statistiques dans la forme d'onde, d'artefacts propres à certains décodeurs, ou encore d'incohérences dans la respiration, la prosodie et les micro-variations de la voix. Pour la parole, cette approche s'inscrit dans la continuité des travaux d'anti-usurpation (*anti-spoofing*) utilisés contre les attaques visant les systèmes de vérification du locuteur ; pour la musique, certains travaux cherchent à relier un morceau généré au modèle qui l'a produit en exploitant les artefacts spécifiques des décodeurs.

Cette détection reste toutefois fragile. Elle dépend des modèles connus au moment de l'entraînement des détecteurs, des bases d'apprentissage disponibles et de la stabilité des artefacts recherchés. Or les transformations ordinaires du fichier (compression, ré-encodage, changement de hauteur, étirement temporel, bruit de fond, égalisation ou diffusion sur une plateforme) peuvent affaiblir les indices exploités. L'approche par empreinte audio peut être très efficace dans des cas fermés, lorsqu'il existe une base de référence fiable de contenus authentifiés, par exemple des archives officielles ; mais elle ne permet pas, à elle seule, d'identifier tout nouveau contenu généré ou modifié. C'est pourquoi le rapport de la Commission européenne sur les contenus audio distingue bien la vérification d'une marque, l'identification par empreinte (cf. 6.2.2) et l'identification du modèle génératif, et recommande de les combiner avec des métadonnées protégées et des filigranes imperceptibles plutôt que de s'appuyer uniquement sur l'analyse forensique de contenus non marqués[417].

Texte

Pour le texte, certaines méthodes de détection fondées sur la régularité statistique sont aujourd'hui insuffisantes. La détection par *perplexité* qui quantifie l'étendue de la surprise dans un texte, c'est-à-dire à quel point l'enchaînement des mots et des phrases est attendu en moyenne, reposait sur l'idée qu'un humain écrit de façon irrégulière, tandis que les modèles produisent des textes très réguliers, à faible perplexité[418]. Toutefois, les modèles d'IA peuvent désormais reproduire l'irrégularité d'un texte humain[419]. Une autre technique basée sur la perplexité mesurée à travers différents modèles permet la détection sans

[414] Yinlin Guo et al., 'Audio Deepfake Detection with Self-Supervised Wavlm and Multi-Fusion Attentive Classifier', *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, 12702–6, https://ieeexplore.ieee.org/abstract/document/10447923/.

[415] Orchid Chetia Phukan et al., 'Towards Neural Audio Codec Source Parsing', arXiv:2506.12627, preprint, arXiv, 14 June 2025, https://doi.org/10.48550/arXiv.2506.12627.

[416] Nicolas M. Müller et al., 'Mlaad: The Multi-Language Audio Anti-Spoofing Dataset', *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, 1–7, https://ieeexplore.ieee.org/abstract/document/10650962/.

[417] Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

[418] Eric Mitchell et al., 'Detectgpt: Zero-Shot Machine-Generated Text Detection Using Probability Curvature', *International Conference on Machine Learning*, 2023, 24950–62, https://proceedings.mlr.press/v202/mitchell23a.html.

[419] Biyang Guo et al., 'How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection', arXiv:2301.07597, preprint, arXiv, 18 January 2023, https://doi.org/10.48550/arXiv.2301.07597.

entraînement mais n'a pas été étudiée à large échelle[420]. Une alternative consiste à entraîner des modèles d'IA à reconnaître un texte produit par IA[421] ce que proposent plusieurs détecteurs commerciaux. Le problème est que ces détecteurs doivent être entraînés sur chaque nouveau modèle. De surcroît, la précision de cette approche est limitée, notamment par de nombreux faux positifs : un texte peut être signalé comme créé par IA alors qu'il ne l'est pas[422]. Des travaux récents explorent la détection par cohérence discursive globale, en analysant la structure argumentative et la cohérence à longue distance. Ces méthodes reposent sur l'hypothèse que les modèles génératifs présentent des signatures spécifiques dans l'organisation globale du discours, bien que ces différences tendent à se réduire avec l'amélioration des modèles[423]. Des travaux récents argumentent même que la détection précise de texte généré par LLM pourrait être « mathématiquement impossible » à mesure que les modèles s'améliorent, certains chercheurs déconseillant désormais de se fier aux détecteurs actuels[424].

Cette approche peut être utile lorsque le texte circule sans métadonnées, lorsqu'il a été copié-collé hors de son format initial, ou lorsqu'aucune coopération du fournisseur n'est disponible. Ses limites sont cependant importantes. Les détecteurs textuels dépendent fortement des corpus utilisés pour les entraîner, des langues, des domaines et des styles couverts. Un détecteur entraîné sur certains modèles, certains types d'instructions génératives (*prompts*) ou certains genres d'écriture peut perdre en fiabilité face à de nouveaux modèles, à des textes spécialisés ou à des textes hybrides mêlant contributions humaines et génération automatique. Le rapport de la Commission européenne sur le texte[425] souligne aussi que la paraphrase, la traduction aller-retour, l'édition humaine ou l'utilisation d'un second modèle peuvent réduire fortement la performance des méthodes fondées sur les traces statistiques. À la différence d'un filigrane, dont on peut parfois mesurer la dégradation, il est plus difficile de savoir précisément quelle quantité de modification suffit à faire disparaître les indices utilisés par un détecteur passif. Cette méthode peut donc servir d'outil de tri ou d'indice complémentaire, mais elle ne devrait pas être présentée comme une preuve autonome et stable du caractère généré d'un texte.

Conclusions

Plusieurs pistes de recherche structurent le champ de la détection en 2026. Une première consiste à développer des systèmes de détection plus généraux, capables d'identifier des contenus synthétiques sans avoir été spécifiquement entraînés sur chaque nouvel outil de génération. Il s'agirait d'une sorte de détecteur universel plutôt qu'un détecteur dédié à chaque modèle. Une seconde piste porte sur la détection multimodale : au lieu d'analyser l'image, le son ou le texte séparément, ces approches vérifient la cohérence entre ces différentes dimensions. Par exemple, si les mouvements de lèvres, la voix et le contenu textuel sont parfaitement synchronisés d'une façon qui trahit une fabrication artificielle.

Ces avancées restent toutefois structurellement fragiles. Chaque amélioration des détecteurs stimule en retour une amélioration des générateurs pour y échapper, selon une logique de course aux armements qui, à ce stade, favorise systématiquement les producteurs de contenus synthétiques. Les modèles de génération progressent plus vite que les outils de détection, et certains chercheurs estiment que la détection précise et généralisable de contenus générés par IA pourrait être fondamentalement impossible à mesure que ces modèles s'améliorent. La détection ne peut donc constituer, à elle seule, une réponse suffisante : elle doit s'inscrire dans une stratégie plus large intégrant la traçabilité à la source, les métadonnées d'authenticité et les obligations de transparence.

---

[420] Abhimanyu Hans et al., 'Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text', arXiv:2401.12070, preprint, arXiv, 13 October 2024, https://doi.org/10.48550/arXiv.2401.12070.

[421] Zae Myung Kim et al., 'Threads of Subtlety: Detecting Machine-Generated Texts through Discourse Motifs', *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, 5449–74, https://aclanthology.org/2024.acl-long.298/.

[422] Weixin Liang et al., 'GPT Detectors Are Biased against Non-Native English Writers', *Patterns* 4, no. 7 (2023), https://www.cell.com/patterns/fulltext/S2666-3899(23)00130-7?ref=maginative.com; 'New AI Classifier for Indicating AI-Written Text', OpenAI, 13 March 2024, https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/.

[423] Kim et al., 'Threads of Subtlety'.

[424] Mingmeng Geng and Thierry Poibeau, 'On the Detectability of LLM-Generated Text: What Exactly Is LLM-Generated Text?', arXiv:2510.20810, preprint, arXiv, 23 October 2025, https://doi.org/10.48550/arXiv.2510.20810.

[425] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.

La stratégie de détection d'hypertrucages en aval, en l'absence de marquage, est donc encore parfois efficace mais semble peu pérenne à mesure que les modèles progressent. La solution durable réside davantage dans le déploiement de stratégies en amont (cf. 6.4). **De plus, la distinction entre contenus humains et contenus synthétiques ne permet pas de repérer la ressemblance d'un contenu avec un existant ;** d'autres techniques de détection sont alors nécessaires (6.3.2).

6.3.2. Evaluer la ressemblance avec une personne réelle : la détection d'identité

La transparence sur la manière dont le contenu a été créé, par l'IA ou non, ne peut constituer un indicateur suffisant (*proxy*) pour restaurer la confiance du public. **Savoir si un contenu est généré par l'IA n'équivaut pas à savoir s'il est fiable et digne de confiance** (par ex. une scène historique signalée comme générée par IA, mais présentant un caractère sérieux alors qu'elle véhicule un message historiquement faux). Bien que ces deux questions puissent se recouper, tous les contenus issus de l'IA générative ne sont pas trompeurs ou problématiques, et tous les contenus trompeurs ou problématiques ne sont pas générés par l'IA.

De plus, au-delà de la question « ce contenu est-il généré par IA ? », la spécificité des hypertrucages nécessite de répondre à une seconde question : « ce contenu représente-t-il fidèlement une identité réelle ? ». Les méthodes de détection d'identité ne se contentent pas de classer un contenu comme humain ou synthétique : elles vérifient si un visage, une voix ou un style correspondent effectivement à une personne donnée.

Reconnaissance faciale et vérification d'identité visuelle

Dans le cas des images et des vidéos, la détection d'usurpation d'identité repose principalement sur des techniques de reconnaissance faciale et de comparaison biométrique. Ces méthodes consistent à extraire des représentations vectorielles du visage à partir d'images ou de vidéos de référence d'une personne, puis à les comparer aux visages présents dans des contenus suspects afin d'évaluer la correspondance identitaire. Des architectures de reconnaissance faciale profonde telles qu'ArcFace ou FaceNet permettent aujourd'hui d'atteindre une précision élevée pour vérifier si un visage correspond à une identité donnée[426]. Des travaux récents étendent ces approches en intégrant la dynamique temporelle du visage et la cohérence des caractéristiques identitaires dans les vidéos : il est possible de comparer l'évolution temporelle des caractéristiques faciales d'une vidéo suspecte à celles observées dans des vidéos de référence authentiques[427].

Ces approches s'inscrivent dans le prolongement des travaux sur la biométrie faciale[428] et des recherches récentes sur la cohérence identitaire dans les hypertrucages[429]. En pratique, des systèmes de comparaison faciale à grande échelle sont déjà utilisés par certaines plateformes en ligne pour détecter l'usurpation d'identité ou les contenus non autorisés représentant des personnalités publiques, bien que leur fonctionnement reste largement opaque et réservé aux grandes plateformes.

Vérification d'identité vocale

Dans le domaine audio, la détection d'usurpation d'identité repose sur les techniques de vérification du locuteur, qui comparent la signature vocale d'un enregistrement suspect à des empreintes vocales de référence. Ces méthodes s'appuient sur des représentations acoustiques profondes permettant de

[426] Jiankang Deng et al., 'Arcface: Additive Angular Margin Loss for Deep Face Recognition', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4690–99, http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

[427] Davide Cozzolino et al., 'Id-Reveal: Identity-Aware Deepfake Video Detection', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 15108–17, http://openaccess.thecvf.com/content/ICCV2021/html/Cozzolino_ID-Reveal_Identity-Aware_DeepFake_Video_Detection_ICCV_2021_paper.html.

[428] Mei Wang and Weihong Deng, 'Deep Face Recognition: A Survey', *Neurocomputing* 429 (March 2021): 215–44, https://doi.org/10.1016/j.neucom.2020.10.081.

[429] Baojin Huang et al., 'Implicit Identity Driven Deepfake Face Swapping Detection', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 4490–99, http://openaccess.thecvf.com/content/CVPR2023/html/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.html.

caractériser le timbre, les propriétés spectrales et certaines caractéristiques physiologiques de la voix d'un individu. Les modèles de représentation numérique vocale issus de la recherche en reconnaissance du locuteur constituent la base de nombreux systèmes d'authentification vocale et de détection d'usurpation[430]. Ces technologies sont déjà déployées à grande échelle dans certains contextes industriels, notamment pour l'authentification biométrique dans les services bancaires ou pour l'identification d'enregistrements dans les bases de données musicales.

Des techniques d'empreinte numérique audio sont utilisées par les plateformes de diffusion en flux (*streaming*) et de partage de contenus pour identifier automatiquement des œuvres musicales protégées[431], et des systèmes similaires pourraient être étendus à la détection d'usurpation d'identité vocale. L'utilisation d'empreinte numérique consiste à extraire une signature compacte à partir du contenu brut d'un fichier (les motifs (*patterns*) de pixels, les fréquences, la structure du signal) qui permet d'identifier sa réutilisation (la même musique, le même enregistrement, ou une copie légèrement dégradée) dans d'autres fichiers. En pratique, un artiste ou un ayant droit peut constituer une base d'empreintes vocales de référence et utiliser des outils de comparaison à grande échelle pour repérer des contenus diffusés en ligne qui correspondent à sa voix ou à son timbre. Toutefois, ces dispositifs restent principalement accessibles aux grandes plateformes ou aux ayants droit disposant d'outils spécialisés ; peu de services ouverts permettent aux individus de surveiller automatiquement l'usage non autorisé de leur identité vocale sur internet. L'apparition de parangonnages d'usurpation vocale[432] et de systèmes anti-usurpation prenant en compte le locuteur (*speaker-aware anti-spoofing*) témoigne de la structuration progressive de ce champ[433].

Stylométrie et attribution d'auteur

Dans le cas des contenus textuels, la détection d'usurpation d'identité repose principalement sur des méthodes de stylométrie computationnelle visant à déterminer si un texte correspond effectivement au style d'un auteur donné. Ces approches analysent des caractéristiques telles que le choix lexical, la syntaxe, la structure argumentative ou des représentations vectorielles globales du style, afin de comparer un texte suspect à un corpus de référence[434]. Les méthodes récentes d'attribution d'auteur ne reposent plus sur le comptage de traits stylistiques prédéfinis (fréquence de certains mots, longueur des phrases, richesse du vocabulaire) mais sur des modèles de langage entraînés à reconnaître globalement le "profil stylistique" d'un auteur, de la même façon qu'un modèle de reconnaissance faciale reconnaît un visage sans décomposer explicitement ses traits. Ces modèles produisent une représentation numérique compacte du style d'un auteur à partir d'un corpus de référence, puis vérifient si un texte inconnu correspond à ce profil. Les systèmes les plus récents vont plus loin en intégrant directement cette capacité dans des grands modèles de langage, qui peuvent alors raisonner explicitement sur les ressemblances stylistiques et en fournir une justification en langage naturel[435]. Ces approches correspondent étroitement à ce qui serait nécessaire pour une protection active des artistes dans les industries culturelles : disposer d'un corpus authentifié de l'écriture, de la voix ou du style d'un artiste, et vérifier automatiquement si une œuvre candidate en est une reproduction fidèle ou une imitation non autorisée — qu'il s'agisse d'un texte dans le style d'un auteur vivant, d'une chanson dans le timbre d'un interprète, ou d'une image dans le registre visuel d'un plasticien.

---

[430] Ye Jia et al., 'Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis', *Advances in Neural Information Processing Systems* 31 (2018), https://proceedings.neurips.cc/paper/2018/hash/6832a7b24bc06775d02b7406880b93fc-Abstract.html.

[431] Pedro Cano et al., 'A Review of Audio Fingerprinting', *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* 41, no. 3 (2005): 271–84, https://doi.org/10.1007/s11265-005-4151-3.

[432] Xin Wang et al., 'Asvspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech', *Computer Speech & Language* 95 (2026): 101825.

[433] Xin Wang and Junichi Yamagishi, 'Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders', *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5, https://ieeexplore.ieee.org/abstract/document/10094779/; Jiangyan Yi et al., 'Audio Deepfake Detection: A Survey', *arXiv Preprint arXiv:2308.14970*, 2023, https://arxiv.org/abs/2308.14970.

[434] Efstathios Stamatatos, 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538–56, https://doi.org/10.1002/asi.21001.

[435] Maël Fabien et al., 'BertAA: BERT Fine-Tuning for Authorship Attribution', *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, 127–37, https://aclanthology.org/2020.icon-main.16/; Baixiang Huang et al., 'Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges', *ACM SIGKDD Explorations Newsletter* 26, no. 2 (2025): 21–43, https://doi.org/10.1145/3715073.3715076; Javier Huertas-Tato et al., 'PART: Pre-Trained Authorship Representation Transformer', arXiv:2209.15373, preprint, arXiv, 9 May 2025, https://doi.org/10.48550/arXiv.2209.15373.

Protection personnalisée des artistes

Des travaux récents vont plus loin vers la protection personnalisée des personnalités publiques et des artistes. C'est le cas de VIPGuard[436]. Cette ligne de recherche ajuste finement un grand modèle multimodal pour qu'il comprenne des caractéristiques faciales très détaillées, puis effectue un apprentissage discriminant au niveau de l'identité et enfin une personnalisation par utilisateur, en apprenant un « jeton VIP » léger qui encode l'identité d'une personne donnée. Les détecteurs personnalisés raisonnent ensuite sur des décalages subtils entre une image suspecte et les images antérieures du VIP, en fournissant à la fois une décision et une explication. Pour l'industrie culturelle, cette approche ouvre la voie à des systèmes de surveillance personnalisés permettant aux artistes et à leurs ayants droit de détecter automatiquement les usurpations de leur identité à travers les plateformes. La détection de l'identité repose donc sur des techniques qui paraissent fiables et déployées à grande échelle.

Conclusions

Nous avons couvert deux problèmes techniques souvent confondus dans les débats sur les hypertrucages, et qui n'ont pas le même niveau de maturité.

- Le premier consiste à **détecter la présence d'une identité dans un contenu. Il s'agit d'un problème largement résolu, sur le plan technique.** Les systèmes modernes de reconnaissance faciale permettent aujourd'hui de déterminer avec une grande fiabilité qu'une vidéo ou une image représente bien une personne donnée, **à partir d'un corpus de référence authentifié**. Cette capacité, robuste même face à des contenus dégradés, compressés ou recadrés, est déjà déployée. Pour les artistes et leurs représentants, cela signifie qu'il est techniquement possible, dès aujourd'hui, de surveiller automatiquement la diffusion non autorisée de leur image à travers les plateformes, à condition de disposer d'un corpus de référence authentifié et d'un accès aux outils appropriés.

- Le second problème qui est de **déterminer si ce contenu est authentique ou fabriqué, est en revanche beaucoup plus difficile** et fait l'objet de la recherche décrite dans les sections précédentes. La détection de l'hypertrucage en tant que tel reste, en l'absence de marquage, un problème ouvert, structurellement asymétrique : les techniques de génération progressent plus vite que les techniques de détection, et certains travaux théoriques suggèrent que la détection parfaite est mathématiquement impossible à terme. **La détection, en l'absence de marquage en amont, reste donc peu fiable, avec des perspectives dégradées du fait de l'avancement des technologies.** C'est pourquoi le marquage, dès la production, offre des perspectives décisives.

6.4. Rendre détectables les contenus synthétiques dès la génération de contenus, par le marquage en amont

Parce qu'il devient de plus en plus difficile de distinguer le contenu généré par l'IA de celui créé par l'être humain, non seulement en raison du réalisme croissant des résultats de l'IA générative, mais aussi des progrès de la multimodalité et de la commercialisation des outils (par exemple, l'intégration de l'édition par l'IA dans l'appareil photo des téléphones Pixel), des acteurs de l'industrie, de la société civile et du monde universitaire ont investi dans des outils permettant de marquer les contenus synthétiques dès l'amont pour faciliter la détection en aval. Par exemple, les technologies de traçabilité peuvent aider à déterminer si une photo a été prise avec un appareil photo, modifiée par un logiciel ou produite par une IA générative.

L'approche la moins onéreuse et la plus simple consiste à produire des métadonnées pour les contenus numériques produits par des humains et, idéalement, par des modèles d'IA. Les métadonnées correspondent à des informations structurées sur le contenu (date, auteur, outil de création, historique des modifications) contenu dans le ficher intégrant les données d'intérêt. Des contenus produits par IA peuvent ainsi être enregistrés dans des formats de fichiers qui stockent des métadonnées sur la manière dont ils ont été générés. C'est par exemple du texte associé à une image. Un téléphone peut par exemple enregistrer des images et des fichiers audios, en utilisant un format qui contient des informations sur l'heure de la photographie ; de la même manière, il est possible de stocker des métadonnées permettant de savoir si le

[436] Kaiqing Lin et al., 'Guard Me If You Know Me: Protecting Specific Face-Identity from Deepfakes', arXiv:2505.19582, preprint, arXiv, 3 December 2025, https://doi.org/10.48550/arXiv.2505.19582.

contenu a été généré par IA. Ces formats sont standardisés, par exemple, le format IPTC pour l'image qui indique les champs qui doivent être renseignés, comme le créateur et les contributeurs[437]. Les métadonnées pourraient donc en principe être exploitées pour étiqueter un contenu comme étant généré par IA. De surcroît, elles peuvent être signées cryptographiquement, en ajoutant un identifiant unique permettant d'identifier l'authenticité d'un contenu. Parmi les problèmes des métadonnées, qui limitent leur portée, on trouve le fait qu'il est aisé de les supprimer ; elles sont souvent éliminées systématiquement lors d'un partage de contenu sur les réseaux sociaux de façon à préserver la vie privée et pour limiter la bande passante. Dès lors, marquer une vidéo de manière plus profonde, plus indélébile, paraît décisif. On distinguera les méthodes de marquage explicites et implicites.

6.4.1. Les méthodes explicites de marquage

Les méthodes explicites comprennent les filigranes visibles (par exemple, logos ou icônes intégrés dans une image), les étiquettes de contenu incorporées dans une image, ou des champs de divulgation séparés pouvant inclure des avertissements et des mises en garde sur le contenu[438]. L'objectif principal de ces marquages visibles est d'informer directement les usagers sur la nature du contenu. La conception et la mise en œuvre de divulgations visibles efficaces demeurent un défi complexe et multidimensionnel[439], dans la mesure où il est difficile de s'assurer que les utilisateurs ordinaires comprennent pleinement les marquages indiquant la nature générée par IA d'un contenu. De plus, les marquages visibles intégrés dans une image se heurtent à des limites pratiques : ils peuvent être facilement supprimés ou nuire à l'expérience de lecture du contenu. Cette approche semble donc limitée ; d'autres méthodes de marquage implicite proposent un filigranage invisible.

6.4.2. Les méthodes implicites de marquage

Les techniques de marquage implicites intègrent des informations lisibles par machine dans le contenu d'une image, destinées à être interprétées par des systèmes techniques plutôt que par des humains. Il existe une variété de techniques pour des solutions de filigranages lisibles par la machine[440]. La méthode du filigranage (*watermarking*) et des signatures embarquées consiste en des schémas proactifs qui insèrent un filigrane ou une marque d'identité soit dans le générateur (filigranage neuronal des sorties) au moment de la génération, soit dans le média protégé. Les détecteurs vérifient ensuite la présence ou l'intégrité de cette marque dans le contenu analysé dont on cherche à évaluer l'authenticité. Ces outils incrustent des filigranes numériques directement dans les images, les audios, les textes ou les vidéos générés par l'IA. Les filigranes traditionnels ne suffisent pas à identifier les images générées par l'IA, car ils sont souvent apposés comme un tampon sur l'image et peuvent être facilement modifiés. Par exemple, des filigranes discrets dans le coin d'une image peuvent être coupés à l'aide de techniques d'édition de base.

Filigranage des images et vidéos

Le marquage des images et vidéos peut être intégré directement dans le processus de génération par IA, de façon invisible à l'œil nu[441]. Il peut être rendu spécifique à chaque générateur, permettant l'attribution de la source[442]. Ce marquage est robuste aux transformations de l'image (compression, recadrage, filtrage) et

---

[437] 'IPTC Photo Metadata Standard 2025.1', 2025, https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata-2025.1.html#contributor.

[438] National Institute of Standards and Technology, *Reducing Risks Posed by Synthetic Content*, 2024, https://dpo-india.com/Resources/NIST/Reducing-Risks-Synthetic-Content-Overview-Technical-Digital-Content-Transparency.pdf; Rijsbosch et al., 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act'.

[439] Abdallah El Ali et al., 'Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How', *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2 May 2024, 1–11, https://doi.org/10.1145/3613905.3650750.

[440] National Institute of Standards and Technology, *Reducing Risks Posed by Synthetic Content*.

[441] Pierre Fernandez et al., 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 22466–77, http://openaccess.thecvf.com/content/ICCV2023/html/Fernandez_The_Stable_Signature_Rooting_Watermarks_in_Latent_Diffusion_Models_ICCV_2023_paper.html; Zhao et al., 'Sok'.

[442] Ning Yu et al., 'Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14448–57, http://openaccess.thecvf.com/content/ICCV2021/html/Yu_Artificial_Fingerprinting_for_Generative_Models_Rooting_Deepfake_Attribution_in_Training_ICCV_2021_paper.html.

détectable avec une forte confiance statistique. Toutefois, il nécessite la coopération du producteur du contenu synthétique et n'est pas compatible avec des modèles open source non contrôlés. Le marquage neuronal dans les modèles de diffusion (par exemple *Tree-Ring Watermarking*[443]) reste encore émergent. Il est également possible de marquer de manière invisible une image après génération par IA ou par humain (par exemple StegaStamp[444]).

Le rapport de la Commission européenne consacré à l'image et à la vidéo[445] distingue quatre familles de solutions qui doivent être articulées plutôt que confondues : les signatures cryptographiques et *Content Credentials*, les métadonnées non cryptographiques, les filigranes numériques intégrés aux pixels ou aux trames, et la détection passive par analyse d'artefacts (cf. 6.4). Les signatures cryptographiques et *Content Credentials*, développées notamment dans le cadre du standard C2PA, consistent à associer au fichier des informations de provenance : outil de création, étapes de transformation, date de production, identité du signataire ou indication du caractère généré ou modifié par IA. Lorsqu'elles sont protégées par une signature cryptographique, ces informations deviennent vérifiables : la plateforme, le lecteur ou l'autorité de contrôle peuvent vérifier qu'elles n'ont pas été modifiées depuis leur apposition. Cette logique est proche d'un certificat attaché au contenu. Les métadonnées non cryptographiques reposent aussi sur l'ajout d'informations dans le fichier ou dans son environnement de diffusion, par exemple une balise indiquant qu'une image a été générée par IA. Elles sont plus faciles à déployer, mais aussi plus faciles à retirer ou à modifier, puisqu'elles ne sont pas protégées par un mécanisme d'authentification comparable. Les filigranes numériques intégrés aux pixels ou aux trames fonctionnent autrement : ils inscrivent une marque directement dans l'image ou dans les images successives d'une vidéo, souvent de manière imperceptible pour l'œil humain, afin qu'un détecteur puisse ensuite reconnaître cette marque. Ils peuvent donc résister à certaines suppressions de métadonnées, mais restent sensibles à la compression, au recadrage, au réencodage, aux filtres et aux attaques adversariales (*i.e.* une modification volontaire du contenu destinée à tromper un système de détection ou de marquage). L'intérêt est donc de combiner ces approches plutôt que de choisir une solution unique : les *Content Credentials* assurent la provenance lorsque la chaîne de confiance est conservée, les métadonnées simples facilitent l'information dans les environnements compatibles, et les filigranes peuvent compléter cette traçabilité lorsque les métadonnées sont absentes ou supprimées.

Pour les images et les vidéos, les critères d'évaluation doivent aussi tenir compte de la continuité de la chaîne de provenance. Les *Content Credentials* de type C2PA apportent une preuve vérifiable lorsque la signature est conservée, mais ils peuvent être perdus lors d'une capture d'écran, d'un réencodage ou d'une republication. Les filigranes sont plus persistants face à certaines transformations, mais ils doivent rester imperceptibles, résister à la compression, au recadrage et au montage, et éviter de créer de nouveaux risques pour la vie privée ou la sécurité. Les vidéos ajoutent une difficulté propre : la marque doit rester détectable malgré les coupes, recompressions, changements de cadence, sous-titres, filtres et réutilisations partielles.

Dans cette modalité, le rapport image-vidéo insiste sur l'intérêt d'une approche en couches. Les métadonnées signées apportent une information de provenance et un point d'ancrage pour les vérificateurs ; les filigranes intégrés au signal ajoutent une persistance lorsque les métadonnées sont perdues ; la détection passive peut servir de filet de sécurité pour les contenus non marqués. Aucune de ces méthodes n'est suffisante isolément : les métadonnées peuvent être retirées, les filigranes peuvent être attaqués ou devenir incompatibles entre fournisseurs, et les détecteurs passifs se dégradent lorsque les modèles de génération évoluent.

L'évaluation des images et vidéos doit enfin inclure la facilité d'usage pour les créateurs, plateformes, journalistes, autorités et utilisateurs finaux. Une solution très robuste mais fermée, coûteuse ou non interopérable risque de peu circuler ; une solution très simple mais non cryptographique peut être contournée ou falsifiée. Le rapport ajoute donc aux critères de l'article 50 des considérations pratiques :

---

[443] Yuxin Wen et al., 'Tree-Ring Watermarks: Fingerprints for Diffusion Images That Are Invisible and Robust', arXiv:2305.20030, preprint, arXiv, 4 July 2023, https://doi.org/10.48550/arXiv.2305.20030.

[444] Matthew Tancik et al., 'Stegastamp: Invisible Hyperlinks in Physical Photographs', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 2117–26, http://openaccess.thecvf.com/content_CVPR_2020/html/Tancik_StegaStamp_Invisible_Hyperlinks_in_Physical_Photographs_CVPR_2020_paper.html.

[445] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*.

coût d'intégration, charge pour les petits fournisseurs, sécurité, protection de la vie privée et lisibilité de l'information fournie au public[446].

Filigranage audio

Comme pour la vidéo et l'image, il est possible d'insérer une signature imperceptible dans le signal sonore synthétique[447]. Deux grandes approches se distinguent selon le moment où la signature est insérée.

La première consiste à marquer le son *après* sa génération, en ajoutant une perturbation imperceptible au signal produit. DeAR [448] est conçu pour résister aux situations où l'audio marqué est enregistré physiquement par un microphone — la signature survit aux distorsions acoustiques de la pièce et aux imperfections du matériel. WavMark[449] encode jusqu'à 32 bits d'information par seconde d'audio grâce à une architecture qui traite le marquage et la détection comme des opérations symétriques, ce qui préserve la qualité sonore ; le détecteur localise automatiquement la signature dans un enregistrement long sans connaître sa position à l'avance. AudioSeal[450], aujourd'hui la référence du domaine, va plus loin en entraînant simultanément le module qui cache la signature et celui qui la détecte, de sorte que la détection fonctionne au niveau de chaque fragment de quelques millisecondes. Cela permet non seulement de savoir si un enregistrement contient de la parole synthétique, mais de localiser précisément les segments concernés.

La deuxième approche intègre le marquage *pendant* la génération elle-même, ce qui la rend beaucoup plus difficile à contourner. Dans une approche génératrice, la signature est ancrée dans la façon dont le modèle a appris à produire des sons. Des chercheurs proposent ainsi de marquer les données d'entraînement du modèle, de sorte que toutes ses sorties portent nativement la signature quelle que soit la façon dont le son est ensuite décodé ce qui confère un avantage décisif pour les modèles dont le code est rendu public[451]. Ces marqueurs sont détectables statistiquement et résistent aux transformations standards telles que la compression ou le changement de format, bien que leur robustesse face à des adversaires actifs reste un sujet de recherche ouvert.

Une limite de portée importante doit être soulignée : ces techniques s'appliquent au signal audio sous sa forme d'onde (*waveform*), c'est-à-dire à l'enregistrement final tel qu'il est diffusé. Les formats symboliques de la musique (partitions numériques, MIDI, MusicXML) relèvent en pratique d'une logique de marquage proche de celle du texte, puisqu'ils encodent des symboles discrets plutôt qu'un signal continu[452]. Cette distinction est appelée à gagner en importance avec la diffusion des modèles génératifs travaillant directement sur des représentations symboliques de la musique. Cependant, dans les lignes directrices de l'article 50 de la Commission européenne, l'obligation de marquage ne s'applique qu'aux textes sémantiques, composés de caractères, excluant de facto les partitions musicales, et ce pour une raison non précisée.

Le rapport de la Commission européenne souligne aussi que les métadonnées sont faciles à intégrer et utiles pour la transparence, mais elles disparaissent facilement lors d'une conversion de format, d'une extraction ou d'une republication. Le filigranage audio peut survivre à certaines compressions et transformations, mais il doit rester inaudible et robuste aux usages ordinaires du fichier. L'empreinte audio est efficace pour reconnaître un contenu déjà connu dans une base de référence, mais elle ne permet pas, à elle seule, d'identifier un nouveau contenu synthétique ou une voix clonée qui n'a jamais été enregistrée. De ce fait, le

[446] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*.
[447] Chang Liu et al., 'Dear: A Deep-Learning-Based Audio Re-Recording Resilient Watermarking', *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 11 (2023): 13201–9, https://ojs.aaai.org/index.php/AAAI/article/view/26550.
[448] Liu et al., 'Dear'.
[449] Guangyu Chen et al., 'WavMark: Watermarking for Audio Generation', arXiv:2308.12770, preprint, arXiv, 7 January 2024, https://doi.org/10.48550/arXiv.2308.12770.
[450] Robin San Roman et al., 'Proactive Detection of Voice Cloning with Localized Watermarking', arXiv:2401.17264, preprint, arXiv, 6 June 2024, https://doi.org/10.48550/arXiv.2401.17264.
[451] Robin San Roman et al., 'Latent Watermarking of Audio Generative Models', *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, 1–5, https://ieeexplore.ieee.org/abstract/document/10889782/.
[452] Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

rapport audio distingue le marquage, la détection et l'identification dont l'usage diffère. Le marquage peut intervenir pendant la génération ou lors de la publication ; la détection est utile sur les plateformes et services de distribution ; l'identification intervient ensuite lorsque des titulaires de droits, des médias ou des autorités doivent analyser un contenu sans marque fiable (comme discuté dans la section sur la détection aval). Cette chaîne de vie est essentielle pour les contenus musicaux et vocaux, car un même fichier peut être généré, compressé, remixé, extrait d'une vidéo puis rediffusé dans un autre environnement technique.

Filigranage textuel

Pour le texte, le marquage en amont est plus délicat que pour l'image, la vidéo ou l'audio, car le texte peut être copié, reformulé, traduit ou extrait de son fichier sans conserver son support technique initial. Le rapport de la Commission européenne consacré au texte distingue cinq méthodes[453] : le filigranage, le marquage structurel, les métadonnées, la journalisation (*logging*) et la détection de texte généré. Parmi elles, les trois premières relèvent plus directement du marquage du contenu ou de son support, tandis que le *logging* et la détection statistique jouent plutôt un rôle de vérification complémentaire (cf. 6.2.1).

Le marquage textuel fonctionne presque exclusivement pendant la génération elle-même : le modèle est modifié pour biaiser discrètement ses choix de mots selon une règle secrète, de sorte que le texte produit porte une régularité statistique détectable. Des chercheurs introduisent ainsi l'approche de référence, dite des listes verte/rouge : à chaque mot généré, le modèle favorise légèrement certains mots parmi ceux qu'il juge possibles, sans que le lecteur humain ne perçoive la différence[454]. D'autres ont renforcé cette idée en faisant dépendre le choix de la liste non plus du mot précédent mais du sens global de la phrase en cours — une reformulation préserve davantage la signature si elle conserve le sens[455]. Ces deux approches sont donc des marquages intégrés au processus de génération, et non appliqués après coup sur un texte déjà produit.

Des approches post-hoc existent en principe (e.g. modifier un texte déjà généré pour y insérer des régularités cachées) mais elles sont nettement moins robustes et moins utilisées en pratique, car le texte ne dispose pas d'un support physique stable comme un signal audio ou une image : une simple reformulation suffit à briser la signature. C'est précisément là que réside la fragilité fondamentale du marquage textuel. D'une part, les paraphrases humaines ou automatiques effacent le signal, puisqu'elles remplacent les mots marqués par des équivalents qui ne respectent plus la règle secrète[456]. D'autre part, un adversaire disposant d'un accès à l'API du modèle peut reconstruire la règle secrète par interrogations répétées (ce qu'on appelle le vol de filigrane) et l'utiliser ensuite soit pour supprimer la signature de ses propres textes, soit pour en apposer une fausse sur des textes qu'il n'a pas produits. Cette attaque est réalisable pour moins de 50 dollars, avec un taux de succès supérieur à 80 % sur les schémas considérés jusqu'alors comme robustes[457]. Pour le texte, un marquage fiable reste donc difficile : le support est instable par nature, et les propriétés mêmes qui rendent un filigrane robuste (sa persistance à travers les reformulations) en font aussi une cible plus facile à localiser et à neutraliser. **Si des solutions fiables de marquage des images, des sons et des vidéos existent, pour les contenus textuels, les réponses paraissent plus difficiles.**

Le rapport de la Commission européenne sur le marquage textuel[458] distingue cinq méthodologies, dont le filigrane statistique présenté ci-dessus ne constitue qu'une famille parmi d'autres. Le marquage structurel (*structural marking*) consiste à coder une signature non pas dans le choix des mots mais dans la mise en forme du document — micro-espaces insécables, variations typographiques imperceptibles, ponctuation Unicode quasi identique. Ces signatures survivent à la copie brute mais sont triviales à effacer par un simple recopiage manuel ; elles trouvent toutefois une utilité dans les modèles à poids ouverts où le filigrane

---

[453] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.

[454] John Kirchenbauer et al., 'A Watermark for Large Language Models', *International Conference on Machine Learning*, 2023, 17061–84, https://proceedings.mlr.press/v202/kirchenbauer23a.html.

[455] Aiwei Liu et al., 'A Semantic Invariant Robust Watermark for Large Language Models', arXiv:2310.06356, preprint, arXiv, 19 May 2024, https://doi.org/10.48550/arXiv.2310.06356.

[456] Kalpesh Krishna et al., 'Paraphrasing Evades Detectors of Ai-Generated Text, but Retrieval Is an Effective Defense', *Advances in Neural Information Processing Systems* 36 (2023): 27469–500.

[457] Nikola Jovanović et al., 'Watermark Stealing in Large Language Models', arXiv:2402.19361, preprint, arXiv, 24 June 2024, https://doi.org/10.48550/arXiv.2402.19361.

[458] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.

intégré à la génération est inaccessible. Les métadonnées (*metadata*) embarquées dans des formats riches (PDF, DOCX, EPUB) permettent de signer cryptographiquement l'origine du fichier ; combinées au standard C2PA, elles offrent la solution la plus fiable mais aussi la plus fragile en cas de simple copier-coller du texte brut. La journalisation (*logging*) côté fournisseur consiste à enregistrer les générations effectuées et leur empreinte numérique, permettant en aval une vérification *a posteriori* par interrogation des journaux. Cette approche se heurte aux contraintes de protection des données personnelles et de coût de stockage. La détection sans marquage préalable, enfin, (déjà abordée *supra*) repose sur l'apprentissage automatique de classifieurs entraînés à reconnaître les textes générés, avec les limites de précision et de faux positifs déjà discutées.

Aucune de ces cinq méthodologies ne satisfait simultanément les cinq exigences réglementaires : les schémas robustes (filigrane statistique) sont vulnérables à l'usurpation (*spoofing*) par un tiers qui éditerait le texte sans effacer la marque, tandis que les schémas fiables (métadonnées signées) sont triviaux à supprimer par effacement (*scrubbing*)[459]. Le rapport de la Commission européenne recommande en conséquence des combinaisons différenciées selon l'usage applicatif. Pour la génération généraliste (de robots conversationnels (*chatbots*), moteurs de recherche augmentés), la combinaison de métadonnées et de filigranes statistiques offre le meilleur compromis. Pour les applications spécialisées (assistants produit, agents sectoriels), le filigrane suffit, complété par une détection automatique *a posteriori*. Pour les contenus sensibles (documents médicaux générés, actes juridiques, contrats), la métadonnée cryptographique signée est seule appropriée, l'altération du contenu étant rédhibitoire. Pour les modèles à poids ouverts enfin, où le filigrane intégré à la génération est techniquement contournable par un simple retrait du module concerné, le marquage structurel demeure l'option résiduelle, faute de mieux. **Cette différenciation par cas d'usage tranche avec l'idée d'une solution unique imposée à tous les fournisseurs.**

Un déploiement industriel limité

Un seul système multimodal a franchi le seuil du déploiement à grande échelle : SynthID, développé par Google DeepMind. Lancé en 2023 pour les images générées par Imagen sur Vertex AI, il a été étendu à l'audio (modèle Lyria, NotebookLM), à la vidéo (Veo) et au texte (Gemini), et rendu accessible en open source pour le texte en octobre 2024 via Hugging Face. SynthID-Image a marqué plus de dix milliards d'images et de trames vidéo à travers les services Google. Pour le texte, la méthode a fait l'objet d'une publication dans *Nature* [460], ce qui témoigne d'un niveau de maturité peu commun dans ce domaine. Du côté de Meta, AudioSeal était déjà en production dans les démonstrations publiques d'Audiobox et Seamless, servant 100 000 utilisateurs par jour, avant même sa publication scientifique en 2024[461].

Ces déploiements industriels partagent trois caractéristiques : ils sont intégrés pendant la génération et non en post-traitement, ils fonctionnent sur des infrastructures fermées où le producteur de contenu maîtrise l'ensemble de la chaîne, et ils ne prétendent pas résoudre le problème de l'adversaire actif — SynthID lui-même indique explicitement qu'il n'est pas conçu pour arrêter un adversaire déterminé, mais pour créer un obstacle suffisant dans les usages courants.

Pour les images, les méthodes Stable Signature et Tree-Ring sont disponibles en open source, fonctionnelles et bien documentées. Elles atteignent des performances solides contre les transformations standard (compression, recadrage, filtrage) avec des taux de fausse détection très bas. Elles sont utilisables dans des contextes où le producteur de contenu peut imposer le marquage et où les adversaires ne disposent pas des outils spécialisés d'effacement par régénération avec un modèle de diffusion. Pour l'audio, WavMark et DeAR offrent des performances comparables face aux attaques de signal standard, et AudioSeal constitue aujourd'hui la référence du domaine pour la parole synthétique. Pour le texte, le schéma de Kirchenbauer et ses variantes sont largement utilisés en recherche, disposent d'implémentations publiques testées, et fonctionnent bien sur des textes longs à contenu varié.

Toutefois, plusieurs obstacles empêchent la généralisation de ces méthodes[462].

---

[459] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.
[460] Sumanth Dathathri et al., 'Scalable Watermarking for Identifying Large Language Model Outputs', *Nature* 634, no. 8035 (2024): 818–23.
[461] Cao, 'Watermarking for AI Content Detection'; Zhao et al., 'Sok'.
[462] Cao, 'Watermarking for AI Content Detection'; Zhao et al., 'Sok'.

Le premier est l'absence de standards communs. Un filigrane SynthID n'est pas détectable par un outil conçu pour AudioSeal, et inversement. Il n'existe pas d'infrastructure de détection interopérable permettant à n'importe quel acteur de vérifier si un contenu a été marqué, quelle qu'en soit la source. Cela rend le marquage quasi inopérant pour la régulation ou la traçabilité à l'échelle de l'internet, où les contenus circulent entre plateformes et perdent leur contexte de génération.

Le second est le problème structurel des modèles open source. Tous les systèmes robustes intègrent le marquage pendant la génération, ce qui suppose que le producteur contrôle l'étape de génération. Pour les modèles dont les poids sont rendus publics — ce qui représente une part croissante de l'écosystème — rien n'oblige un utilisateur à conserver l'étape de marquage dans son propre déploiement. Les approches par marquage des données d'entraînement (San Roman et al., 2024) constituent une piste prometteuse pour contourner ce problème, mais elles sont encore en phase expérimentale.

Le troisième obstacle est la faiblesse spécifique du marquage textuel, que les deux articles qualifient de problème non résolu. Les contraintes sur le type de contenu marquable sont sévères : les réponses courtes, les textes factuels, les contenus à faible variabilité lexicale sont pratiquement impossibles à marquer de façon fiable. Google reconnaît explicitement cette limite pour SynthID Text, qui fonctionne mal sur les réponses à des questions factuelles. Or ce sont souvent ces contenus — affirmations factuelles directes, infox condensées — qui présentent les risques informationnels les plus sérieux.

Enfin, **même les systèmes les plus avancés n'ont été évalués que contre un ensemble limité d'attaques. Lorsqu'on élargit le panel d'attaques — en particulier les attaques par régénération avec un modèle de diffusion pour les images, par codec neuronal pour l'audio, ou par paraphrase automatique pour le texte — les performances chutent significativement.** Une évaluation systématique contre un large spectre d'attaques, avec des protocoles standardisés, fait encore défaut dans la littérature.

6.4.3. Certifier la provenance de l'authentique

Le marquage peut être appliqué pour un contenu synthétique, mais également pour un contenu humain, authentique. Une stratégie complémentaire consiste donc non pas à détecter le synthétique, mais à certifier l'authentique, en attachant une preuve de provenance au contenu dès sa création. Ces techniques permettent d'embarquer des métadonnées ou des filigranes dans les images générées, rendant leur origine authentique ou artificielle détectable par des outils automatisés.

Pour les images et la vidéo, le marquage peut être intégré à un contenu numérique au moment de l'enregistrement (prise de photo ou de vidéo) ou après. Au moment de la capture, le contenu peut être signé cryptographiquement, ce qui devient le standard industriel[463]. Après l'enregistrement, il est possible de marquer un contenu authentique pour détecter sa réutilisation future. Par exemple, des chercheurs encodent un filigrane dans les caractéristiques d'identité, de sorte que toute manipulation de visage altérant l'identité puisse être détectée via la vérification du filigrane[464]. Cette approche proactive est particulièrement pertinente pour la protection des artistes ou célébrités. Pour l'audio, la stratégie consiste également à certifier l'origine des contenus audio par des mécanismes de provenance et de signature cryptographique. La certification de l'origine des textes par des mécanismes de provenance et de signature cryptographique[465] intègre signature cryptographique à la création, métadonnées inviolables et historique des modifications.

Des standards de provenance de contenu s'imposent progressivement comme le protocole C2PA (Coalition for Content Provenance and Authenticity) qui enregistre chaque étape du cycle de vie d'un contenu.

Encadré – C2PA, une norme concertée visant à authentifier et tracer la provenance des contenus

[463] Irene Amerini et al., 'Deepfake Media Forensics: Status and Future Challenges', *Journal of Imaging* 11, no. 3 (2025): 73.

[464] Yuan Zhao et al., 'Proactive Deepfake Defence via Identity Watermarking', *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, 4602–11, https://openaccess.thecvf.com/content/WACV2023/html/Zhao_Proactive_Deepfake_Defence_via_Identity_Watermarking_WACV_2023_paper.html.

[465] Jaiden Fairoze et al., 'Publicly-Detectable Watermarking for Language Models', arXiv:2310.18491, preprint, arXiv, 4 January 2025, https://doi.org/10.48550/arXiv.2310.18491.

Le référentiel C2PA constitue désormais le standard industriel de référence pour la certification de provenance. La spécification C2PA v2.2 (avril 2025) introduit la notion de « *durable Content Credentials* », qui combine un *hard binding* (hachage cryptographique) avec un *soft binding* (filigranage) pour assurer que la provenance survit même si les métadonnées sont supprimées (C2PA, *Explainer v2.2*, 2025). Contrairement à un simple détecteur de contenus synthétiques, le C2PA fournit une chaine de traçabilité cryptographique qui certifie l'origine du fichier, les outils utilisées pour le créer et les modifications apportées. Au lieu de prouver qu'un contenu est synthétique, le système permet de prouver à l'utilisateur final qu'il est authentique ou certifié humainement. Le C2PA permet d'attacher une identité numérique au contenu dès sa création (auteur, date, historique des modifications).

Cette solution normative regroupe de nombreux adhérents (porté par Adobe, Microsoft, BBC, …), rejoints plus tardivement par certaines plateformes comme Meta, Google et les acteurs IA comme OpenAI, ou des agences de presse (AFP-Reuters). L'approche est ouverte (licence gratuite), mais son adoption nécessite des investissements d'intégration technique. Les initiatives européennes (DSA, RIA, *Coalition for Media Provenance*) soutiennent publiquement l'authentification des médias et l'adoption de ces standards[466]. L'objectif premier est d'interdire la suppression ou modification des informations de provenance, effectuée sciemment et dans un but de tromperie ou de gain financier et plus généralement de garantir que la provenance des contenus soit traçable, de leur production jusqu'à leur diffusion, afin de limiter les risques de tromperie et de renforcer la confiance des publics. Son intérêt est majeur pour les médias afin de certifier que le contenu diffusé provient bien du site officiel des adhérents au C2PA.

Le C2PA soulève cependant des critiques. Techniques tout d'abord. Des fraudeurs pourraient simuler les comportements des outils d'inspection pour authentifier abusivement des contenus ou pirater les métadonnées d'un contenu pour intégrer à un autre. Le C2PA ne traite pas directement de la question des hypertrucages, et son application au texte reste limitée. Ethiques, ensuite, en portant les risques d'atteinte à la vie privée lorsque par exemple le lieu et la personne ayant pris une photo peuvent être identifiés. Economiquement, pour être efficace, le système doit être adopté par le plus grand nombre d'éditeurs et de producteurs de contenus afin de marginaliser les fichiers dépourvus de ces métadonnées de provenance (comme un consommateur se détournerait d'un produit dépourvu du label de transparence nutriscore). Or l'échange avec ces acteurs est actuellement faible. De plus, la complexité d'implantation du standard peut empêcher les acteurs modestes de produire des métadonnées et le label de confiance risquerait alors de détourner les consommateurs de leurs contenus et de renforcer le pouvoir de marché des groupes dominants. La mise en œuvre du standard est limitée à quelques projets, l'adoption inexistante à l'échelle du web, et la visibilité pour le grand public réduite.

Conclusions

Le marquage est donc techniquement disponible et opérationnel pour les images et l'audio dans des contextes fermés avec un producteur coopératif. Il est fragile pour le texte et contournable pour les modèles open source. Si le principe du marquage préventif en amont est posé, son efficacité dépendra de la standardisation des méthodes de marquage et de signalement. Le risque est de voir émerger une multiplicité de systèmes distincts, développés par différents acteurs, sans interopérabilité ni cohérence. Ces techniques en amont nécessitent pour se développer la coopération des fournisseurs d'IA, ce qui constitue leur principale limite face aux modèles open source non contrôlés ou aux modèles « maison ». Rien ne pousse en effet ces derniers à procéder au marquage des contenus générés. Toutefois, la coexistence de contenus authentiques labellisés comme tels et de contenus artificiels labellisés comme tels réduirait la crédibilité des contenus non marqués (avec cependant un risque d'usurpation). **La détection et le signalement auprès du public** ne peuvent cependant reposer uniquement sur quelques acteurs volontaires ; elles **nécessitent une coopération entre les développeurs de solutions technologiques, les déployeurs d'outils, les diffuseurs de contenus culturels et les institutions publiques afin notamment de répartir la charge des investissements nécessaires.**

Une harmonisation européenne, voire internationale, permettrait de faciliter l'identification automatique par des outils techniques et de simplifier la vérification des contenus par les utilisateurs et par les plateformes. Le marquage et le signalement systématique des contenus synthétiques devront par ailleurs

[466] Amerini et al., 'Deepfake Media Forensics'.

nécessairement être résilients et robustes, de sorte qu'il ne soit pas aisé de retirer intentionnellement les indications en question mais également accessibles afin de ne pas désavantager les PME.

Pour pallier l'absence de standards d'interopérabilité, principal frein opérationnel à l'adoption du marquage, les rapports de la Commission européenne[467] proposent deux scénarios d'infrastructure de vérification.

- Le premier, dit du *vérificateur centralisé*, repose sur une autorité de confiance unique, comme par exemple le Bureau européen de l'IA, qui mettrait à disposition une interface publique permettant à n'importe quel utilisateur, plateforme ou régulateur de soumettre un contenu et d'en faire vérifier la provenance. Le vérificateur interrogerait successivement les métadonnées authentifiées du fichier puis, le cas échéant, les systèmes de détection mis à disposition par chaque fournisseur participant. Cette architecture suppose une participation volontaire des fournisseurs et une standardisation forte des formats de vérification.

- Le second scénario, dit du *vérificateur décentralisé*, confie ce rôle à plusieurs tiers de confiance accrédités plutôt qu'à un point unique, ce qui allège la contrainte de standardisation tout en distribuant les risques d'indisponibilité. Ces architectures partagent une même logique : déplacer le coût de l'interopérabilité du producteur de contenu vers une infrastructure publique mutualisée, à l'image de ce qui existe déjà pour la certification des sites web (autorités de certification SSL/TLS) ou pour la gestion des noms de domaine.

6.5. Synthèses et recommandations européennes pour la gouvernance technique de la transparence

Dès 2021, l'étude *Tackling Deepfakes in European Policy* commandée par le panel STOA du Parlement européen[468] élaborait un cadre conceptuel de réflexion[469]. Cette étude qui a structuré une grande partie de la pensée européenne ultérieure sur le sujet, **distingue cinq dimensions du cycle de vie d'un hypertrucage que toute politique publique cohérente doit articuler** :

- la dimension technologique (les modèles d'IA et leurs fournisseurs en amont, traités principalement par l'article 50(2) du RIA via les obligations de marquage) ;
- la dimension création (les utilisateurs et leurs usages, couverts par l'article 50(4) sur l'étiquetage des hypertrucages) ;
- la dimension circulation (la diffusion sur les plateformes et les intermédiaires, principalement régulée par le DSA) ;
- la dimension cible (la protection des personnes représentées, qui relève du RGPD, des droits de la personnalité et du droit d'auteur) ;
- la dimension audience (la littératie médiatique du public, qui relève notamment des politiques d'éducation).

Ce découpage rappelle qu'aucune dimension prise isolément ne suffit : la robustesse du dispositif européen de transparence dépend de la cohérence d'ensemble entre ces cinq aspects, comme le confirment quatre ans plus tard des rapports techniques de la Commission. Les trois tableaux ci-dessous, organisés par modalité (texte, image et vidéo, audio), synthétisent les évaluations comparatives présentées dans ces rapports (Fritz, 2025 ; Puccetti, 2026 ; Serra et al., 2026)[470].

---

[467] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*; Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

[468] van Huijstee et al., *Tackling Deepfakes in European Policy*.

[469] Voir aussi le rapport pour le CSPLA (2019), *Vers une application effective du droit d'auteur sur les plateformes numériques de partage, état de l'art et propositions sur les outils de reconnaissance des contenus* et les rapports de l'ARCOM (2022 et 2024) sur l'évaluation des mesures de protection.

[470] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*; Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

Tableaux de synthèse : techniques de marquage et de détection évaluées selon les cinq critères de l'article 50 du RIA

Chaque famille technique y est analysée selon les cinq critères dérivés de l'article 50(2) du RIA. L'échelle utilisée est : *Élevée*, *Moyenne*, *Faible*, *Très faible*. Les notes sont indicatives et destinées à donner une vue d'ensemble. Elles ne se substituent pas aux évaluations détaillées des rapports source.

Texte (Puccetti, 2026)

Technique	Effectivité	Robustesse	Fiabilité	Accessibilité	Interopérabilité
Filigrane statistique (type Kirchenbauer)	Moyenne (fragile sur textes courts ou factuels)	Moyenne	Faible (vulnérable au spoofing pour moins de 50 $)	Faible	Faible
Marquage structurel (Unicode, typographie)	Moyenne	Très faible (effacée par recopiage)	Faible	Moyenne	Faible
Métadonnées (C2PA en PDF, DOCX, EPUB)	Élevée si conservées	Très faible (perdues au copier-coller du texte brut)	Élevée (signature cryptographique)	Élevée	Moyenne
Journalisation côté fournisseur	Moyenne (requiert accès aux journaux)	Élevée (indépendante du contenu)	Moyenne (dépend de l'honnêteté du fournisseur)	Faible	Faible
Détection forensique (classifieurs)	Faible à moyenne et déclinante	Faible (spécifique au modèle)	Faible (taux de faux positifs élevé)	Moyenne (détecteurs commerciaux)	Faible

Image et vidéo (Fritz, 2025)

Technique	Effectivité	Robustesse	Fiabilité	Accessibilité	Interopérabilité
Signatures cryptographiques (C2PA, content credentials)	Élevée si présent	Faible (perdues à la recompression, recadrage, capture d'écran)	Élevée	Moyenne (lecteur C2PA requis)	Moyenne à élevée (standard émergent)
Métadonnées non cryptographiques (IPTC, EXIF)	Élevée si présent	Très faible (supprimées par les plateformes sociales)	Faible (sans signature)	Élevée (outils standards)	Élevée
Filigrane invisible intégré (SynthID, Stable Signature, Tree-Ring)	Élevée	Élevée (résiste à compression, recadrage, filtrage)	Moyenne (risque de spoofing documenté)	Faible (détecteur tenu par le producteur)	Faible (pas de format de détection commun)
Fingerprinting et détection passive (forensique)	Moyenne et déclinante	Faible (sensible aux mises à jour de modèles)	Moyenne (faux positifs)	Moyenne	Faible (spécifique aux modèles cibles)

Audio (Serra et al., 2026)

Technique	Effectivité	Robustesse	Fiabilité	Accessibilité	Interopérabilité
Métadonnées audio (BWF, ID3)	Élevée si présent	Faible (perdues au transcodage et au repackaging)	Faible (sans signature)	Élevée	Élevée
Filigrane audio (AudioSeal, WavMark, DeAR)	Élevée pour la parole synthétique	Élevée (survit à compression, capture micro)	Moyenne	Faible (détecteur tenu par le producteur)	Faible (pas de standard de détection)
Fingerprinting audio (référence sur contenu connu)	Élevée pour le contenu déjà répertorié, nulle pour le nouveau	Élevée	Élevée	Moyenne (outils commerciaux)	Élevée (technologie mature)
Identification forensique de modèle génératif	Moyenne et déclinante	Faible (sensible aux mises à jour de modèles)	Moyenne	Faible (stade recherche)	Faible (spécifique aux modèles cibles)

Trois régularités structurent ces tableaux.

- Premièrement, aucune technique n'obtient un profil simultanément élevé sur les cinq critères, ce qui confirme la nécessité de l'approche en couches recommandée par les rapports.
- Deuxièmement, les techniques de marquage signées cryptographiquement (métadonnées, C2PA) inversent toujours le profil des filigranes invisibles : les premières sont fiables et accessibles mais fragiles à la robustesse, les seconds sont robustes mais peu accessibles et peu interopérables.
- Troisièmement, **les techniques forensiques de détection sans marquage sont systématiquement les moins performantes sur l'effectivité et tendent à se dégrader à mesure que les générateurs progressent, ce qui les disqualifie comme solution unique mais les maintient comme filet de sécurité pour les contenus non marqués.**

Au sein des trois modalités, **le texte est la plus problématique** : aucune technique n'y atteint un score élevé sur l'effectivité combinée à la robustesse, et le filigrane est intrinsèquement plus fragile que ses équivalents image et audio. L'image et la vidéo disposent des solutions les plus matures (SynthID combiné à C2PA), tandis que l'audio présente un profil intermédiaire avec une particularité forte : l'existence d'une technique de prise d'empreinte (*fingerprinting*) mature mais limitée au répertoire connu, qui n'a pas d'équivalent direct dans les autres modalités.

Les trois rapports techniques publiés par la Direction générale des réseaux de communication, du contenu et des technologies (CNECT) en 2026[471] convergent vers un même diagnostic : **aucune solution technique ne suffit à elle seule à satisfaire l'ensemble des exigences de l'article 50 du RIA, et toute stratégie viable doit reposer sur la complémentarité de plusieurs mécanismes.** Ce diagnostic doit toutefois être lu avec une précision supplémentaire : les rapports ne recommandent pas une technique unique, mais une combinaison ajustée selon la modalité, le type d'acteur, le degré de risque et le moment de vérification. Les critères d'efficacité, de robustesse, de fiabilité, d'accessibilité et d'interopérabilité doivent donc servir de grille minimale pour toute solution proposée.

Lue à l'aune des cinq critères de l'article 50, l'évaluation comparative des techniques fait apparaître une asymétrie structurelle qui justifie l'approche en couches. Les signatures cryptographiques et *content*

[471] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*; Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*; SPECTRA, *Analysis of the Stakeholder Consultation on Transparency Requirements for Certain AI Systems under Article 50 AI Act* (Publications Office of the European Union, 2026), https://data.europa.eu/doi/10.2759/0753439.

credentials (type C2PA) obtiennent les meilleures notes en effectivité, fiabilité et accessibilité, mais souffrent d'une robustesse faible (suppression triviale par recompression, copier-coller ou repartage en basse résolution) et d'une interopérabilité encore partielle. Le filigrane invisible (image, audio, texte) inverse précisément ce profil : robustesse élevée aux transformations standards, mais effectivité très inégale selon les modalités (très bonne pour l'image, bonne pour l'audio, fragile pour le texte) et interopérabilité limitée faute de standard commun de détection. La prise d'empreinte (*fingerprinting*) passive et la détection forensique *a posteriori* présentent une effectivité décroissante au fil de l'amélioration des générateurs et restent surtout utilisables comme filets de sécurité. Le marquage structurel textuel et la journalisation côté fournisseur ne sont pertinents que pour des cas d'usage précis (modèles ouverts pour le premier, contenus sensibles avec exigence de traçabilité pour le second). Aucune technique n'obtient une note simultanément élevée sur les cinq critères ; cette asymétrie est précisément ce qui justifie les orientations qui suivent.

Première orientation : rendre obligatoire l'approche multi-couches. Pour les images et la vidéo, cela signifie l'adoption conjointe du standard C2PA pour la provenance et d'un filigrane invisible robuste de type SynthID intégré dans le pipeline de génération[472]. Pour l'audio, la combinaison recommandée associe métadonnées cryptographiquement signées, filigrane imperceptible (par exemple AudioSeal) et analyse forensique des artefacts résiduels lorsque les marques explicites sont absentes[473]. Pour le texte, où le marquage demeure structurellement fragile, l'association du filigrane statistique avec des métadonnées C2PA embarquées dans des formats riches (PDF, DOCX) est jugée la plus prometteuse, même si elle ne résout pas entièrement les vulnérabilités à l'usurpation (*spoofing*)[474]. Cette logique en couches obéit à un principe de défense en profondeur : chaque mécanisme compense les faiblesses des autres, et un adversaire devrait simultanément contourner plusieurs dispositifs hétérogènes pour produire un contenu parfaitement non détectable.

Deuxième orientation : interdire légalement le retrait des marques. Le rapport sur l'audio[475] recommande explicitement de prohiber, à l'instar de la directive 2001/29/CE sur le droit d'auteur (qui interdit déjà le contournement des mesures techniques de protection), tout retrait, altération ou contournement intentionnel des marques d'origine IA[476]. Cette interdiction transformerait le contournement d'une marque — aujourd'hui techniquement aisé — en infraction sanctionnée, modifiant ainsi le calcul économique des acteurs malveillants. Elle suppose toutefois de distinguer le retrait intentionnel des dégradations involontaires occasionnées par les traitements légitimes (compression, transcodage par codec neuronal), ce qui plaide en retour pour la poursuite de la recherche sur la robustesse intrinsèque des filigranes.

Troisième orientation : créer une infrastructure publique de soutien aux PME et aux utilisateurs. L'adoption d'outils sophistiqués comme C2PA ou SynthID a un coût d'intégration significatif qui désavantage structurellement les petits acteurs[477]. Trois dispositifs sont proposés : une autorité de certification européenne mutualisée permettant aux créateurs indépendants d'obtenir la signature de leurs contenus authentiques sans monter eux-mêmes une infrastructure cryptographique ; une bibliothèque open source de référence pour le filigranage et la détection, librement intégrable par n'importe quel développeur ; des extensions de navigateur et applications mobiles permettant au grand public de vérifier d'un clic l'authenticité d'une image ou d'une vidéo. Sans cette mutualisation, le marquage risquerait de renforcer la position dominante des grandes plateformes capables d'absorber le coût de conformité, en marginalisant les PME.

Quatrième orientation : instaurer un programme permanent de monitorage (*monitoring*). Les rapports recommandent que le Bureau européen de l'IA (ou une entité équivalente) mette en place un dispositif de

---

[472] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*.
[473] Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.
[474] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.
[475] Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.
[476] Le CPI, quant à lui, sanctionne déjà le retrait des mesures techniques de protection mais aussi des mesures techniques d'identification pour les œuvres ce qui pourrait être étendu aux contenus synthétiques.
[477] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*.

surveillance échantillonnant régulièrement les contenus en circulation pour mesurer le taux effectif de marquage et de détection, complété par un programme structuré de tests d'intrusion confiés à des chercheurs indépendants et à des journalistes[478]. Les vulnérabilités identifiées alimenteraient une mise à jour annuelle des spécifications techniques et des bonnes pratiques. Les résultats des tests d'intrusion pourraient être transmis aux autorités de surveillance des marchés (ARCOM et DGCCRF en France). Cette logique itérative reconnaît implicitement que la transparence n'est pas un état stable mais un processus en équilibre instable avec les capacités génératives, qui progressent plus vite que les outils de détection.

Cinquième orientation : standardiser au niveau européen et international. Le constat est unanime : la prolifération de schémas de marquage propriétaires et non interopérables réduit à néant l'utilité du marquage à l'échelle de l'internet, où les contenus circulent entre plateformes et perdent leur contexte de production[479]. Trois leviers sont identifiés : la standardisation par les organismes européens (CEN-CENELEC, ETSI) des formats de détection et de l'API publique de vérification ; la convergence avec les démarches volontaires, notamment américaines et chinoises, pour permettre la reconnaissance mutuelle des marques ; l'engagement des secteurs non-IA particulièrement exposés (médias, justice, finance) dans l'élaboration de lignes directrices sectorielles. À défaut, le risque est de voir coexister durablement plusieurs écosystèmes incompatibles, chacun ne couvrant qu'une fraction du marché et aucun ne permettant une vérification universelle.

Sixième orientation : porter l'application des règles du RIA au niveau des modèles fondamentaux (fondationnels) via la qualification de « risque systémique ». Une enquête empirique sur l'adoption effective du marquage par les générateurs d'images grand public révèle que seuls 38 % des systèmes implémentent une forme de marquage lisible par machine et que cette adoption se concentre presque entièrement dans les acteurs intégrés de bout en bout[480]. Or l'écosystème des modèles fondationnels est extrêmement concentré : quelques fournisseurs (Stability AI, Black Forest Labs, OpenAI) couvrent l'essentiel des déploiements en aval. Les auteurs proposent en conséquence d'utiliser l'article 51 du RIA pour qualifier les modèles d'image les plus avancés de « modèles d'IA à usage général à risque systémique », ce qui imposerait directement à leurs développeurs l'intégration de techniques de marquage robustes et non désactivables (par exemple le marquage embarqué dans le bruit initial des modèles de diffusion). Ce déplacement des obligations juridiques vers l'amont présente trois avantages opérationnels : il réduit le nombre d'acteurs à contrôler de plusieurs milliers à quelques dizaines ; il neutralise la stratégie de désactivation observée chez les déployeurs en aval ; et il aligne la responsabilité technique sur celle des seuls acteurs disposant de l'expertise et des ressources pour concevoir des schémas robustes.

Ces six orientations dessinent une trajectoire crédible mais exigeante. Elles supposent un investissement public soutenu, une coopération inédite entre acteurs concurrents et une articulation fine avec les autres pans du droit numérique européen. Surtout, **elles invitent à abandonner l'idée que la transparence pourrait être garantie par un seul dispositif technique[481]: c'est la cohérence d'un écosystème — technique, juridique, économique et éducatif — qui conditionnera la capacité à restaurer la confiance dans les contenus numériques produits par l'IA générative.**

De plus, la consultation conduite par la Commission européenne auprès des parties prenantes en 2025[482] consacre une part substantielle de ses recommandations aux **modalités concrètes par lesquelles l'information sur l'origine IA d'un contenu doit parvenir à l'utilisateur final.** Plusieurs principes se dégagent. La notification doit d'abord intervenir à la première exposition au contenu (première diffusion pour la vidéo, mention en tête d'article pour le texte, signal sonore en ouverture pour les interactions vocales) plutôt qu'enfouie dans les conditions générales ou un menu secondaire. Elle doit ensuite être persistante pour les contenus longs ou les interactions prolongées, sans tomber dans la répétition aveuglante qui produit l'effet inverse (cécité aux bannières (*banner blindness*)). L'accessibilité impose une

---

[478] Fritz, Mario, *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*; Serra, Xavier et al., *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*.

[479] Puccetti, Giovanni, *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*.

[480] Rijsbosch et al., 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act'.

[481] Zhao et al., 'Sok'.

[482] SPECTRA, *Analysis of the Stakeholder Consultation on Transparency Requirements for Certain AI Systems under Article 50 AI Act*.

déclinaison multimodale : étiquette textuelle en langage simple, icône reconnaissable, équivalent audio descriptif, contraste suffisant pour les personnes malvoyantes, compatibilité avec les lecteurs d'écran. Les répondants insistent sur le risque de sur-étiquetage qui banaliserait la mention « contenu IA » et la rendrait inopérante : ils recommandent de privilégier un étiquetage *au niveau du segment* indiquant précisément quelles portions d'une œuvre composite sont synthétiques, et de définir clairement les flux de travail hybrides humain/IA dans lesquels la responsabilité du marquage doit être imputée. L'articulation avec les autres pans du droit numérique européen est mentionnée comme une priorité opérationnelle : les métadonnées de provenance peuvent constituer des données personnelles au sens du RGPD ; les obligations des plateformes face aux contenus marqués doivent être coordonnées avec le DSA ; les règles sectorielles sur la publicité politique et le droit de la consommation introduisent des obligations supplémentaires qu'il faut éviter de dupliquer ou de contredire.

La consultation confirme la convergence des attentes dans les industries culturelles et médiatiques, qui figurent parmi les secteurs les plus représentés dans les réponses. Trois enseignements se dégagent.

D'abord, les répondants (éditeurs, producteurs, plateformes, associations d'artistes) rejettent l'idée qu'une technique unique puisse répondre à l'obligation de transparence et plaident pour une **approche en couches** combinant étiquetage visible, métadonnées de provenance signées cryptographiquement, filigrane invisible et journalisation des générations.

Ensuite, ils demandent un **étiquetage segmenté** plutôt que global, afin de signaler uniquement les portions effectivement synthétiques d'une œuvre composite et **d'éviter le sur-étiquetage qui décrédibiliserait les contenus authentiques mêlés à des éléments générés**.

Enfin, ils s'accordent sur la nécessité de distinguer **l'édition assistée** (correction grammaticale, transcription, recadrage automatique), qui ne devrait pas déclencher d'obligation de marquage, de la véritable génération ou manipulation.

La mise en œuvre de l'article 50(2), articulée avec l'article 50(4), du RIA recouvre en effet trois questions juridiquement et techniquement distinctes qu'il importe de séparer (Fritz, 2026 ; Serra et al., 2026 ; Puccetti, 2026).

- La première porte sur l'identification d'un contenu généré ou substantiellement manipulé par IA, qui déclenche l'obligation générale de marquage lisible par machine pesant sur les fournisseurs au titre de l'article 50(2).
- La deuxième concerne l'hypertrucage au sens de l'article 50(4), qui ajoute une obligation d'étiquetage explicite à destination du public lorsqu'un contenu audiovisuel artificiel ou manipulé présente faussement des personnes, lieux, objets, entités ou événements comme authentiques.
- **La troisième est la plus délicate : il s'agit de tracer la frontière entre l'édition assistée —** correction grammaticale, débruitage audio, recadrage automatique, transcription — qui ne devrait pas déclencher d'obligation au titre de la réserve d'« usage assisté » de l'article 50(2), **et la véritable manipulation** modifiant le sens ou la portée du contenu. Le RIA n'en propose pas de définition opérationnelle. Le rapport audio de la Commission cite ainsi un contraste éclairant : la réduction de bruit de fond sur un enregistrement existant ne nécessite manifestement aucun marquage, tandis que la fabrication d'une fausse déclaration prêtée à une voix clonée relève du cœur même de l'obligation (Serra et al., 2026). La consultation publique[483] confirme que les répondants des industries culturelles et médiatiques demandent en priorité des seuils opérationnels permettant de séparer ces trois situations, ainsi qu'une harmonisation des labels au niveau du segment plutôt que du fichier entier, afin d'éviter à la fois le sur-étiquetage et l'omission trompeuse dans les œuvres composites mêlant productions humaines et passages synthétiques.

6.6 Le marquage en pratique : adoption empirique et chaîne d'imputation

Au-delà de la maturité technique des solutions disponibles, **leur adoption effective dépend de la configuration de la chaîne de valeur dans laquelle elles s'inscrivent.**

[483] SPECTRA, *Analysis of the Stakeholder Consultation on Transparency Requirements for Certain AI Systems under Article 50 AI Act*.

Des auteurs distinguent quatre scénarios typiques de déploiement des générateurs d'images, qui posent chacun des questions différentes d'imputation de l'obligation de marquage[484].

Le premier rassemble les systèmes intégrés de bout en bout (Categorie 1), dans lesquels le même acteur développe le modèle de base et l'expose à l'utilisateur final (OpenAI avec DALL-E dans ChatGPT, ou Google avec Imagen dans Gemini). Cet acteur est à la fois fournisseur et déployeur au sens du RIA, et porte donc, seul, la double obligation de marquage technique (article 50(2)) et d'étiquetage de l'hypertrucage (article 50(4)).

Le deuxième scénario regroupe les applications qui s'appuient sur des modèles tiers via API (Categorie 2), créant une chaîne d'imputation plus floue entre le fournisseur de l'API et l'éditeur de l'application aval.

Le troisième, porte sur les déploiements open source via plateformes d'hébergement de modèles comme Hugging Face (Categorie 3), où les vignette actives (*widgets*) d'inférence directement utilisables soulèvent la question de savoir si la plateforme elle-même devient fournisseur du système au sens du RIA.

Le quatrième couvre les applications rebadgées (Categorie 4) qui déploient des modèles open source téléchargés sans en créditer l'origine, fréquentes parmi les applications mobiles hors UE, et où le marquage initialement intégré au code est souvent retiré (voir aussi[485]).

L'enquête empirique conduite par les mêmes auteurs sur 50 générateurs d'images grand public entre janvier et mai 2025 donne une **mesure concrète du fossé entre obligation légale et pratique observée**. Seuls 19 systèmes sur 50 (38 %) implémentent une forme de marquage lisible par machine, et seulement 9 sur 50 (18 %) appliquent une mention visible d'origine IA[486]. L'adoption se concentre presque entièrement dans la première catégorie : 8 systèmes de bout en bout sur 10 implémentent un marquage, contre une faible part des déploiements Hugging Face et des applications rebadgées. Lorsqu'un marquage est présent, il s'agit le plus souvent de métadonnées (12 systèmes), aisément supprimables ; les filigranes invisibles ne sont détectés que dans 8 systèmes. Deux constats structurels complètent ce tableau. D'abord, l'écosystème des modèles de fondation est extrêmement concentré (Stability AI, Black Forest Labs et OpenAI couvrent l'essentiel des déploiements non intégrés), ce qui rend une intervention régulatoire portée à ce niveau techniquement gérable. Ensuite, les marques intégrées par ces fournisseurs en amont ne se propagent pas systématiquement aux déploiements en aval : la bibliothèque de filigrane embarquée dans les modèles Stability AI et Black Forest Labs est, selon les auteurs, « *tout à fait facile à désactiver, par exemple en commentant une ligne de code* », et cette désactivation est effectivement observée dans la plupart des systèmes en aval.

Les enjeux économiques de la répartition concrète de la charge de mise en conformité dans toute la chaîne de valeur sont considérables. Des entreprises comme la startup française Label4.ai émanant du CNRS et de l'INRIA proposent une solution mobilisant à la fois les techniques amont et aval. Au-delà de ces initiatives publiques, **parallèlement au développement du marché des hypertrucages (cf. Chap.2), un marché en miroir émerge autour des outils défensifs** : détection automatique, marquage cryptographique, authentification de provenance, surveillance des usurpations. Les estimations placent ce marché autour de 1,5 milliard de dollars en 2025 avec un taux de croissance attendu supérieur à 40 % par an. Les acteurs principaux incluent des entreprises spécialisées (Reality Defender, Sensity AI, Pindrop, Truepic, DeepMedia), des services intégrés par les grandes plateformes (Deezer Radar, YouTube Content ID adapté à l'IA, Meta watermark detection), et des initiatives normatives ouvertes (Coalition for Content Provenance and Authenticity – C2PA, Content Authenticity Initiative). **Ce marché reste pour l'heure majoritairement aval (réaction post-production) plutôt qu'amont (prévention à la génération).**

---

[484] Rijsbosch et al., 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act'.

[485] Joëlle Farchy et Bastien Blain, *Rémunération Des Contenus Culturels Utilisés Par Les Systèmes d'IA*, 2025, https://strossburi.eu/wp-content/uploads/2025/05/CSPLA_Rapport-economique_Remuneration-des-contenus-mai-2025.pdf.

[486] Rijsbosch et al., 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act'.

7. AU-DELA DE LA TRANSPARENCE, RESPONSABILISER LES ACTEURS FACE AUX HYPERTRUCAGES

Au-delà de l'obligation de transparence, les hypertrucages soulèvent de nombreuses questions, qui se traduisent par l'ensemble des normes potentiellement invocables. En miroir de cet encadrement juridique se posent des enjeux de responsabilisation des acteurs de la chaîne de valeur, qui supposent de tenir compte des autres réglementations au regard desquelles peut être déterminé ce qui relève de l'hypertrucage admissible ou non (**7.1**), condition préalable à la détermination des responsabilités des différents acteurs de la chaîne de valeur autour de ces contenus (**7.2**).

7.1. La transparence des hypertrucages, une approche partielle des responsabilités

La transparence est le point d'entrée essentiel des hypertrucages dans le RIA. A cet égard, on relèvera que, selon l'article 1er, paragraphe 2, point *d*, **le RIA ne procède à l'harmonisation que des règles en matière de transparence applicables à certains systèmes d'IA** : si les Etats membres ne sont plus ainsi plus en mesure d'intervenir pour poser des règles de transparence sur les contenus générés ou manipulés par IA, dont les hypertrucages, l'ensemble des autres normes auxquels ces contenus sont soumis peuvent encore faire l'objet de dispositions nationales, sous réserve des autres domaines d'intervention d'une règlementation européenne.

Pourtant, ainsi que cela a déjà été exposé, les obligations sont potentiellement plus étendues, et elles le sont d'autant plus qu'il est nécessaire d'être clair sur **la portée du RIA : celui-ci définit les responsabilités des différents acteurs de la chaîne de valeur indépendamment, pour l'essentiel, de la licéité des contenus générés par les systèmes d'IA**. Cette licéité relève d'une autre appréciation, qui croise celle issue du RIA pour dresser un panorama des responsabilités, dont il convient de préciser, sans prétendre à l'exhaustivité, les principaux éléments et leur articulation avec ce type de contenu.

7.1.1. La transparence, un enjeu indépendant du contenu généré ou manipulé

Ainsi que cela a été indiqué, le considérant 137 apporte une précision importante, qui ne ressort pas expressément des termes du RIA : le respect des obligations de transparence applicables aux systèmes d'IA relevant du RIA « *ne devrait pas être interprété comme indiquant que l'utilisation du système d'IA ou de ses sorties est licite en vertu du présent règlement ou d'autres actes législatifs de l'Union et des Etats membres* ».

En effet, l'obligation de transparence vise les fournisseurs et les déployeurs d'hypertrucages, sans que cela suffise à retenir que le contenu généré soit licite. Il y a, à cet égard, **une différence d'approche avec le règlement sur les services numériques (DSA)**[487] : si celui-ci identifie des « *contenus illicites* », définis comme des « *informations* » non conformes au droit de l'UE ou d'un Etat membre[488], le RIA s'intéresse au produit de l'utilisation de l'IA pour responsabiliser les acteurs de la chaîne de valeur, sans se focaliser sur le produit de cette utilisation, sauf à travers l'évaluation des risques et les éventuelles pratiques interdites. Mais ces dernières sont elles-mêmes définies à travers l' « *utilisation* », l'utilisateur n'étant responsable que dans les limites de l'article 2, paragraphe 10, notamment. La licéité d'un hypertrucage sort pour l'essentiel du champ du RIA.

Avant d'examiner la question de la licéité d'un hypertrucage et les personnes responsables en cas de méconnaissance des différentes règles applicables, il y a lieu de préciser la **portée des obligations de marquage et d'étiquetage sur l'hypertrucage lui-même**. Dans ses lignes directrices sur les pratiques interdites en matière d'IA, la Commission européenne avance en effet, afin de « *clarifier l'interaction entre l'interdiction énoncée à l'article 5, paragraphe 1, point a), du règlement sur l'IA et les obligations de marquage des hypertrucages et de certaines publications textuelles générées par l'IA sur des questions d'intérêt public incombant au déployeur au titre de l'article 50, paragraphe 4* » que « *Le marquage visible des hypertrucages et des dialogueurs réduit le risque de tromperie susceptible de survenir une fois que le contenu généré par l'IA est diffusé au public et diminue le risque d'effets d'altération préjudiciables sur la formation de l'opinion et des convictions et sur le comportement de la personne* » (paragr. 71).

---

[487] 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).

[488] Art. 3 : « *h) «contenu illicite»: toute information qui, en soi ou par rapport à une activité, y compris la vente de produits ou la fourniture de services, n'est pas conforme au droit de l'Union ou au droit d'un État membre qui est conforme au droit de l'Union, quel que soit l'objet précis ou la nature précise de ce droit* ».

En reportant l'appréciation de l'illicéité d'une pratique au stade de la diffusion du contenu, alors que la conception même des systèmes d'IA doit respecter les interdictions posées, et en la rendant dépendante d'une personne – le diffuseur – qui n'est potentiellement pas soumise aux obligations d'étiquetage, une telle affirmation apparaît contradictoire avec l'objet même de l'interdiction de l'article 5 et avec les dispositions du RIA, en particulier les articles 50, 53 et 55, destinées à responsabiliser les fournisseurs de systèmes d'IA. En effet, les pratiques visées à l'article 5 sont interdites de manière générale et invoquer l'effet potentiel d'un étiquetage pour rendre possible des utilisations qui sont prohibées car subliminales, délibérément manipulatrices ou trompeuses, rend possible des contournements des dispositions de l'article 5, quand bien même il est impossible, pour le fournisseur d'un système d'IA, de prévenir l'ensemble des utilisations de son outil. Par ailleurs, alors que l'article 5, paragraphe 1, *b*, interdit l'utilisation des systèmes d'IA exploitant les éventuelles vulnérabilités d'une personne, il semble peu adapté de compter sur un simple étiquetage dont la compréhension ne sera pas accessible à tous et qui ne répondra pas à un modèle uniformisé en dehors des adhérents au code de bonnes pratiques. A tout le moins, cela renforce la nécessité de donner une pleine portée aux obligations des fournisseurs de systèmes d'IA permettant de générer des hypertrucages.

Par ailleurs, il a déjà été relevé par plusieurs auteurs que **le respect des obligations de marquage et d'étiquetage ne saurait rendre admissible un hypertrucage portant atteinte aux droits des personnes concernées** : le cas de la génération d'un hypertrucage représentant une personne dans un scène dégradante, portant atteinte à sa réputation[489], montre qu'**il ne faut pas confondre la licéité du système d'IA et celle du contenu généré**, ni procéder à des reports de responsabilités entre les différents acteurs de la chaîne de valeur. Les dispositions qu'il est envisagé d'introduire concernant, en particulier, les contenus à caractère sexuel dans le cadre du règlement omnibus en cours de discussion vont d'ailleurs, pour ces contenus, en ce sens.

La mission considère donc que les dispositions du RIA ne peuvent en aucune manière être interprétées comme rendant licite l'utilisation des systèmes d'IA du simple fait que les obligations de transparence ont été satisfaites et que les responsabilités prévues par les dispositions de ce règlement doivent être appréciées dans leur plénitude.

7.1.2. L'illicéité d'un hypertrucage : une qualification aux contours imprécis

Les hypertrucages, en tant que contenus générés par IA et faisant l'objet d'une diffusion, se situent à l'intersection d'un ensemble de normes, au-delà de celles protégeant les droits des personnes et dont on s'efforcera de retracer les principales lignes, sans prétendre à l'exhaustivité, notamment à l'égard de législations sectorielles qui pourraient avoir pour effet de prohiber l'utilisation de ce type de contenu. La question revêt néanmoins deux aspects, le premier se trouvant dans le RIA, le second en dehors de lui.

a) <u>Les conséquences incertaines du non-respect des obligations de transparence sur la licéité de l'hypertrucage</u>

Ainsi que cela a déjà été indiqué, le RIA régit les (seules) règles de transparence applicables aux hypertrucages et ne préjuge pas de leur licéité. L'effet du non-respect de ces règles de transparence, bien qu'encadré par le règlement, repose largement sur les mesures prises par les Etats membres et ce régime de sanction, en cohérence avec les règles de fond posées, ne vise pas les contenus, mais les fournisseurs et déployeurs. En effet, l'article 99 comporte plusieurs niveaux de sanctions :

- Le paragraphe 1 renvoie aux Etats membres la responsabilité de déterminer le régime de sanctions « *et autres mesures d'exécution* », applicables aux violations du RIA commises par des opérateurs, c'est-à-dire les fournisseurs et déployeurs concernant les obligations de l'article 50[490] ;
- Le paragraphe 3 précise que le non-respect de l'interdiction des pratiques mentionnées à l'article 5 fait l'objet d'amendes administratives pouvant aller jusqu'à 35 M€ ou 7 % du chiffre d'affaires annuel mondial pour les entreprises ;
- Le paragraphe 4 sanctionne, notamment, le non-respect des obligations de transparence pesant sur les fournisseurs et déployeurs en application de l'article 50 d'une amende administrative pouvant aller jusqu'à 15 M€ ou 3 % du chiffre d'affaires annuel mondial.

[489] Voy. l'exemple de l'hypertrucage à caractère pornographique pris par Mateusz Łabuz, « Deep Fakes and the Artificial Intelligence Act—An Important Signal or a Missed Opportunity? », *Policy & Internet*, vol. 16, n° 4, 2024, p. 783.
[490] Selon le 8 de l'article 3, un opérateur est « *un fournisseur, fabricant de produits, déployeur, mandataire, importateur ou distributeur* ».

Le RIA laisse en revanche en suspens le sort de l'hypertrucage qui ne respecterait pas les règles de transparence, question qui est donc renvoyée aux Etats membres, puisqu'elle porte sur la licéité du contenu diffusé.

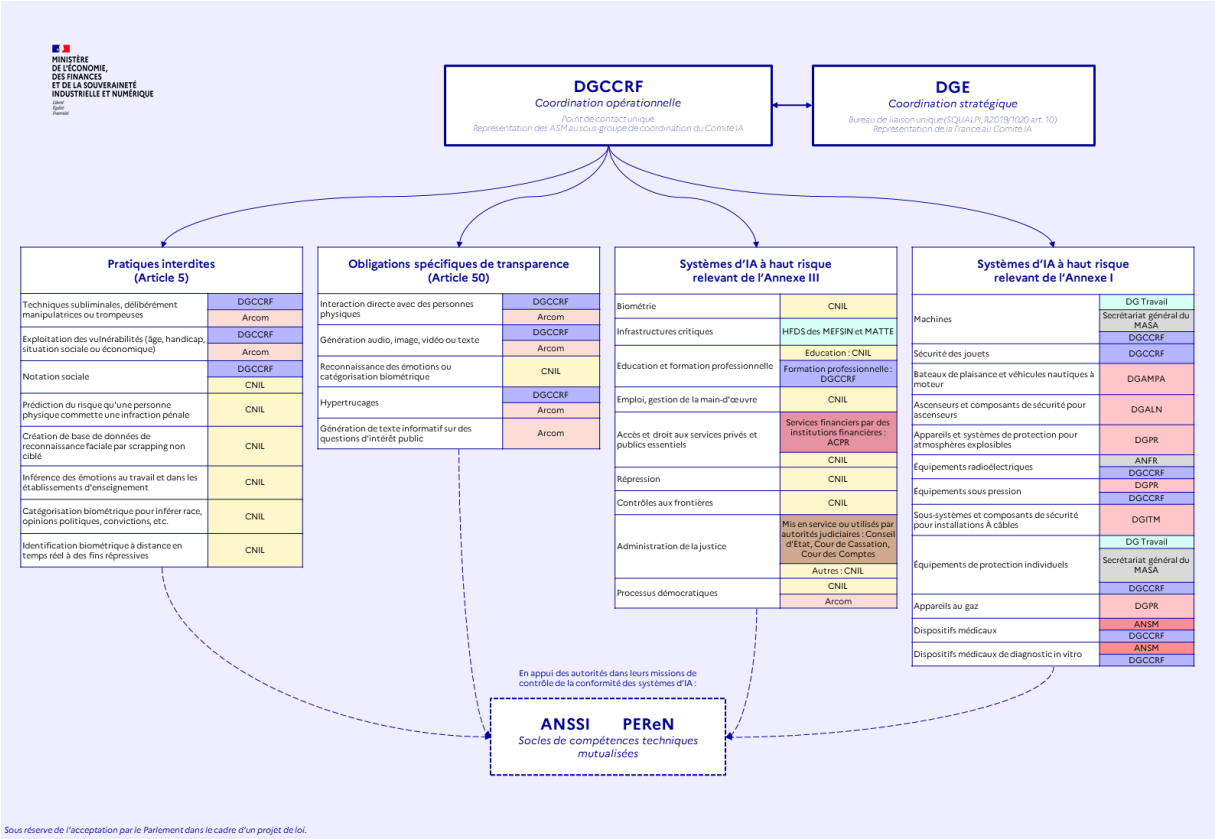
Le projet de lignes directrices sur l'article 50, soumis à consultation par la Commission européenne en mai 2026 (paragr. 91) retenait une lecture ambitieuse : est illicite au sens du DSA un contenu soumis aux obligations des paragraphes 2 ou 4 de l'article 50 ne répondant pas aux obligations de transparence prévues par ces dispositions. Une telle conclusion n'est pas immédiate à la lecture du RIA et peut prêter à hésitations, d'autant que le RIA ne régit pas, en soi, la licéité des contenus générés par IA. Cette interprétation n'est d'ailleurs plus celle retenue dans la version finale des lignes directrices (paragraphe 96).

En France, le projet de loi portant diverses dispositions d'adaptation au droit de l'Union européenne en matière économique, financière, environnementale, énergétique, d'information, de transport, de santé, d'agriculture et de pêche, en cours d'examen au Parlement, prévoit de confier, essentiellement, à l'Arcom la surveillance du respect des dispositions de l'article 50, paragraphes 2 et 4 (voy. le schéma ci-après). Le V de l'article 20-12 introduit dans la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication par l'article 24 *terdecies* du projet de loi confère, dans sa version adoptée en première lecture par le Sénat, des pouvoirs à l'Arcom qui sont essentiellement structurels mais qui sont vraisemblablement susceptibles d'être mobilisés en réponse à la diffusion d'un hypertrucage. L'Autorité pourra en effet, entre autres, enjoindre à l'opérateur concerné de mettre fin à la mise sur le marché, à la mise en service ou à *l'utilisation* d'un système d'intelligence artificielle dans un délai déterminé ou de prendre toute mesure corrective de nature structurelle ou comportementale proportionnée au manquement et nécessaire pour faire cesser dans un délai déterminé celui-ci ; elle pourra également, en application de l'article 20-13, après cette injonction ou immédiatement, prononcer une sanction administrative, ainsi que le prévoit le RIA. Dans ce cadre, compte tenu de la référence à l'utilisation du système d'IA, à laquelle elle pourra demander qu'il soit mis fin, **l'Autorité devrait pouvoir demander le retrait d'un contenu hypertruqué ne répondant pas aux obligations de transparence, auprès, selon le cas, du fournisseur ou du déployeur qui aura manqué à ses obligations.** Ce pouvoir est, en cohérence avec le RIA, cantonné aux professionnels.

Néanmoins, en sanctionnant le seul non-respect des obligations de transparence, **l'ensemble de ces dispositions ne permet pas de considérer que l'hypertrucage ainsi diffusé serait en lui-même illicite** ; c'est le défaut de paramétrage du système (pour le fournisseur) et la diffusion de l'hypertrucage (pour le déployeur) qui sont sanctionnés. L'harmonisation à laquelle procède le règlement est d'ailleurs, ainsi que nous l'avons indiqué, limitée aux systèmes d'IA. Il peut à cet égard être relevé *a contrario*, et à titre d'exemple, que l'article 1-3 de la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (LCEN) qualifie expressément d'illicite au sens du DSA les contenus ne respectant l'obligation de marquage prévue à cet article (avertissement du caractère illégal des comportements représentés par des contenus à caractère pornographique simulant la commission d'un crime ou d'un délit) ou que l'article L. 335-3-2 du code de la propriété intellectuelle sanctionne pénalement le fait de supprimer sciemment, en particulier, le tatouage numérique relatif aux droits de propriété intellectuelle prévu à l'article L. 331-11.

Or, le caractère illicite du contenu conditionne l'ensemble des procédures prévues par le droit de l'UE, en particulier aujourd'hui dans le cadre du DSA, et par le droit national, notamment la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique[491], qu'il s'agisse des procédures administratives ou judiciaires[492], même si ce sont désormais, depuis l'entrée en vigueur du DSA, tous les contenus illicites qui sont couverts[493].

[491] Suzanne Lequette, *Droit du numérique*, Paris, LGDJ-Lextenso, 2024, p. 177 et s.
[492] Sur ces procédures, voy. Myriam Quéméner, V° « Sites internet. Retrait, blocage, déréférencement », *JurisClasseur Communication*, fasc. 690, mise à jour 2025.
[493] Grégoire Loiseau, « Le *Digital Services Act* », *Communication Commerce électronique*, n° 2, fév. 2023, étude 3.



Source : site internet de la direction générale des entreprises[494]

Avant de revenir plus précisément sur ces procédures lorsque le contenu hypertruqué est qualifiable d'illicite, il faut relever que **le principal moyen d'action face à un hypertrucage ne respectant pas les règles de l'article 50 du RIA, hors action de l'Arcom telle que prévue par le projet de loi DDADUE en cours d'examen, pour une personne s'estimant victime d'un tel contenu, est d'user de la procédure spéciale de l'article 6-3 de la LCEN** (art. 6, I, 8 avant la loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique). Cette disposition, dont la portée a été élargie par le législateur[495] et qui – ce qui est déterminant – permet d'agir sans que l'action soit directement dirigée contre le responsable d'un contenu dommageable[496], permet de saisir le président du tribunal judiciaire afin de prescrire, selon la procédure accélérée au fond, à toute personne susceptible d'y contribuer toutes les mesures « *propres à prévenir un dommage ou à faire cesser un dommage occasionné par le contenu d'un service de communication au public en ligne* ». La notion de « dommage » n'est pas précisée[497] et la jurisprudence n'a pas encore eu l'occasion de délimiter entièrement les contours de cette notion. En l'état des textes, la jurisprudence retient, lorsque la procédure n'est pas contradictoire avec l'auteur et au regard, par suite, de la nécessité d'identifier un préjudice évident[498], la **nécessité de démontrer le caractère manifestement illicite du contenu**[499], en cohérence avec la définition des responsabilités des acteurs susceptibles d'être contraints à agir dans ce cadre (ces responsabilités étant définies en référence au caractère illicite du contenu). Revenant à la logique prévalant pour le référé, et même si certains auteurs

494 https://www.entreprises.gouv.fr/priorites-et-actions/transition-numerique/soutenir-le-developpement-de-lia-au-service-de-0

495 Art. 39 de la loi n° 2021-1109 du 24 août 2021 confortant le respect des principes de la République.

496 Voy. à ce sujet Eva Lee, Benoît Huet, « L'immunité paradoxale offerte aux auteurs anonymes de contenus diffamatoires » *Légipresse*, 26 juillet 2024, n° 427.

497 Ce qui a été largement regretté : outre l'article précité d'Eva Lee et Benoît Huet, Christophe Bigot, « La procédure accélérée au fond prévue par l'article 6-I-8 de la LCEN : un risque de déstabilisation profonde des fragiles équilibres du droit de la presse », *Légipresse* 2023, p. 601.

498 Emmanuel Binoche, « L'articulation entre référé de droit commun et référé de la loi du 21 juin 2004, art. 6-I.8 LCEN », *Legicom* 2006/1, n° 35 ; Christophe Bigot, art. préc..

499 Cass. 1re Civ., 26 février 2025, pourvoi n° 23-16.762, au Bull. Rappr. Cass. 1re Civ., 26 février 2025, pourvoi n° 23-22.386, au Bull.

ont largement critiqué cette perspective[500], le dommage est établi à la lumière de l'illicéité l'ayant provoqué – ce qui, de notre point de vue, est cohérent puisque l'identification de la faute à l'origine (au moins probable) du dommage est nécessaire pour déterminer les mesures pertinentes et apprécier leur proportionnalité, au regard de la balance des droits en présence ; c'est en outre nécessaire pour que cette action corresponde au cadre européen posé par le DSA (voy. *infra*).

Dès lors, sauf à pouvoir invoquer un autre motif d'illicéité (et l'hypothèse est loin d'être négligeable au regard des termes de l'article 226-8 du code pénal qui sanctionne dans certains cas les montages, sauf s'ils sont marqués ou si leur caractère artificiel est évident), le seul manquement aux obligations de transparence ne suffira pas, en général, à actionner cette procédure puisque l'illicéité, telle qu'elle résulte de la lettre de l'article 50 du RIA, porte sur le comportement du fournisseur du système ou le déployeur, et non sur le contenu généré en litige ; l'enjeu tenant à la qualification d'illicite au sens du DSA d'un hypertrucage ne répondant pas aux obligations du RIA, ainsi que le proposait le projet de lignes directrices de la Commission, est donc fondamental.

A la lumière de ces éléments, compte tenu du renvoi fait par le RIA au droit des Etats membres pour décider de l'illicéité d'un contenu, **la mission estime indispensable, sans attendre une confirmation du législateur européen dans le RIA ou par la Cour de justice de l'UE, d'engager une réflexion sur l'opportunité d'ancrer dans le droit national une telle qualification à l'égard d'hypertrucages qui ne satisferaient pas aux obligations de l'article 50 du RIA**.

Une telle qualification élargirait le champ des responsabilités préexistantes à l'égard des contenus illicites à la méconnaissance de ces obligations. La mission, bien que consciente de la difficulté opérationnelle et juridique[501] que pose le caractère asymétrique de l'obligation d'étiquetage qui pèse sur les seuls déployeurs, estime qu'une telle voie apparaît comme la plus pertinente pour lutter efficacement contre la diffusion de tels hypertrucages. Au demeurant, les dispositions du RIA, et notamment l'article 2, paragraphe 10, ne permettent pas de considérer que le législateur européen aurait entendu exclure que les Etats membres réglementent les hypertrucages générés et diffusés par des utilisations non-professionnels. Ces éléments renforcent d'autant plus la nécessité de donner son plein effet à l'obligation de marquage pesant sur les fournisseurs, doublée à celle de mettre à disposition des outils de détection.

Par ailleurs, on relèvera que, contrairement aux obligations de l'article 50, celles des articles 53 et 55, pourtant indissociables pour appréhender l'hypertrucage dans le RIA, ne font pas l'objet de mesures de sanctions dans ce règlement ; seuls des pouvoirs de demande d'informations sont prévus par les articles 91 et 93 au bénéfice de la Commission européenne. Si, là aussi, l'article L. 336-2 du CPI peut compenser une partie du défaut de prise en compte des droits de propriété intellectuelle, **il semble nécessaire que les Etats membres se saisissent de la question, même si elle est circonscrite à l'utilisation des systèmes d'IA à caractère général,** dès lors que le règlement harmonise les règles relatives à leur mise sur le marché.

b) <u>L'illicéité de l'hypertrucage au regard de son contenu</u>

L'élément central de l'équilibre juridique retenu par le législateur dans le DSA est, comme on l'a déjà indiqué, la qualification d'un contenu comme étant *illicite*[502], bien que ce règlement ne procède pas en lui-même à cette qualification. A cet égard, l'articulation entre le RIA, qui s'intéresse aux systèmes d'IA et aux acteurs de la chaîne de valeur, et le DSA, qui se focalise sur les contenus et les procédures permettant de lutter contre les contenus illicites, est essentielle. L'article 2, paragraphe 5, du RIA, selon lequel ce règlement n'affecte pas l'application des dispositions relatives à la responsabilité des fournisseurs de services intermédiaires énoncées au chapitre II du DSA[503], est central pour déterminer les responsabilités des différents acteurs à l'égard des hypertrucages.

---

[500] Notamment Emmanuel Dreyer, « Errements dans la procédure accélérée au fond engagée contre un hébergeur », *Legipresse,* n° 406, p. 481. Voy. également Nicolas Verly, « Suppression de contenus diffamatoires par l'hébergeur dans le cadre d'une procédure accélérée au fond : une première lumière au bout du tunnel ? », *Légipresse,* 2025, p. 278.

[501] Il ne s'agit toutefois pas de créer une infraction pénale, mais d'assurer l'opérationnalité des procédures afin de sanctionner le non-respect d'obligations de nature civile.

[502] S. Lequette, *op. cit.*, p. 177-178.

[503] 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).

Pour autant, dès lors que le RIA ne pose pas de principe de fond de licéité des hypertrucages, à l'exclusion des pratiques interdites de l'article 5 et sous réserve des éléments exposés ci-dessus quant à la portée de l'obligation de transparence au regard du DSA, il faut, pour déterminer les responsabilités, **procéder au préalable à la qualification de l'hypertrucage comme étant illicite**. Cette opération de qualification ne revêt pas de particularité par rapport aux autres contenus circulant en ligne – et d'autant moins si le non-respect des obligations de transparence ne rend pas l'hypertrucage illicite. Compte tenu du vaste champ des illicéités susceptibles de toucher les hypertrucages, on formulera ici quelques observations axées essentiellement sur le champ de la mission, à savoir les secteurs culturel et créatif[504].

Au préalable, il faut néanmoins **préciser ce qu'on entend par un contenu illicite**, au-delà de la définition donnée par le DSA d'un contenu non conforme à des règles de droit de l'UE ou national. En effet, s'il ne fait guère de doute que le caractère illicite découle nécessairement de la méconnaissance d'une disposition pénalement sanctionnée (ce qui est le cas par exemple d'une diffamation, d'un acte de contrefaçon ou d'une violation des articles 226-8 et 226-8-1 du code pénal), on peut avoir une hésitation sur l'étendue de la notion s'agissant de la méconnaissance de droits de nature civile. La notion d'acte illicite n'est pas inconnue en droit et en procédure civils ; elle figure même à l'article 835 du code de procédure civile. La cessation de l'illicite est identifiée comme une des composantes de la responsabilité civile extracontractuelle[505], bien que sa place exacte soit parfois discutée[506]. Pour autant, cette responsabilité peut être engagée indépendamment d'une faute, et donc d'une illicéité au sens strict. En outre, la place précise de l'illicéité dans l'engagement de la responsabilité n'est pas aussi claire dans la jurisprudence, dans la mesure où elle détermine surtout les mesures que le juge peut (ou, dans certains cas, doit) prononcer[507].

Mais, si l'illicite en droit français recouvre le champ civil, le doute sur l'intention du législateur européen est permis dès lors que, à la lecture du DSA, il apparaît qu'alors que celui-ci se réfère abondamment aux contenus illicites, il définit les obligations des fournisseurs de services intermédiaires aux articles 5 à 7 par référence à la notion d'*infraction*, notion reprise au considérant 25. La lecture combinée des articles 4, paragraphe 3, 5, paragraphe 2, et 6, paragraphe 4, ainsi que 9, semble aller dans le sens d'une assimilation d' « infraction » à « contenu illicite ». Pour autant, on observera que le DSA renvoie expressément au droit national la définition des cas d'illicéité et que d'autres dispositions semblent montrer que les deux termes sont utilisés de manière distincte, l'un au regard du risque de responsabilité pénale (voy. en particulier les articles 18 et 51), l'autre pour définir un champ plus large d'intervention des fournisseurs de services intermédiaires.

A la lumière de ces éléments, **la mission considère que la faute civile relève du champ de l'illicéité au sens du DSA.** Cela est déterminant au regard des enjeux posés par les hypertrucages. En effet, parmi les principaux risques qui, au regard du champ du présent rapport, sont susceptibles de se rencontrer sur les hypertrucages, nombreux sont ceux qui relèvent du champ pénal, et qui ne relèvent pas des infractions listées au A du IV de l'article 6 de la LCEN :

- Contrefaçon, telle que définie et sanctionnée par L. 335-3 et L. 335-3 du code de la propriété intellectuelle ;
- Infractions de presse, telle que définies par la loi du 29 juillet 1881 relative à la liberté de la presse. On rappellera à cet égard que le Conseil constitutionnel a estimé que les dispositions existantes, tant dans cette loi que dans le code pénal, réprimant les abus de la liberté d'expression constituaient un ensemble suffisant pour réprimer ces faits, y compris lorsqu'ils sont commis par l'utilisation d'un service de communication au public en ligne, de sorte qu'il a censuré le nouveau délit que le législateur avait voulu introduire dans la loi SREN[508] ;

---

[504] Nous n'évoquons notamment par les dispositions spéciales en matière de lutte contre le terrorisme, la pédocriminalité ou concernant les contenus à caractère pornographique, ni les dispositions en matière commerciale ou portant sur des droits propres à certains domaines, comme en matière de santé.

[505] Notamment Cyril Bloch, *La cessation de l'illicite. Recherche sur une fonction méconnue de la responsabilité civile extracontractuelle*, Paris, Dalloz, Nouvelle bibliothèque des thèses, vol. 71, 2008. Également : Geneviève Viney, « Cessation de l'illicite et responsabilité civile », *in Mélanges Goubeaux*, Paris, Dalloz-LGDJ, 2009, p. 547 ; Cyril Bloch et Philippe Stoffel-Munck, « La cessation de l'illicite », *in* François Terré (dir.), *Pour une réforme du droit de la responsabilité civile*, Paris, Dalloz, 2011, p. 87.

[506] Philippe Brun, V° « Responsabilité du fait personnel », *Répertoire de droit civil*, Dalloz, mai 2015, paragr. 161 ; Fabrice Leduc, V° « Régime de la réparation – modalités de la réparation – Règles communes aux responsabilités délictuelle et contractuelle – principes fondamentaux », *Jurisclasseur Responsabilité civile et Assurances*, fasc. 201, mai 2024, paragr. 27.

[507] Fabrice Leduc, fasc. préc., paragr. 28.

[508] Décision n° 2024-866 DC du 17 mai 2024, paragr. 75.

- Infraction aux dispositions du code électoral destinées à lutter contre la désinformation, issues de la loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (art. L. 112 du code électoral, renvoyant à l'article L. 163-1). S'y ajoute également le délit de diffusion de fausses nouvelles de l'article 332-14 du code pénal ;
- Escroquerie, abus de confiance, potentiellement chantage et extorsion, étant précisé que, sauf pour le chantage, le recours à des outils numériques n'est pas spécialement pris en compte par le législateur, y compris dans le cadre de circonstances aggravantes.

Il est toutefois au moins un domaine où l'illicéité relève pour l'essentiel d'une méconnaissance des règles du droit civil : il s'agit de celui portant sur les droits de la personnalité et la vie privée. Si certaines violations sont pénalement sanctionnées, les dispositions de l'article 226-1 du code pénal sont restrictives par rapport au champ de ces droits. La prise en compte des fautes civiles pour la qualification d'un contenu comme illicite revêt alors une importance particulière, même si elle **implique que les différents acteurs chargés de la mise en œuvre du DSA interviennent dans des questions relevant de relations de droit privé**. Cette inclusion est, pour la mission, d'autant plus importante, qu'elle conditionne, en l'état des textes, la capacité à agir contre des hypertrucages méconnaissant ces droits.

En revanche, alors que l'action de l'article 6-3 de la LCEN devrait être la plus large pour agir face à ce type de contenu (point suivant), on ne peut manquer de relever un miroitement entre la condition de mise en œuvre de cette procédure (le dommage) et le critère d'illicéité fixé par le DSA : le dommage résultant d'un fait non fautif n'est pas susceptible de relever de l'illicéité au sens du DSA et des dispositions prises pour son application. C'est là un enjeu important concernant une procédure qui permet d'agir à l'égard d'autres personnes que l'auteur du contenu, de sorte qu'il faut les responsabiliser afin qu'elles soient fautives en cas d'inaction.

c) Au-delà de l'illicéité, (re)penser la responsabilité des fournisseurs

Au-delà de ces différents aspects qui concernent directement les hypertrucages, la mission a relevé avec intérêts des **propositions tendant à développer un droit de la responsabilité de l'IA spécifique, s'inspirant de la responsabilité des produits défectueux**[509].

Pour des opérateurs qui, étant fournisseurs de services intermédiaires, s'exposaient à une responsabilité limitée, cela constituerait un changement important, mais qui répond à l'évolution de leur position : ce changement est général[510], mais, au regard du champ de la présente mission, il s'agit de tenir compte du fait qu'en mettant à disposition un système d'IA générative, ils offrent un outil susceptible de causer directement des dommages dont ils sont les seuls à connaître l'ensemble des modalités de fonctionnement et de paramétrage. Face à ce qui relève d'une « boîte noire », on peut certes estimer que les responsabilités des différents acteurs de la chaîne de valeur ayant conduit à la diffusion d'un hypertrucage préjudiciable sont invocables, mais cela suppose la mise en cause d'une multitude d'acteurs[511], dans une configuration procédurale qui s'oppose au caractère potentiellement massif des atteintes en jeu. **Or, la protection contre les hypertrucages ne peut être efficace, sans préjudicier aux droits et libertés des personnes, que si elle repose sur une intervention aisée du juge** et, malgré les apports substantiels de l'article 6-3 de la LCEN, celui-ci reste perfectible et ne permet pas d'obtenir réparation du préjudice subi.

Il faut néanmoins relever que la **nouvelle directive sur la responsabilité du fait des produits défectueux**[512] comporte déjà des avancées. Son champ d'application est en effet étendu aux logiciels et, bien que l'IA ne soit pas mentionnée dans les articles de la directive, mais seulement au considérant 13[513], cette notion de logiciel semble, dans ce cadre précis, ainsi que le suggère le considérant 13, pouvoir être le

[509] Kenneth S. Abraham, Catherine M. Sharkey, « Untangling AI Liability », *Virginia Public Law and Legal Theory Research Paper No. 2026-19*, 23 février 2026, à paraître à la *California Law Review.*

[510] Olivier Sylvain, *Reclaiming the Internet. How Big Tech took control and how we can take it back*, Columbia Global Reports, 2026, p. 106.

[511] Voy. à ce sujet, Miriam Buiten, « Product liability for defective AI », *European Journal of Law and Economics*, 2024, vol. 57, p. 239 ; Fabrice Leduc, « Dommages causés par l'intelligence artificielle - La réparation des dommages causés par l'intelligence artificielle », *Responsabilité civile et assurances,* n° 1, janvier 2026, étude 1.

[512] Directive (UE) 2024/2853 du Parlement européen et du Conseil du 23 octobre 2024 relative à la responsabilité du fait des produits défectueux et abrogeant la directive 85/374/CEE du Conseil.

[513] Jonas Knetsch, « La nouvelle responsabilité du fait des produits défectueux : état des lieux et analyse critique », JCP G 2025, doctr. 125, n° 7

support de prise en compte de l'IA eu égard au contexte de la directive et notamment à la référence aux capacités du produit à « *poursuivre son apprentissage* » après sa mise sur le marché ou sa mise en service, de sorte que la directive peut être considérée comme applicable aux systèmes d'intelligence artificielle[514]. Pour autant, cette avancée est limitée par les conditions permettant l'engagement de la responsabilité sans faute de ce fait, en particulier s'agissant des dommages concernés[515] : il s'agit, selon l'article 6, des dommages corporels, y compris psychologiques, et des dommages causés à des biens, ainsi que de la destruction ou la corruption de données. **Les préjudices liés aux droits des personnes ne sont donc pas susceptibles d'être invoqués pour obtenir réparation**, sauf pour ce qui relève des conséquences psychologiques.

En outre, si la nouvelle directive évoque les produits poursuivant leur apprentissage pour évaluer la défectuosité, elle est silencieuse sur l'effet réel de cet apprentissage sur la responsabilité, le fournisseur du SIA pouvant être tenté de faire valoir que l'évolution de son produit n'est pas de son fait. En revanche, elle comporte des mécanismes de présomption en cas de refus de répondre à l'injonction de divulgation de preuves qui peut s'avérer adapter à l'IA[516].

Il n'en reste pas moins que l'IA soulève des enjeux spécifiques pour la rendre pleinement opérationnelle. Ainsi, la notion de « fabricant » laisser une incertitude sur son identification entre le fournisseur ou, dans certains cas, potentiellement le déployeur[517]. De plus, le caractère international des fournisseurs de SIA interroge la capacité à les mettre en cause de manière effective, alors qu'il a été constaté que la précédente directive sur les produits défectueux a généré un contentieux très peu fréquemment transfrontière[518]. Or, la directive s'en tient à une approche unifiée, ne semblant pas tenir compte du RIA négocié en parallèle, alors que celui-ci mériterait une réflexion plus spécifique sur les enjeux de la responsabilité en matière d'IA[519], à tout le moins, pour le champ du présent rapport, s'agissant de l'IA générative.

Une telle évolution ne pourrait, en tout état de cause, être portée qu'au niveau européen.

7.2. Responsabiliser pleinement l'ensemble des acteurs autour de la génération et la diffusion des hypertrucages

Les textes européens et nationaux organisent un ensemble complexe de procédures destinées à responsabiliser les différents opérateurs des services numériques, à partir de l'idée fondamentale que, compte tenu des particularités de la société de l'information, les responsabilités ne peuvent se borner à être engagées à l'égard des seuls auteurs de l'illégalité ou du dommage, de sorte qu'il est nécessaire d'obtenir le concours de l'ensemble des opérateurs afin de prévenir et limiter autant que possible les effets de cette illégalité ou dommage.

La confrontation de cet ensemble aux hypertrucages et aux enjeux qu'ils soulèvent pour les secteurs de la culture et de la création met en évidence des responsabilités très variables et une marge d'amélioration afin de protéger efficacement les droits des personnes face à ces contenus.

7.2.1. Une multiplicité de procédures pour des opérateurs aux responsabilités diverses

[514] Grégoire Loiseau, « Responsabilité du fait des produits défectueux : l'adaptation au numérique de la responsabilité du fait des produits défectueux », *Communication Commerce électronique*, n° 5, mai 2025, comm. 44 ; Jean-Denis Pellier, « Regard sur la nouvelle directive relative à la responsabilité du fait des produits défectueux », *Revue des contrats*, n° 2, juin 2025.

[515] Voy. à ce sujet Julie Charpenet, « Les nouvelles conditions de la responsabilité des produits défectueux à l'aune de l'IA », *Dalloz IP/IT*, 2025, p. 320.

[516] F. Leduc, art. préc.

[517] Luc Grynbaum, « Responsabilité du fait des produits défectueux : Les produits numériques et systèmes d'IA, Colloque à la Cour de cassation (24 septembre 2025) : les 40 ans de la directive du 25 juillet 1985 relative à la responsabilité du fait des produits défectueux », *Responsabilité civile et assurances,* n° 12, décembre 2025, dossier 26.

[518] Jean-Sébastien Borghetti, « Taking EU Product Liability Law Seriously : How Can The Product Liability Directive Effectively Contribute to Consumer Protection ? », *French Journal of Legal Policy*, 2023.

[519] Voy. par ex. Beatriz Botero Arcila, « AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight? », *Computer Law & Security Review*, 2024, vol. 54, 106012 ; Amedeo Rizzo, Giorgio Hassan, « On the Distinction Between Tort-Based and Risk-Based AI Regulation: A Transatlantic Perspective », *Transatlantic Technology Law Forum*, Stanford Law School, n° 131, 2025.

Selon le droit commun de la responsabilité, les opérateurs qui agissent pour leur propre compte, et non comme simples services intermédiaires, engagent leur responsabilité dans les conditions de droit commun s'ils mettent à disposition des contenus illicites[520]. Le chapitre II du DSA, en cohérence avec le principe de neutralité et l'exemption de responsabilité qui en découle, ne pose pas d'obligation pour les fournisseurs de services intermédiaires, tels que définis au *g* de l'article 3, y compris les plateformes en ligne, d'agir d'initiative à l'encontre d'un contenu illicite. L'article 4, paragraphe 3 (transport), l'article 5, paragraphe 2 (mise en cache) et l'article 6, paragraphe 4 (hébergement) réservent la possibilité pour une autorité judiciaire ou administrative, conformément au droit national, d'exiger du fournisseur qu'il mette fin à une *infraction* ou la prévienne, sans préjudice de la faculté de procéder à des enquêtes volontaires ou de détecter, identifier et retirer d'initiative des contenus illicites (art. 7). Ils sont en revanche tenus d'agir à réception d'une injonction prononcée par les autorités administratives ou judiciaires compétentes en vertu du droit national (art. 9 et 10). L'article 2 du RIA réserve expressément l'application de ces dispositions lorsque les opérateurs qui relèvent de ses dispositions sont également des fournisseurs de services intermédiaires[521].

Or, les dispositions impliquant une initiative des fournisseurs de services intermédiaires, et notamment les fournisseurs de services d'hébergement et plus encore les fournisseurs de plateformes figurent au chapitre III, relatif aux obligations de diligence des fournisseurs de services intermédiaires, notamment en ce qu'ils doivent permettre à tout particulier ou à toute entité de leur signaler la présence au sein de leur service d'éléments d'information spécifiques que le particulier ou l'entité considère comme du contenu illicite, la notification ainsi reçue étant assimilée à l'information permettant à un service d'hébergement d'agir (art. 16). Les fournisseurs de plateforme en ligne sont en outre tenus de mettre en place un système interne de traitement des réclamations (art. 20). Ce silence du RIA à l'égard du chapitre III du DSA ne semble toutefois pas être une exclusion : le considérant 11 mentionne qu'il est, sans autre restriction, « *sans préjudice des dispositions relatives à la responsabilité des fournisseurs de services intermédiaires* » prévues par le DSA et le considérant 118 présente le RIA comme « *complétant* » les obligations incombant à ces fournisseurs.

Il en résulte que :

- Les fournisseurs de services intermédiaires doivent agir, selon les procédures prévues par le chapitre II du DSA et précisées en droit national, s'ils sont saisis par une autorité administrative ou judiciaire d'un hypertrucage constituant un contenu illicite ;
- Les mécanismes de notification et de traitement des réclamations que, conformément au chapitre III du DSA, les fournisseurs de services d'hébergement et les fournisseurs de plateformes en ligne sont, respectivement, tenus de mettre en place au bénéfice des particuliers, s'appliquent à un hypertrucage constituant un contenu illicite.

En revanche, **les fournisseurs de système d'IA générative ne sont soumis à ces obligations que pour autant qu'ils puissent, de manière cumulative, être qualifiés de fournisseurs de services intermédiaires, de services d'hébergement ou de plateformes en ligne**. En principe, le DSA s'intéressant à la responsabilité des opérateurs proposant des services de la société de l'information, tandis que le RIA s'intéresse à la licéité des contenus, les perspectives sont distinctes et **le cumul de responsabilités au titre des deux règlements ne s'infère que du cumul de qualifications de l'opérateur**.

A cet ensemble, il faut ajouter la responsabilité propre des fournisseurs de très grandes plateformes en ligne et très grands moteurs de recherche, en particulier les obligations d'atténuation des risques déjà évoquées (art. 35). Si la notion de risque dans le DSA diffère de celle retenue dans le RIA[522], il n'en reste pas moins que les risques systémiques au sens du DSA couvrent à la fois la diffusion de contenus illicites et ceux portant atteinte aux droits fondamentaux (art. 34, paragr. 1), qui sont au cœur des enjeux soulevés par les hypertrucages pour les secteurs culturels et créatifs. **Les obligations pesant sur ces très grands opérateurs sont les plus à même de prévenir ou, à défaut, limiter la diffusion d'hypertrucages**

---

[520] Agnès Lucas-Schloetter, V° La responsabilité des fournisseurs de services de partage de contenus en ligne (CPI ; art. L. 137-1 à L. 137-4), *JurisClasseur Propriété littéraire et artistique*, fasc. 1615, juin 2025.

[521] Voy. également Anne-Sophie Chavent-Leclère, « Lutte contre les contenus illicites en ligne : la mutation de la responsabilité des fournisseurs de services intermédiaires », *AJ Pénal*, 2025, p. 181.

[522] Felipe Romero Moreno, « Generative AI and deepfakes: a human rights approach to tackling harmful content », *International Review of Law, Computers & Technology*, vol. 38, 2024, p. 297 ; Łabuz, Mateusz. « Deep Fakes and the Artificial Intelligence Act—An Important Signal or a Missed Opportunity? » *Policy & Internet* 16, n° 4, 2024, p. 783-800.

illicites, y compris ceux dont des personnes agissant à titre non-professionnel (et potentiellement anonymes) sont les auteurs. La mission en déduit que **ces fournisseurs ont une responsabilité particulière à l'égard de la diffusion de ces contenus** et qu'ils doivent intégrer les risques s'y rapportant dans leur politique d'évaluation et d'atténuation des risques. Elle observe d'ailleurs, en ce sens, que dans le cadre de l'enquête ouverte par le parquet du tribunal judiciaire de Paris concernant la plateforme X, en particulier du fait des possibilités ouvertes par son outil d'IA Grok, les investigations portent notamment sur le délit d'administration d'une plateforme en ligne illicite en bande organisé, prévu à l'article 323-3-2 du code pénal. Il faut à cet égard rappeler que le rapport Delaporte-Vojetta a préconisé d'étendre cette infraction aux contenus illicites[523], ce qui serait de nature à renforcer la capacité d'action contre les contenus circulant sur les plateformes.

Au niveau national, on peut distinguer deux grandes séries de procédures permettant de répondre à des contenus illicites. Il s'agit d'une part de l'ensemble des dispositions prises pour des infractions précises, qu'il s'agisse du terrorisme, de la pédocriminalité ou de celles énumérées au A du IV de l'article 6 de la LCEN : des obligations détaillées sont alors mises à la charge des différents opérateurs, en particulier les fournisseurs de services d'hébergement, afin, au regard de la particulière gravité des enjeux, de concourir à la lutte contre la diffusion de contenus criminels ou délictuels. Ces infractions n'ont, pour l'essentiel, pas d'incidence sur la problématique du présent rapport et les procédures y afférentes ne sont pas évoquées ; elles sont applicables aux hypertrucages dès lors que ceux-ci entrent dans les qualifications énumérées, ce qui correspond à des atteintes à l'ordre et à la sécurité publics qui sortent du champ de l'étude.

D'autre part, et en dehors de ces infractions limitativement énumérées, la mise en œuvre des mesures permettant de prévenir ou mettre fin à la diffusion de contenus illicites repose sur deux piliers :

- **Une action administrative, confiée à l'Arcom par l'article 8-1 de la LCEN**. A ce titre, l'Autorité peut notamment adresser des injonctions aux fournisseurs de services intermédiaires, afin de « *mettre fin à un ou plusieurs manquements* » ou de « *prendre toute mesure corrective de nature structurelle ou comportementale* ». Elle peut également « *adopter des injonctions à caractère provisoire lorsque le manquement constaté paraît susceptible de créer un dommage grave* » et saisir l'autorité judiciaire (A du V de l'article 9-1), ainsi que prononcer une sanction pécuniaire en cas de non-respect d'une de ses injonctions (art. 9-2). On pourrait également ajouter l'action administrative de la DGCCRF dans le cadre des pouvoirs qu'elle tient du code de la consommation, mais ceux-ci intéressent moins le champ de la présente étude ;
- **Une intervention judiciaire, prévue à l'article 6-3 de la LCEN**, dont les termes ont déjà été évoqués et qui permet, à toute personne subissant ou susceptible de subir un dommage du fait du contenu d'un service de communication au public en ligne, d'obtenir les mesures pertinentes.

L'équilibre et l'intérêt de l'ensemble de ces procédures reposent sur la possibilité d'obtenir des mesures correctives de la part d'acteurs qui, sans être responsables du contenu illicite, sont en mesure d'agir pour éviter sa diffusion. C'est ce qui explique[524] qu'outre l'efficacité de la procédure accélérée au fond pour obtenir une décision ayant force de chose jugée, l'essentiel des procédures utiles relève, dans les catégories du droit français, du champ civil, et non du champ pénal[525].

Cette logique est également reprise, afin de protéger les droits de propriété intellectuelle, à l'article L. 336-2 du CPI, transposant l'article 8, paragraphe 3, de la directive 2001/29/CE du Parlement

---

[523] *Influence et réseaux sociaux. Face aux nouveaux défis, structurer la filière de la création, outiller l'Etat et mieux protéger*, rapport au Premier ministre, janvier 2026, recommandations n° 63.

[524] On rappellera que dans la décision n° 2024-866 DC du 17 mai 2024 précitée sur la loi SREN, le Conseil constitutionnel s'est prononcé sur l'ajout d'une infraction pénale et qu'il a retenu, entre autres, pour conclure à l'inconstitutionnalité qu' « *en incriminant le simple fait de diffuser en ligne tout contenu transmis au moyen d'un service de plateforme en ligne, d'un service de réseaux sociaux en ligne ou d'un service de plateformes de partage de vidéo, au sens des dispositions auxquelles elles renvoient, les dispositions contestées n'exigent pas que le comportement outrageant soit caractérisé par des faits matériels imputables à la personne dont la responsabilité peut être engagée. D'autre part, en prévoyant que le délit est constitué dès lors que le contenu diffusé soit porte atteinte à la dignité de la personne ou présente à son égard un caractère injurieux, dégradant ou humiliant, soit crée à son encontre une situation intimidante, hostile ou offensante, les dispositions contestées font dépendre la caractérisation de l'infraction de l'appréciation d'éléments subjectifs tenant à la perception de la victime. Elles font ainsi peser une incertitude sur la licéité des comportements réprimés* » (paragr. 78).

[525] Ce qui a pu être regretté par certains auteurs comme Romain Ollard, « Loi SREN : entre régulation (primaire) et répression (secondaire) », *Droit pénal*, n° 7, juillet-août 2024, étude 15.

européen et du Conseil du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information : le tribunal judiciaire peut ordonner, à la demande des titulaires de droit (d'auteur ou voisins) sur les œuvres et objets protégés, de leurs ayants-droit, des organismes de gestion collective, des organismes de défense professionnelle et du Centre national du cinéma et de l'image animée, toute mesure susceptible de prévenir ou faire cesser ces atteintes « *à l'encontre de toute personne susceptible de contribuer à y remédier* », aux frais de cette dernière[526]. La particularité de ce dispositif tient à la définition très large des demandeurs et des personnes potentiellement visées[527], afin de donner toute son efficacité au dispositif.

7.2.2. Des responsabilités asymétriques face à des hypertrucages illicites

La mise en rapport de l'ensemble de ces dispositions avec les enjeux soulevés pour les secteurs de la culture et la création par les hypertrucages conduit à relever une protection imparfaite selon le type d'enjeu soulevé. Ces enjeux sont évoqués successivement, sauf, ainsi que cela a déjà été mentionné, ceux liés à un hypertrucage qui constituerait un crime ou un délit couvert par une disposition spéciale (terrorisme, pédocriminalité et infractions énumérées au A du IV de l'article 6 de la LCEN).

a) En cas d'atteinte aux droits de propriété intellectuelle, **le code de la propriété intellectuelle permet d'ores et déjà d'agir à l'égard d'un contenu numérique, dont les hypertrucages**.

On rappellera ainsi que, face aux difficultés qui avaient été identifiées[528], les articles L. 137-1 à L. 137-4 comportent des dispositions spéciales pour certains fournisseurs de services de partage de contenus en ligne, afin de responsabiliser ceux-ci et, selon les termes de la CJUE qui a validé la légalité de l'article 17 de la directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique, dont ces dispositions assurent la transposition en droit français, « *d'encourager le développement du marché de l'octroi de licences équitables entre les titulaires de droits et ces fournisseurs*[529] ». Les droits de PLA étant ainsi opposables à ces plateformes, les fournisseurs sont conduits à devoir collaborer avec les ayants-droit, notamment pour empêcher la présence d'œuvres non autorisées (c'est-à-dire contrefaisantes) sur leurs plateformes : il s'agit ainsi d'agir *ex ante* à l'égard d'acteurs en mesure d'agir utilement (et le cas échéant, en cas de manquement, *ex post*) et non plus seulement, dans le cadre de la responsabilité de droit commun, *ex post*[530].

La **procédure de l'article L. 336-2 du CPI** permet ainsi à toutes les personnes mentionnées – et cette liste dépasse les seules victimes d'un dommage direct – de saisir le président du tribunal judiciaire selon la procédure accélérée au fond tant pour prendre des mesures à l'encontre de toute personne, ce qui couvre tant les fournisseurs de services visés par les articles L. 137-1 à L. 137-4, que les fournisseurs de services intermédiaires relevant du DSA (l'illicéité résultant de la méconnaissance des droits de propriété intellectuelle) et, pour ce qui concerne les hypertrucages, les fournisseurs de systèmes d'IA générative et les déployeurs ainsi que, le cas échéant, les personnes agissant à titre personnel. En outre, la seule condition de fond est l'atteinte à un droit d'auteur ou un droit voisin, ce qui est à la fois large et sans exigence quant au degré de l'atteinte, son caractère vraisemblable ou plausible étant suffisant[531].

---

[526] Voy. à cet égard l'arrêt de la CJUE, 27 mars 2014, *UPC Telekabel Wien GmbH*, C-314/12, ainsi que les décisions CE, 24 avril 2019, *Société Free* , n°425941, 428381, inédit (QPC) et CE, 13 novembre 2020, *Société Free* , n°425941, 428381, inédit.

[527] Christophe Caron, « Exégèse de l'article L. 336-2 du Code de la propriété intellectuelle concernant les injonctions en droit d'auteur », *in Mélanges André Lucas*, Paris, LexisNexis, 2014, p. 107 ; Caroline Le Goffic, V° « Injonctions contre les intermédiaires techniques n'ayant pas participé à l'atteinte au droit d'auteur ou au droit voisin – (CPI, art. L. 336-2) », *JurisClasseur Propriété littéraire et artistique*, fasc. 1616, janvier 2026.

[528] Valérie-Laure Benabou et Sophie Goossens, « Où en est le *value gap* ? Le transfert de valeur dans le projet de directive sur le droit d'auteur dans le marché unique numérique », *Propriétés Intellectuelles,* 2017, n° 65, p. 6-11 ; Alexandra Bensamoun, « *Value gap* : une adaptation du droit d'auteur au marché unique numérique », *Dalloz*, 2018, p. 122-123.

[529] CJUE, gr. ch., 26 avril 2022, *Pologne c. Parlement européen et Conseil*, C-401/19, point 29.

[530] Michel Vivant, « Une responsabilité ad hoc pour les sites de partage », *Communication Commerce électronique*, 2019, dossier 8, n° 12 ; Alexandra Bensamoun, « L'article 17 de la directive 2019/790 ou le retour de l'opposabilité des droits : une mesure d'équilibre, en équilibre », *RIDA*, 2020/2, n° 264, p. 72.

[531] Cass. 1re Civ., 12 juillet 2012, pourvoi n° 11-20.358, *Bull*. 2012, I, n° 168. Voy. également, et concernant les limites quant aux mesures pouvant être ordonnées : Cass. 1re Civ., 6 juillet 2017, pourvois nos 16-18.595, 16-17.217, 16-18.298, 16-18.348, *Bull*. 2017, I, n° 170.

Cet ensemble paraît suffisamment complet pour permettre d'agir de manière pertinente face à la diffusion d'un hypertrucage méconnaissant les droits d'auteur ou droits voisins. Néanmoins, il est circonscrit à la protection des droits de propriété intellectuelle, ce qui est insuffisant au regard des questions soulevées par les hypertrucages, ainsi que cela a déjà évoqué au chapitre 3.

b) En cas d'hypertrucage participant d'une manipulation, plusieurs outils[532] sont particulièrement orientés vers les enjeux de désinformation :

- L'article 58 de la loi n° 86-1067 du 30 septembre 1986 relative à la liberté de communication permet à l'Arcom d'adresser des recommandations aux fournisseurs de plateformes en ligne, aux moteurs de recherche en ligne et aux fournisseurs de services de plateformes de partage de vidéos (c'est-à-dire, contrairement aux éditeurs de presse, des services dont le fournisseur n'a pas de responsabilité éditoriale sur les contenus mais en détermine l'organisation selon l'article 2 de cette loi) afin d'améliorer la lutte contre la diffusion de fausses informations. Selon l'article 60, les fournisseurs de plateformes de partage de vidéos doivent mettre en place des procédures de résolution des réclamations, l'Arcom pouvant, selon l'article 17, être saisie de tout différend à cet égard entre un utilisateur et un fournisseur. Il s'agit donc d'une action préventive à l'égard d'acteurs disposés à agir.
- L'article 27 de la loi du 29 juillet 1881 sur la liberté de la presse prévoit, de manière ancienne, que « *La publication, la diffusion ou la reproduction, par quelque moyen que ce soit, de nouvelles fausses, de pièces fabriquées, falsifiées ou mensongèrement attribuées à des tiers lorsque, faite de mauvaise foi, elle aura troublé la paix publique, ou aura été susceptible de la troubler, sera punie d'une amende de 45 000 euros* ». Bien que n'étant pas limitée aux éditeurs de presse (qui sont aisément identifiables compte tenu des obligations qui leur sont applicables), cette disposition répressive ne permet toutefois que de sanctionner, sans intervenir de manière systémique sur le contenu concerné.
- En matière électorale, ainsi que cela a déjà été évoqué, l'article L. 163-2 du code électoral, issu de la loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, permet de saisir le juge des référés du tribunal judiciaire pour faire cesser la diffusion d'informations dont il est manifeste qu'elles sont inexactes ou trompeuses et de nature à altérer la sincérité du scrutin et qui sont diffusées de manière délibérée, artificielle ou automatisée et massive, par le biais d'un service de communication au public en ligne. Cette action, qui est indépendante de la réalisation du dommage, est limitée dans sa finalité et, en quelque sorte, à éclipse, puisqu'elle est pertinente à l'approche d'un scrutin.

C'est en réalité dans le cadre de la mise en œuvre du RIA qu'une protection plus large contre les manipulations, y compris celles portant sur l'information, pourra être mise en œuvre : le *a* de l'article 5, paragraphe 1, qualifiant de pratique interdite l'utilisation d'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. Sous réserve de la réunion de l'ensemble de ces conditions – qui sont nombreuses – le contenu ainsi qualifié est illicite et entre dans le cadre des procédures prévues par le DSA à l'égard de ce type de contenus. Les critères de qualification d'une telle pratique d'IA comme illégale au sens de l'article 5 du RIA recoupent en outre, en droit français, les critères de mise en œuvre de la procédure de l'article 6-3 de la LCEN.

En tenant compte des évolutions prévues dans le projet de loi portant diverses dispositions d'adaptation au droit de l'Union européenne en matière économique, financière, environnementale, énergétique, d'information, de transport, de santé, d'agriculture et de pêche, en cours d'examen au Parlement, il en résulte que :

- Les fournisseurs de services intermédiaires peuvent ainsi être saisis, au moins par l'Arcom et, pour ceux relevant du chapitre III du DSA, dans le cadre de réclamations d'utilisateurs ;

[532] Comme indiqué précédemment, nous n'abordons pas ici ce qui relève en particulier des pratiques commerciales déloyales, telles que définies par la directive 2005/29/CE du Parlement européen et du Conseil du 11 mai 2005 relative aux pratiques commerciales déloyales des entreprises vis-à-vis des consommateurs dans le marché intérieur et les articles L. 121-1 et suivants du code de la consommation.

- Toute personne pouvant se prévaloir d'un dommage ou d'un risque de dommage peut saisir l'autorité judiciaire afin de prendre des mesures appropriées auprès des fournisseurs de services intermédiaires, des fournisseurs de système d'IA, des déployeurs, voire potentiellement auprès de tout opérateur si les conditions de l'article 6-3 de la LCEN sont satisfaites (ce qui, eu égard aux critères de qualification de la pratique, devrait être le cas dès lors que la qualification peut manifestement être retenue). Les autorités administratives autorisées peuvent, le cas échéant, utiliser les pouvoirs de l'article 6-4 ;
- En application de l'article 24 *terdecies* du projet de loi, l'Arcom est en outre compétente pour connaître spécifiquement des pratiques interdites en matière d'IA relevant du *a* de l'article 5, paragraphe 1, du RIA. Elle pourra à ce titre prononcer les injonctions déjà évoquées, le cas échéant accompagnées d'astreintes, et des amendes administratives.

Cet ensemble devrait permettre ainsi aux personnes concernées de disposer d'outils pour réagir face à ce type d'hypertrucages.

c) S'agissant d'hypertrucages méconnaissant les règles de transparence de l'article 50, paragraphe 2 et 4, du RIA, la mise en œuvre des procédures résultant des DSA et prévues par le droit national en conséquence est **conditionnée à la qualification de leur caractère illicite**, dont il a été souligné qu'elle résultait de l'économie générale du texte, auquel elle donne son effet utile, même si cela mériterait d'être clarifié. Cette qualification est déterminante, sauf à ce que, la seule possibilité d'action puisse s'exercer auprès de l'Arcom, dans les conditions déjà évoquées et prévues par le projet de loi en cours d'examen au Parlement.

En tout état de cause, on peut estimer que **cette hypothèse recoupera fréquemment** (mais pas systématiquement, ce qui est une difficulté) **une méconnaissance des dispositions de l'article 226-8 du code pénal**, qui, elle, permet la reconnaissance de l'illicéité. En effet, cette disposition réprime :

- le fait de porter à la connaissance du public ou d'un tiers, par quelque voie que ce soit, le montage réalisé avec les paroles ou l'image d'une personne sans son consentement, s'il n'apparaît pas à l'évidence qu'il s'agit d'un montage ou s'il n'en est pas expressément fait mention ;
- le fait de porter à la connaissance du public ou d'un tiers, par quelque voie que ce soit, un contenu visuel ou sonore généré par un traitement algorithmique et représentant l'image ou les paroles d'une personne, sans son consentement, s'il n'apparaît pas à l'évidence qu'il s'agit d'un contenu généré algorithmiquement ou s'il n'en est pas expressément fait mention.

L'illicéité fondée sur l'article 226-8 du code pénal est donc restrictive par rapport aux hypertrucages : elle n'est actionnable que pour les personnes, à l'égard de deux attributs de leur personnalité (voie et image) et à condition qu'il y ait, outre une absence de consentement, une intention manipulatrice[533], tandis que le caractère évident du montage ou du caractère artificiel ne permet pas de sanctionner l'absence de marquage et d'étiquetage.

Du point de vue procédural, **l'appréhension d'un hypertrucage diffusé en méconnaissance des obligations de l'article 50 du RIA est donc largement dépendant de la qualification de l'illicéité en résultant, ce qui renforce la nécessité d'une clarification**. Au demeurant, alors que le projet de lignes directrices de la Commission allait en ce sens, cette lecture n'a *in fine* pas été retenue. Elle est d'une importance telle qu'elle justifie une intervention législative pour la soutenir.

d) Concernant les hypertrucages qui pourraient être critiqués sur le terrain de la méconnaissance des règle relatives à la protection des données personnelles, on relèvera que l'article 124-9 de la loi Informatique et libertés, que l'article 24 *duodecies* du projet de loi portant diverses dispositions d'adaptation au droit de l'Union européenne en matière économique, financière, environnementale, énergétique, d'information, de transport, de santé, d'agriculture et de pêche, en cours d'examen au Parlement prévoit d'introduire, explicite que les dispositions de mise en œuvre du RIA sont sans préjudice de l'ensemble des dispositions relatives à la protection des données.

Dès lors que celles-ci sont en jeu, la CNIL dispose ainsi de l'ensemble des moyens d'investigation et de sanction pertinents, mais ceux-ci portent seulement à l'égard des responsables de traitement, ce qui est

533 Cass. Crim., 30 mars 2016, pourvoi n° 15-82.039, Bull. crim. 2016, n° 112.

une limite intrinsèque, pour le sujet du présent rapport, aux pouvoirs d'action sur des contenus diffusés par voie numérique. Cela permet en outre de qualifier une illicéité rendant d'autant plus opérante la voie de la procédure accélérée au fond de l'article 6-3 de la LCEN qu'elle permet d'agir dans certains cas directement contre le responsable de traitement – notion large qui peut intégrer des opérateurs comme l'exploitant d'une place de marché en ligne[534] – ce qui permet de réduire l'exigence d'un caractère manifestement illicite, bien que cette procédure soit limitée à la personne victime de l'utilisation irrégulière de ses données.

d) Enfin, si la DSA entend agir face à des contenus illicites de manière générale, parmi les nombreuses illicéités qu'un hypertrucage est susceptible de constituer, deux hypothèses particulières sont à signaler car, en l'absence de législation spéciale les couvrant, elles manifestent les limites du dispositif actuel pour traiter des hypertrucages illicites.

α) Il s'agit d'abord du cas où l'hypertrucage constituerait une infraction aux dispositions de la loi du 29 juillet 1881 relative à la liberté de la presse. On peut ainsi envisager qu'un hypertrucage soit susceptible de constituer une infraction de diffamation.

Cette hypothèse souligne les limites des moyens permettant d'agir face à une illicéité, que ce soit par le biais d'une action auprès d'un opérateur ou devant le juge civil au titre de l'article 6-3 de la LCEN. En effet, pour que ces voies soient pertinentes, l'illicéité doit être suffisamment caractérisée, ce qui peut soulever des difficultés, en particulier pour l'infraction de diffamation, qui peut être écartée si la preuve de la vérité est rapportée ou l'excuse de bonne foi admise. Or, c'est précisément dans cette hypothèse que la Cour de cassation a pu souligner les limites du dispositif de l'article 6-3 et la nécessité de caractériser une illicéité manifeste. Elle a en effet retenu que, si, aux termes de cette disposition, le président du tribunal judiciaire, statuant selon la procédure accélérée au fond, peut prescrire à toute personne susceptible d'y contribuer toutes mesures propres à prévenir un dommage ou à faire cesser un dommage occasionné par le contenu d'un service de communication au public en ligne, pour faire usage, *en l'absence de débat contradictoire avec les auteurs des propos litigieux*, du pouvoir qui lui est ainsi conféré de retrait de contenu ou de blocages de sites, il doit constater le caractère manifestement illicite des propos critiqués constitutifs d'un abus de la liberté d'expression. Ce caractère manifestement illicite n'est pas établi par la seule communication de propos portant atteinte à l'honneur ou à la considération d'une personne, la diffamation alléguée pouvant être écartée si la preuve de la vérité est rapportée ou si l'excuse de bonne foi est admise[535].

En revanche, la procédure de l'article 6-3 permet d'ouvrir les moyens d'action au-delà du seul responsable de la publication du contenu litigieux, ce qui est un élément central de l'équilibre retenu par la Cour de cassation en 2025. Pour autant, l'invocation d'une atteinte à la loi de 1881 suppose que seule la victime puisse agir en demande.

β) Ensuite, en cas d'atteinte aux droits de la personnalité et au droit à la vie privée, violant les dispositions de l'article 9 du code civil, la qualification de l'illicéité peut également poser des difficultés au regard de l'appréciation exigée par les textes, alors que les procédures mises en place, si elles sont efficaces, dérogent, afin d'être expédientes, au droit commun et constituent des limites aux droits et libertés, de sorte que leur champ ne peut s'étendre sans restriction.

A titre liminaire, on relèvera que, s'agissant d'atteintes à la vie privée, le second alinéa de l'article 9 du code civil prévoit une procédure spécifique, permettant d'obtenir de toute personne les mesures utiles afin d'empêcher ou faire cesser l'atteinte : « *Les juges peuvent, sans préjudice de la réparation du dommage subi, prescrire toutes mesures, telles que séquestre, saisie et autres, propres à empêcher ou faire cesser une atteinte à l'intimité de la vie privée : ces mesures peuvent, s'il y a urgence, être ordonnées en référé* ». Sans revenir sur l'articulation entre cette procédure et le référé de l'article 835 du code de procédure civile, on ne peut manquer de s'interroger sur l'articulation entre cette disposition, qui ouvre la possibilité de recourir à une procédure au fond ou au référé, et l'article 6-3 de la LCEN, qui crée une procédure accélérée au fond, les deux dispositions permettant d'agir à l'encontre de personnes autres que l'auteur du contenu litigieux.

Cette question fait écho à celle de l'articulation entre l'article 9 du code civil et la loi du 29 juillet 1881, dont les conséquences procédurales sont plus lourdes[536], la jurisprudence tenant compte assez largement de l'objet du contentieux, selon qu'il se fonde sur l'un ou l'autre fondement : lorsqu'une partie

[534] CJUE, gr. ch., 2 décembre 2025, *Russmedia Digital et Inform Media Press*, C-492/23.
[535] Cass. 1re Civ., 26 février 2025, pourvoi n° 23-16.762, au Bull.
[536] Voy. par ex. Agathe Lepage, V° Droits de la personnalité, *Répertoire Dalloz de droit civil*, juillet 2022.

invoque une atteinte à son image et à sa réputation, ces faits, constitutifs de diffamation, ne pouvaient être poursuivis que sur le fondement de la loi du 29 juillet 1881[537]. Il en va différemment lorsque la victime fonde son action exclusivement sur l'article 9 du code civil et qu'aucun fait constitutif de diffamation n'est invoqué, l'action ne relevant alors pas de la loi du 29 juillet 1881[538]. Une assignation se fondant sur les deux dispositions étant nulle[539], l'absence de clarté de la séparation entre les deux dispositions a pu être critiquée[540]. En revanche, s'agissant d'une atteinte à la présomption d'innocence garantie par l'article 9-1 du code civil, la jurisprudence fait prévaloir cette dernière disposition sur la loi de 1881[541].

Cette approche, par sa complexité, aurait posé difficulté au regard d'hypertrucages qui peuvent cumuler les atteintes à plusieurs dispositions. Néanmoins, la Cour de cassation a consacré l'autonomie de la procédure instaurée par l'article 6-3 de la LCEN : celle-ci peut, sous la réserve déjà évoquée tenant à l'exigence de la démonstration d'une atteinte à un droit, être mobilisée pour obtenir le retrait de contenus diffamatoires[542]. Une telle logique d'affirmation de la primauté des procédures de la LCEN est de nature, dans un contexte où l'identification des auteurs est difficile et où la diffusion numérique des hypertrucages est très rapide, à faciliter une réponse adaptée et diligente. En outre, la mobilisation de l'article 6-3, de préférence au référé de l'article 9 du code civil, s'avère nécessaire pour mobiliser les dispositions de l'article 6-4 de cette loi, sauf à méconnaître l'objectif d'efficacité poursuivi par le législateur. Aussi, si la spécificité de la loi du 29 juillet 1881, tenant à l'identification d'un éditeur, a pu expliquer l'exclusivité de ses dispositions par rapport au code civil, la mission estime qu'en dehors du cas particulier où est engagée la responsabilité d'un éditeur de presse, il y a lieu de considérer que **la procédure de l'article 6-3 de la LCEN devrait permettre d'obtenir toute les mesures utiles face à la diffusion d'un hypertrucage illicite** dans le cadre de la procédure accélérée au fond prévue par la LCEN et au bénéfice des dispositions substantielles de l'article 9 du code civil.

Néanmoins, la juxtaposition de procédures propres au droit de la presse et à l'atteinte à la vie privée, non compatibles entre elles est une source de complexité pour les victimes. S'il est notable que la Cour de cassation a, en février 2025, admis que la procédure de la LCEN puisse être utilisée en cas de dommage résultant d'une diffamation, l'ensemble n'en demeure pas moins une source d'interrogations au moment de l'engagement d'une procédure. Par suite, **la mission considère que la prise en compte des hypertrucages devrait conduire à engager une réflexion pour tenir compte des situations où les atteintes aux droits des personnes sont plurifactorielles et conduisent à invoquer cumulativement plusieurs dispositions**.

Au fond, concernant la procédure de l'article 6-3 de la LCEN, il avait été souligné que « *la plupart des violations du droit au respect de la vie privée ne sont pas non plus manifestement illicites*[543] », faisant obstacle à une action efficace dans le cadre des anciennes dispositions prévoyant le référé. En effet, les mécanismes de la LCEN reposaient initialement sur le caractère *manifeste* de l'illicéité, qui seul justifie la capacité à contraindre un opérateur, qui n'est pas l'auteur, à agir sur ce contenu. C'est une réserve que le Conseil constitutionnel avait posée dès 2004[544] et que le législateur avait ensuite reprise à son compte[545]. Le juge veillait scrupuleusement à cette condition[546].

Aujourd'hui, l'article 6-3 de la LCEN ne prévoit plus expressément une telle contrainte ; les dispositions antérieures à la loi SREN sur la responsabilité des hébergeurs qui la restreignait aux cas d'illicéités manifestes ne sont plus en vigueur. Néanmoins, ainsi qu'on l'a indiqué au point *i*, compte tenu du caractère particulier de la procédure de l'article 6-3, qui ne met pas nécessairement en cause l'auteur, **le**

---

[537] Cass. 1re Civ., 8 novembre 2017, pourvoi n° 16-23.779, Bull. 2017, I, n° 230.
[538] Cass. 2e Civ., 24 avril 2003, pourvoi n° 01-01.186, Bull. 2003, II, n° 114.
[539] Cass. 1re Civ., 4 février 2015, pourvoi n° 13-16.263, Bull. 2015, I, n° 27.
[540] J. Hauser, « Atteinte à l'intimité de la vie privée, droit civil, droit pénal : prescriptions et preuve », *RTD Civ*., 2012, p. 89.
[541] Cass. 1re Civ., 20 mars 2007, pourvoi n° 05-21.541, Bull. 2007, I, n° 124 ; Cass. 1re Civ., 8 novembre 2017, pourvoi n° 16-23.779, Bull. 2017, I, n° 230.
[542] Cass. 1re Civ., 26 février 2025, pourvoi n° 23-16.762, au Bull. ; Cass. 1re Civ., 26 février 2025, pourvoi n° 23-22.386, au Bull. Rappr. Cass. 1re Civ., 16 février 2022, pourvoi n° 21-10.211, au Bull.
[543] L. Marino, V° Responsabilité civile et pénale des fournisseurs d'accès et d'hébergement, *JurisClasseur Communication*, fasc. 670, 2015.
[544] Déc. n° 2004-496 DC du 10 juin 2004, cons. 9.
[545] Art. 17 de la loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet.
[546] Voy. par ex. Cass. 1re Civ., 23 novembre 2022, pourvoi n° 21-10.220, au Bull., retenant le caractère manifestement illicite d'un site internet contrevenant aux dispositions françaises prohibant la gestation pour autrui.

dommage invocable doit être suffisamment manifeste pour bénéficier de cette voie expédiente – ce qui n'est pas requis s'agissant de la procédure de l'article L. 336-2 du CPI lorsque l'illicéité résulte d'une atteinte aux droits d'auteurs et voisins. En outre, cette exigence se rapproche de celle qui figure dans le DSA, qui responsabilise les fournisseurs de services d'hébergement : selon l'article 16 de ce règlement, le mécanisme de notification n'entraîne prise de connaissance aux fins de l'article 6 (et donc responsabilisation de ces fournisseurs) qu'à la condition que cette notification leur permette d'identifier l'illégalité de l'activité ou de l'information concernée « *sans examen juridique détaillé* ». Cette restriction est conforme aux principes posés antérieurement par la CJUE[547] et s'il a pu être relevé qu'elle ne porte que sur l'effet probatoire de la notification dans le cadre du DSA[548], il n'en reste pas moins que l'exigence peut avoir des incidences sur la charge de la preuve dans le cadre de la procédure de l'article 6-3 de la LCEN, même si cette procédure n'exige pas que la personne dont l'action est attendue soit responsable du contenu.

Or, **au regard des hypertrucages, l'accès à cette procédure peut se heurter à certaines difficultés pour faire reconnaître une méconnaissance du droit à la vie privée ou des droits de la personnalité**[549]. La possibilité de mettre en œuvre l'article 6-3 de la LCEN, au-delà même de la question préalable de cumul des dispositions, soulève, en particulier à la lumière de la jurisprudence relative à la procédure de l'article 9 du code civil, **des questions liées à la personne du défendeur, mais aussi à celle du demandeur et à la preuve du trouble manifestement illicite**.

En effet, s'agissant du défendeur, la circonstance que la procédure permette d'agir à l'encontre de toute personne pouvant faire cesser la diffusion a pour effet, potentiellement, de ne pas attraire l'auteur dans le débat, c'est-à-dire celui qui, en premier et en connaissance de cause, a manqué à ses obligations. Or, ainsi que nous l'avons vu sur la diffamation au point α, l'absence de débat contradictoire avec l'auteur, faisant obstacle à la possibilité d'invoquer l'exception de vérité ou l'excuse de bonne foi, la seule atteinte à l'honneur ou à la considération d'une personne ne suffit pas. De même, dès lors que, s'agissant d'une atteinte au consentement, il appartient usuellement, en matière d'édition, à celui qui publie et édite (donc au défendeur) d'établir la réalité de celui-ci, qui peut, au demeurant, parfois être présumé au regard des circonstances[550], un hypertrucage à l'encontre duquel serait invoqué un vice du consentement tiré des droits de la personnalité pourrait se heurter, sauf dans un contexte soumis au droit de la presse, à l'absence de mise en cause de l'auteur (pour autant qu'il soit identifiable) ; il pourrait également devenir difficile, à l'égard de plateformes de services intermédiaires, de discuter de la portée réelle d'un consentement donné antérieurement[551] faute pour elles de disposer des éléments permettant de le connaître, rendant délicate la satisfaction de la condition posée par le DSA d'absence d'examen juridique détaillé. Cette conséquence de la configuration procédurale tenant à la circonstance que le défendeur n'est pas l'auteur du contenu litigieux, si elle assure son efficacité, vise à protéger la liberté d'expression et souligne en creux l'équilibre particulier du droit de la presse tenant à l'identification d'une personne responsable[552]. Il s'en infère, quand bien même la disposition législative prévoit la seule démonstration d'un dommage (le cas échéant imminent), un report de la charge sur le demandeur qui peut prêter à discussion lorsqu'il n'y a pas – nous y reviendrons – une illicéité qui résulte d'une violation des dispositions pénales protégeant le droit à la vie privée. La mise en cause de la responsabilité des plateformes de services intermédiaires, en particulier, en

---

[547] CJUE, 22 juin 2021, *Youtube et Cyando*, aff. C-682/18 et C-683/18.

[548] Maxime Brenaut, Hélène Skrzypniak, V° Responsabilité civile, pénale et para-pénale des fournisseurs de services intermédiaires, *JurisClasseur Communication*, fasc. 670, juin 2025.

[549] La Cour de cassation admet en effet l'utilisation de cette disposition pour protéger les droits de la responsabilité (Cass. Civ., 13 janvier 1998, pourvoi n° 95-13.694, *Bull*. 1998, I, n° 14 concernant la reproduction de l'image d'une personne sous la forme de caricature ; Cass. 1re Civ., 16 juillet 1998, pourvoi n° 96-15.610, *Bull*. 1998, I, n° 259, concernant l'utilisation de l'image d'une personne dans un jeu vidéo dans un sens volontairement dévalorisant).

[550] Cass. 2e Civ., 4 novembre 2004, pourvoi n° 02-15.120, Bull., 2004, II, n° 487 ; 1re Civ., 13 novembre 2008, pourvoi n° 06-16.278, Bull. 2008, I, n° 259.

[551] Voy. Cass. 1re Civ., 14 juin 2007, pourvoi n° 06-13.601, Bull. 2007, I, n° 236 : la participation volontaire de personnes atteintes d'une maladie à une émission télévisée ayant pour objet de sensibiliser le public à celle-ci n'implique pas un consentement des intéressés à voir reproduite dans un livre leur photographie prise en la circonstance et sur laquelle ils demeurent identifiables, l'utilisation du cliché, fût-elle destinée à illustrer un étude d'intérêt général relative à l'affection en cause, étant faite dans un contexte différent.

[552] Comme l'indique l'avocat général dans ses conclusions sur l'arrêt Cass. 1re Civ., 26 février 2025, pourvoi n° 23-16.762, « *l'identification d'un responsable du contenu est l'une des contreparties de la liberté de publication tant pour s'assurer du statut de ce qui est publié, que pour permettre, d'identifier la personne susceptible de répondre de ce contenu et ainsi, le cas échéant, aux personnes visées par cette information, de la contredire* ».

devient plus délicate, d'autant plus que l'arrêt de la Cour de cassation du 26 février 2025 souligne la nécessité de tenter d'identifier l'auteur du contenu litigieux pour agir contre les autres opérateurs[553].

Se pose ensuite une question relative à la personne du demandeur. La jurisprudence retient que les atteintes à la loi du 29 juillet 1881 supposent une identification claire de la victime[554], de même que la reconnaissance d'une atteinte à la vie privée au titre de l'article 9 du code civil[555]. **Une atteinte au droit à l'image suppose également que la représentation permette l'identification de la personne concernée**[556]. Or, cela rend impossible l'action des tiers, y compris les ayants-droit, notamment après le décès de la personne représentée. C'est précisément ce type de restriction que l'article L. 336-2 du CPI a entendu dépasser.

Enfin, pour obtenir les mesures prévues par l'article 6-3 de la LCEN et l'article 9 du code civil sans que l'auteur soit appelé à l'instance (ce qui est le cas de procédure visant en particulier les fournisseurs de services intermédiaires), il est nécessaire de qualifier, au titre de la première disposition, un dommage actuel ou futur, qui, certes, ne saurait s'assimiler, en l'absence de débat contradictoire, à l'exigence de la démonstration complète d'un tel dommage mais, pour déterminer les mesures proportionnées dans le cadre d'une instance de fond (s'agissant de la procédure de l'article 6-3), requiert d'apprécier la cause de ce dommage – et donc l'illicéité manifeste, ce qui tend à la rapprocher de la condition posée pour la procédure prévue au second alinéa de l'article 9 du code civil.

Sur le fondement de l'article 9 du code civil, a été reconnue l'urgence à ordonner des mesures destinées à faire cesser l' « *atteinte intolérable à la personnalité* » résultant d'une photographie truquée représentant des employés des pompes funèbres avec des brassards « SS » en train de porter le cercueil du Président Georges Pompidou[557] et cette disposition permet au juge des référés d'agir face à toute atteinte à la vie privée et à l'image, ainsi qu'aux droits de la personnalité[558]. Il n'en reste pas moins que **le trouble doit être *manifestement* illicite** – condition qui entraîne automatiquement la reconnaissance de l'urgence à statuer pour le juge des référés[559] mais qui assure une certaine continuité avec l'ancien critère jurisprudentiel de l'atteinte *intolérable* à la vie privée[560]. L'enjeu est donc, tant pour l'article 6-3 de la LCEN que l'article 9 du code civil, de parvenir, dans les limites d'un débat contradictoire auquel ne participe pas nécessairement l'auteur du contenu litigieux, le caractère illicite de celui-ci. Deux hypothèses principales peuvent être à cet égard distinguées.

D'une part, cette condition est *a priori* aisément satisfaite lorsque le contenu litigieux constitue une infraction pénale, en particulier au regard des articles 226-1, 226-2 et 226-8 du code pénal, la Cour de cassation écartant alors toute mise en balance avec la liberté d'expression[561]. Pour autant, le débat s'avérera plus complexe s'agissant des hypertrucages relevant de l'infraction spécifique de l'article 226-8, qui comporte une condition tenant au caractère *évident* du montage, en l'absence de mention en ce sens. Les conditions du marquage et de l'étiquetage par le fournisseur du système d'IA et le déployeur, telles que prévues par le RIA, supposent qu'ils soient parties à la procédure, ce qui ne sera pas le cas si les mesures demandées conduisent à s'adresser aux fournisseurs de services intermédiaires.

En revanche, et d'autre part, en dehors de cette hypothèse, dès lors que le fondement de l'atteinte aux droits à la vie privée et de la personnalité d'une personne relève de la seule méconnaissance du code civil, la possibilité d'obtenir des mesures de la part des fournisseurs de services intermédiaires ou de

---

[553] Voy. à ce sujet le commentaire de Grégoire Loiseau, « Responsabilité de l'hébergeur - Injonction de retrait d'un contenu en ligne : l'illicéité doit être certaine », *Communication Commerce électronique*, 2025, n° 4, comm. 31.

[554] Par ex. Cass. Crim., 16 septembre 2003, pourvoi n° 02-85.113, Bull. crim. 2003, n° 161.

[555] Cass. 2e Civ., 8 juillet 2004, pourvoi n° 03-13.260, Bull., 2004, II, n° 390 ; Cass. 1re Civ., 15 février 2005, pourvoi n° 03-18.302, Bull. 2005, I, n° 86.

[556] Cass. 1re Civ., 5 avril 2012, pourvoi n° 11-15.328, Bull. 2012, I, n° 86 ; Cass. 1re Civ., 20 mars 2014, pourvoi n° 13-16.829, Bull. 2014, I, n° 57.

[557] TGI Paris, ord. réf., 20 juin 1974, *D.* 1974, jurisp. P. 751, note R. Lindon.

[558] Jean-Christophe Saint-Pau, V° Jouissance des droits civils. – Droit au respect de la vie privée. – Régime. Actions, *JurisClasseur Code civil*, art. 9, fasc. 20, février 2024.

[559] Cass. 1re Civ., 16 octobre 1984, pourvoi n° 83-11.312, Bulletin 1984 I N° 267 ; Cass. 1re Civ., 19 décembre 1995, pourvoi n° 93-18.939, Bulletin 1995 I N° 479

[560] Cass. 2e Civ., 12 juillet 1966, Bull. civ. II, n° 778.

[561] Voy. même si ces arrêts concernent des affaires dirigées contre l'éditeur : Cass. 1re Civ., 6 octobre 2011, pourvoi n° 10-21.823, Bull. 2011, I, n° 162 ; Cass. 1re Civ., 6 octobre 2011, pourvoi n° 10-21.822, Bull. 2011, I, n° 161 ; Cass. 1re Civ., 2 juillet 2014, pourvoi n° 13-21.929, Bull. 2014, I, n° 122 ; Cass. 1re Civ., 15 janvier 2015, pourvoi n° 14-12.200.

services d'IA générative dépend du motif de l'atteinte, qui la rend plus ou moins facile à retenir « *sans examen juridique détaillé* » : en référé, dans une telle hypothèse, le juge est tenu de procéder à une mise en balance des différents intérêts[562] et, eu égard à la circonstance que les intérêts invocables ne sont pas nécessairement mis dans le débat, le caractère manifeste devient délicat à caractériser.

A la lumière de l'ensemble de ces éléments, la comparaison des procédures envisageables face à un hypertrucage mettant en jeu les droits de propriété littéraire et artistique et les droits voisins et celles qui le sont face à un hypertrucage mettant en cause le droit à la vie privée ou les droits de la personnalité, voire des manipulations ou des atteintes au droit à la protection des données personnelles, **la mission constate que, dans cette seconde série d'hypothèses, les garanties offertes sont, *de facto* et *de jure*, bien moindres, ce qu'elle estime insatisfaisant au regard des enjeux soulevés, qui sont d'une intensité comparable sur l'ensemble de ces terrains et peuvent se cumuler.**

Tout en étant consciente que les questions posées par les droits de la personnalité doivent être abordées avec une grande prudence dans la mesure où elles dépassent largement le champ du présent rapport et où une extension non maîtrisée des dispositions de l'article L. 336-2 du CPI à ce champ pourrait avoir des effets excessifs, **la mission estime néanmoins nécessaire d'envisager la manière dont l'effectivité de la procédure de l'article 6-3 de la LCEN pourrait être améliorée lorsque les droits liés à la vie privée, à la personnalité ou aux données personnelles sont remis en cause par un contenu généré par un système d'IA générative tel qu'un hypertrucage, en particulier quant aux personnes ayant intérêt à agir.**

[562] Cass. 1re Civ., 14 février 2018, pourvoi n° 17-10.499, Bull. 2018, I, n° 31.

ANNEXES

I. Lettre de mission

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Conseil supérieur de la propriété littéraire et artistique

Paris, le 20 JUIN 2025

Madame Célia Zolynski
Madame Joelle FARCHY
Professeures des universités

OBJET : Mission relative aux enjeux pour les secteurs culturels et créatifs des hypertrucages générés ou manipulés par l'intelligence artificielle

Mesdames les Professeures,

Au sens du règlement de l'UE du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (IA), un « hypertrucage » (*deepfake*) est « *une image ou un contenu audio ou vidéo généré ou manipulé par l'IA, présentant une ressemblance avec des personnes, des objets, des lieux, des entités ou événements existants et pouvant être perçu à tort par une personne comme authentiques ou véridiques* » (article 3). Le règlement européen du 13 juin 2024 précise que les hypertrucages ne se limitent pas à la sphère informationnelle et qu'ils peuvent faire partie « *d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue* » (article 50-4).

Réservée auparavant aux professionnels, la réalisation d'hypertrucages devient aujourd'hui, grâce à l'intelligence artificielle, à portée de main du grand public. L'IA permet ainsi de manipuler des images ou des contenus audio ou vidéo en intervertissant et en remplaçant des visages (*face swapping*) ou en reconstituant la voix d'une personne (synchronisation labiale ou *lipsyncing*). Elle permet de créer de toutes pièces des images ou des vidéos réalistes et crédibles. Outre qu'ils sont plus faciles à réaliser, les hypertrucages générés par l'IA présentent une apparence d'authenticité de plus en plus grande. Alors que les contenus manipulés étaient plutôt faciles à détecter auparavant, y compris parfois à l'œil nu, une telle discrimination devient désormais de plus en plus difficile à effectuer.

Le Règlement européen du 13 juin 2024 envisage les hypertrucages principalement sous l'angle des manipulations auxquelles ils peuvent donner lieu, à travers des obligations de transparence et de marquage.

Il fixe tout d'abord un cadre contraignant relatif au marquage des contenus générés par l'intelligence artificielle. En son article 50-2, il dispose que les fournisseurs de systèmes d'IA « *veillent à ce que les sorties des systèmes d'IA soient marquées dans un format lisible par machine et identifiables comme ayant été générées ou manipulées par une IA* ». Les

fournisseurs seront donc tenus de concevoir des systèmes de telle façon que les images, textes, contenus audio et vidéo à caractère synthétique fassent l'objet d'un marquage dans un format lisible par machine et que leur nature de contenus générés ou manipulés artificiellement soit détectable. Il s'agit donc d'une préoccupation qui doit être pleinement intégrée au processus de conception des systèmes d'IA [1].

Par ailleurs, lorsqu'ils présentent une ressemblance sensible avec des personnes, des objets, des lieux, des entités ou des événements existants et peuvent être perçus à tort comme authentiques, les images et les contenus audio ou vidéo devront être clairement identifiés comme tels par les déployeurs de systèmes d'IA en vertu de l'article 50-4 du même Règlement. Des règles particulières sont applicables d'une part dans les cas où le contenu fait partie « *d'une œuvre ou d'un programme manifestement artistique, créatif, satirique, fictif ou analogue* », et, d'autre part, aux « *déployeurs d'un système d'IA qui génère ou manipule des textes publiés dans le but d'informer le public sur des questions d'intérêt public* ».

C'est dans ce contexte que je souhaite vous confier une mission sur les enjeux soulevés par les hypertrucages générés ou manipulés par l'IA pour les secteurs culturels et créatifs.

Cette mission aura tout d'abord pour objet d'identifier les hypertrucages parmi la catégorie plus large des extrants synthétiques (*output*) en recherchant notamment comment l'intention de tromper peut être mobilisée pour entrer dans leur définition.

Votre mission s'attachera ensuite à évaluer les différents enjeux économiques soulevés par les hypertrucages artistiques en se penchant sur la fabrique actuelle de ces hypertrucages (qui les produit et pourquoi), avec les opportunités et les risques qui en découlent. Vous étudierez les enjeux qu'ils soulèvent pour l'ensemble des secteurs culturels et créatifs, notamment les auteurs et les titulaires de droits voisins dont la voix ou l'image peut être exploitée et de manière générale l'ensemble des auxiliaires de la création concernés. Vous vous pencherez sur les enjeux des hypertrucages pour les plateformes qui accueillent ces contenus synthétiques et aux questions qu'ils posent pour la découvrabilité des contenus culturels. Enfin, vous vous intéresserez à leurs enjeux pour les usagers qui se trouvent exposés à des contenus truqués mais ultraréalistes.

Vous pourrez également travailler aux mesures qui pourraient permettre de répondre à ce phénomène. Vous vous attacherez à étudier les modalités de mise en œuvre des obligations de transparence et de marquage prévues par le Règlement européen, dans le respect de la liberté d'expression et de la liberté des arts et des sciences. Vous analyserez les différents corpus juridiques susceptibles d'être mobilisés pour préserver les droits des personnes représentées de façon inauthentique par le *deepfake* (droit d'auteur, droits voisins, droits de la personnalité, action en concurrence déloyale, etc.), ou des personnes, en particulier les doubleurs, dont les prestations orales sont prolongées ou dupliquées. Dans ce cadre, vous

[1] Depuis la loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique, l'article 226-8 du Code pénal sanctionne, au même titre qu'un montage d'image, le fait de porter à la connaissance du public ou d'un tiers « *un contenu visuel ou sonore généré par un traitement algorithmique et représentant l'image ou les paroles d'une personne, sans son consentement, s'il n'apparaît pas à l'évidence qu'il s'agit d'un contenu généré algorithmiquement ou s'il n'en est pas expressément fait mention* ».

analyserez les contrats qui encadrent aujourd'hui l'exploitation de la voix et/ou de l'image de acteurs culturels concernés et les garanties qu'ils prévoient (consentement, rémunération, etc.).

Pour mener cette mission, vous serez assistées d'un, voire de deux rapporteurs. Vous pourrez vous appuyer sur les services du ministère de la culture, et en particulier le secrétariat général (service des affaires juridiques et internationales). Vous procèderez aux auditions des membres du CSPLA ainsi que des entités et personnalités dont vous jugerez les contributions utiles.

Il serait souhaitable que le rapport final issu de vos travaux puisse être présenté lors de la séance plénière du CSPLA de l'été 2026.

Je vous remercie d'avoir accepté cette mission et vous prie de croire, Mesdames les Professeures, à l'expression de ma meilleure considération.

Jean-Philippe MOCHON
Président du CSPLA

II. Questionnaire de la mission

Les objectifs de la mission relative aux enjeux pour les secteurs culturels et créatifs des hypertrucages générés ou manipulés par l'intelligence artificielle sont de faire le point sur les obligations spécifiques, notamment de marquage, posées par le Règlement européen sur l'IA (article 50) ; de cartographier la fabrique des deepfakes : qui les produit et pourquoi ; d'évaluer les enjeux tant pour les artistes et les autres parties prenantes à la création que pour les plateformes confrontées à ce phénomène et pour le public ; de proposer les outils pour y répondre : obligations de consentement et de rémunération, droit d'auteur et droits voisins, droit de la personnalité, concurrence déloyale…
Afin de proposer des pistes de solutions juridiques et économiques, la réflexion pourra le cas échéant être menée en contemplation des besoins et attentes de chaque secteur ou activité.
La finalité du présent questionnaire préliminaire est de recueillir les différents points de vue des structures et entités concernées, afin d'amorcer la phase des auditions dans les meilleures conditions.

1) Présentez succinctement votre activité en liaison avec le sujet de la mission.

2) Dans le cadre de votre activité, que considérez-vous comme relevant de l'hypertrucage ?

3) Quels sont les cas d'usage d'hypertrucages qui vous permettent d'améliorer vos performances et/ou de résoudre des problèmes dans le cadre de votre activité ?

4) Quels sont les cas d'usage d'hypertrucages qui soulèvent des problématiques juridiques ou économiques dans le cadre de votre activité ?

5) En matière d'hypertrucages, avez-vous connaissance de collaborations ou d'oppositions entre fournisseurs et/ou déployeurs d'IA et entreprises du secteur culturel (donnez des exemples concrets) ?

6) Dans quelle mesure le droit positif (RIA, RGPD, DSA, droit d'auteur, droits voisins, droits de la personnalité, action en concurrence déloyale, ...) vous semble-t-il adapté ou inadapté à l'émergence des hypertrucages et aux problématiques que vous avez identifiées ? Quelles évolutions vous paraitraient souhaitables ?

7) Eléments supplémentaires que vous souhaitez porter à notre connaissance.

III. Liste des personnes ayant répondu au questionnaire

Vingt-deux réponses ont été adressées à la suite de la diffusion du questionnaire :

- *Organisations professionnelles de producteurs :*
  - Syndicat national de l'édition phonographique (SNEP) ;
  - Société civile des producteurs de phonogrammes en France (SPPF) ;

- *Organisations professionnelles d'artistes et auteurs* :
  - SNAM-CGT ;
  - Les Voix ;
  - ADAMI ;
  - Société civile des auteurs multimédias (LaScam) ;
  - Société de perception et de distribution des droits des artistes-interprètes (SPEDIDAM) ;
  - Ligue des auteurs professionnels ;
  - Syndicat national des auteurs et des compositeurs (SNAC) ;

- *Groupes de presse et représentants* :
  - SIPA Ouest-France ;
  - Alliance de la presse d'information générale (APIG) ;
  - Groupe Les Echos-Le Parisien ;
  - Radio France ;
  - Groupe TF1 ;
  - Groupe EBRA (PQR) ;

- *Sociétés développant des outils d'IA* :
  - Google ;
  - Mistral AI ;

- *Autres acteurs et institutions* :
  - Autorité de régulation de la communication audiovisuelle et numérique (Arcom) ;
  - Institut national de l'audiovisuel (INA) ;
  - Provenance4Trust (TrustMyContent, UncovAI, Sciences Po Paris, CEPIC, Reporter sans frontières) ;
  - Fondation Allfeat ;
  - Cédric Limousin ;
  - 1D345 (Ideas), entreprise du groupe Compilatio.

IV. Liste des personnes auditionnées

- Jean-Michel Bruguière, professeur à l'Université Grenoble-Alpes
- Grégoire Loiseau, professeur à l'Université Paris 1 Panthéon-Sorbonne
- TF1 : Stéphane Soppelsa, directrice juridique du pôle contenu et information du groupe TF1 et Timothée Widiez, responsable juridique au sein de ce pôle
- LaScam : Nicolas Mazars, directeur des affaires juridiques et institutionnelles, et Théo Florens, conseiller en charge des affaires institutionnelles et européennes
- ADAMI : Elisabeth Le Hot, directrice générale, Benjamin Sauzay, directeur exécutif, Florent Viel, directeur juridique
- Ministère de la justice, Direction des affaires civiles et du sceau : Clara Thouvenot, Emilie Brunet, Manon Fauvernier et Céline Delcoigne
- Collectif Les Voix : Patrick Kuban, fondateur du collectif, coprésident UVA, Jean Vandecasteele, administrateur, Mathilde Croze, avocate
- Centre national du cinéma et de l'image animée : Alexis Goin, directeur juridique et financier ; Pauline Augrain, directrice du numérique, Pauline Dalmasso, Service des industries techniques et de l'innovation
- Anya Schiffrin, professeur à l'Université de Columbia (New York) et Haaris Mateen, Université de Houston (Texas)
- SNEP : Alexandre Lasch, directeur général
- Radio France : Claire Desprez, directrice juridique adjointe, Nicolas Javre, responsable du pôle propriété intellectuelle, Romaine Pereira, juriste
- Syndicat national du jeu vidéo : Lévan Sardjevéladzé, président
- Grégoire Loiseau, professeur à l'Université Paris 1 Panthéon-Sorbonne
- Jonathan Elkaim, avocat, et Gauthier Doleac, avocat
- Commission européenne : Yordanka Ivanova, chef de secteur au bureau de l'IA
- Provenance4Trust : Mathieu Kervenec et Xavier Brunet, co-fondateurs de TrustMyContent, Sylvie Fodor, directrice générale du CEPIC
- Jane Ginsburg, professeur à Columbia Université (New York)
- France Télévisions : Sylvie Courbarien Le Gall, directrice des affaires juridiques et européennes, membre du CSPLA, et Bertrand Scirpo, directeur délégué à la réglementation numérique et délégué à la protection des données
- Antoine Guiziou, doctorant à l'Université Paris I Panthéon-Sorbonne
- Jennifer Rothman, professeur à l'Université de Pennsylvanie
- Google : Arnaud Vergnes
- SACD : Hubert Tilliet, directeur des Affaires juridiques et des contrats audiovisuels
- SNAM CGT : Laurent Tardif, juriste
- Spedidam : Benoît Galopin, directeur des affaires juridiques et internationales
- INA : Jean-François Debarnot, directeur juridique, Barbara Mutz, responsable du département des affaires juridiques et réglementaires, Sandrine Lardeux, responsable du département des droits, acquisitions et royalties, Johanna Dong, responsable de mission au département des droits, acquisitions et royalites, Aude Formage, juriste au département des affaires juridiques et réglementaires
- Motion Picture Association (MPA) : Emilie Anthony, Johanna Baysse
- Commission européenne (bureau de l'IA et unité droit d'auteur)

BIBLIOGRAPHIE

0. 'AI-Generated Porn Site Mr. Deepfakes Shuts down after Service Provider Pulls Support - CBS News'. 5 May 2025. https://www.cbsnews.com/news/ai-generated-porn-site-mr-deepfakes-shuts-down/.

'2023 State Of Deepfakes: Realities, Threats, And Impact | Security Hero'. Accessed 4 June 2026. https://www.securityhero.io/state-of-deepfakes/.

'Addressing Deepfake Image-Based Abuse | eSafety Commissioner'. 24 July 2024. https://www.esafety.gov.au/newsroom/blogs/addressing-deepfake-image-based-abuse.

Adobe. *State of Creativity Report 2024*. Adobe, 2024. https://www.adobe.com/content/dam/cct/creativecloud/business/teams/whitepapers/dotcom-116728/Adobe_StateofCreativity_Report_2024_FV_UE.pdf.

Adobe Communications Team. 'Adobe's AI and the Creative Frontier Study Reveals Creators' Views on the Opportunities and Risks of Generative AI'. Adobe Blog, 8 October 2024. https://blog.adobe.com/en/publish/2024/10/08/adobes-ai-creative-frontier-study-reveals-creators-views-opportunities-risks-generative-ai.

'Advertisers Can Buy Product Placement Ads through The Trade Desk'. 17 June 2025. https://www.thetradedesk.com/press-room/ttd-unveils-ai-tool-rembrand.

Afchar, Darius, Gabriel Meseguer-Brocal, and Romain Hennequin. 'Detecting Music Deepfakes Is Easy but Actually Hard'. arXiv:2405.04181. Preprint, arXiv, 22 May 2024. https://doi.org/10.48550/arXiv.2405.04181.

Ahmed, Saifuddin, Adeline Wei Ting Bee, Sheryl Wei Ting Ng, and Muhammad Masood. 'Social Media News Use Amplifies the Illusory Truth Effects of Viral Deepfakes: A Cross-National Study of Eight Countries'. *Journal of Broadcasting & Electronic Media* 68, no. 5 (2024): 778–805. https://doi.org/10.1080/08838151.2024.2410783.

'AI Dubbing Tools Market'. *Market.Us*, n.d. Accessed 5 June 2026. https://market.us/report/ai-dubbing-tools-market/.

Algan, Yann, and Pierre Cahuc. 'Inherited Trust and Growth'. *American Economic Review* 100, no. 5 (2010): 2060–92. https://doi.org/10.1257/aer.100.5.2060.

Amerini, Irene, Mauro Barni, Sebastiano Battiato, et al. 'Deepfake Media Forensics: Status and Future Challenges'. *Journal of Imaging* 11, no. 3 (2025): 73.

'An Overview of the Launch of JournalismAI's Generating Change Report'. Accessed 4 June 2026. https://www.ftstrategies.com/en-gb/insights/journalismai-launch-event.

Astrid, Marcella, Enjie Ghorbel, and Djamila Aouada. 'Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies'. arXiv:2408.06753. Preprint, arXiv, 14 October 2024. https://doi.org/10.48550/arXiv.2408.06753.

Bagratuni, Mikael, Patrick Müller, and Patrick Planing. 'Innovation in Tune: An Empirical Investigation of User Acceptance of Artificial Intelligence-Generated Music'. *Computers in Human Behavior Reports* 18 (May 2025): 100660. https://doi.org/10.1016/j.chbr.2025.100660.

Barrington, Sarah, Emily A. Cooper, and Hany Farid. 'People Are Poorly Equipped to Detect AI-Powered Voice Clones'. *Scientific Reports* 15, no. 1 (2025): 11004. https://doi.org/10.1038/s41598-025-94170-3.

Cano, Pedro, Eloi Batlle, Ton Kalker, and Jaap Haitsma. 'A Review of Audio Fingerprinting'. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* 41, no. 3 (2005): 271–84. https://doi.org/10.1007/s11265-005-4151-3.

Cao, Lele. 'Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities'. arXiv:2504.03765. Preprint, arXiv, 2 April 2025. https://doi.org/10.48550/arXiv.2504.03765.

Cardon, Dominique. 'Les Vidéos Générées Par IA Ne Font Pas Disparaître l'intérêt Que Nous Portons Au Réel'. *Le Monde*, 11 February 2026. https://www.lemonde.fr/idees/article/2026/02/11/dominique-cardon-sociologue-les-videos-generees-par-ia-ne-font-pas-disparaitre-l-interet-que-nous-portons-au-reel_6666388_3232.html.

Casu, Mirko, Luca Guarnera, Pasquale Caponnetto, and Sebastiano Battiato. 'GenAI Mirage: The Impostor Bias and the Deepfake Detection Challenge in the Era of Artificial Illusions'. *Forensic Science International: Digital Investigation* 50 (September 2024): 301795. https://doi.org/10.1016/j.fsidi.2024.301795.

Centre national de la musique. *Manipulation Des Écoutes En Ligne*. Centre national de la musique, 2023. https://cnm.fr/etude-manipulation-des-ecoutes-en-ligne/.

Centre national de la musique. *Stream Manipulation: 2021*. Centre national de la musique, 2023. https://cnm.fr/wp-content/uploads/2023/01/2023_-CNM-_Manipulation-des-streams_ENG.pdf.

Chad. 'The Nelson Mandela Foundation Authorizes a Definitive Documentary Series on Nelson Mandela's Life. Nelson Mandela's Own Story in His Own Words, Narrated in His Voice.' *Joburgstyle Online*, 20 May 2024. http://www.joburgstyle.co.za/the-nelson-mandela-foundation-authorizes-a-definitive-documentary-series-on-nelson-mandelas-life-nelson-mandelas-own-story-in-his-own-words-narrated-in-his-voice/.

Chen, Guangyu, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 'WavMark: Watermarking for Audio Generation'. arXiv:2308.12770. Preprint, arXiv, 7 January 2024. https://doi.org/10.48550/arXiv.2308.12770.

Chesney, Robert, and Danielle Citron. 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security'. *California Law Review* 107, no. 6 (2019): 1753.

Chia, Suqi, Andree Hartanto, and Eddie M. W. Tong. 'Do Listeners Devalue AI-Generated Pop Music? Exploring Negative Biases in Listeners' Responses to AI-Labelled vs Human-Labelled Pop Music'. *Computers in Human Behavior: Artificial Humans* 6 (December 2025): 100217. https://doi.org/10.1016/j.chbah.2025.100217.

Ching, Didier, John Twomey, Matthew P. Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 'Can Deepfakes Manipulate Us? Assessing the Evidence via a Critical Scoping Review'. *PLOS ONE* 20, no. 5 (2025): e0320124. https://doi.org/10.1371/journal.pone.0320124.

Chmielewski, Dawn, Anna Tong, and Anna Tong. 'Scarlett Johansson Says OpenAI Chatbot Voice "eerily Similar" to Hers'. Technology. *Reuters*, 21 May 2024. https://www.reuters.com/technology/scarlett-johansson-says-openai-chatbot-voice-eerily-similar-hers-2024-05-21/.

CISAC. 'Global Economic Study Shows Human Creators' Future at Risk from Generative AI'. CISAC Newsroom, 4 December 2024. https://www.cisac.org/Newsroom/news-releases/global-economic-study-shows-human-creators-future-risk-generative-ai.

Communications, Reuters. 'Reuters Launches AI-Voiced Packaged Video Content in Spanish and Portuguese'. Media Center. *Reuters*, 15 July 2025. https://www.reuters.com/media-center/reuters-launches-ai-voiced-packaged-video-content-in-spanish-and-portuguese/.

Cooke, Chris. '71% of Music Creators Fear Multi-Billion Dollar Music AI Business Could Stop Songwriters from Earning a Living, Says New Report'. Complete Music Update, 31 January 2024. https://completemusicupdate.com/71-percent-of-music-creators-fear-multi-billion-dollar-music-ai-business-could-stop-songwriters-from-earning-a-living-says-new-report/.

Corsi, Giulio, Bill Marino, and Willow Wong. 'The Spread of Synthetic Media on X'. *Harvard Kennedy School Misinformation Review* 5, no. 3 (2024): 1–19.

Corvi, Riccardo, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 'On The Detection of Synthetic Images Generated by Diffusion Models'. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095167.

Cozzolino, Davide, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 'Id-Reveal: Identity-Aware Deepfake Video Detection'. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 15108–17. http://openaccess.thecvf.com/content/ICCV2021/html/Cozzolino_ID-Reveal_Identity-Aware_DeepFake_Video_Detection_ICCV_2021_paper.html.

Crépel, Maxime, Béatrice Mazoyer, Benjamin Ooghe-Tabanou, et al. *Une Cartographie Web de l'écosystème Médiatique Français*. 2025. https://sciencespo.hal.science/hal-05094289/.

D'Alessandro, Dominic Patten, Anthony. 'The Strike Is Over! SAG-AFTRA & Studios Reach Tentative Deal On New Three-Year Contract'. *Deadline*, 9 November 2023. https://deadline.com/2023/11/sag-strike-ends-actors-studios-deal-contract-1235566470/.

DataInsightsMarket. *AI Dubbing Tools Market's Evolution: Key Growth Drivers 2026–2034*. DataInsightsMarket, 2025. https://www.datainsightsmarket.com/reports/ai-dubbing-tools-1372047.

Dathathri, Sumanth, Abigail See, Sumedh Ghaisas, et al. 'Scalable Watermarking for Identifying Large Language Model Outputs'. *Nature* 634, no. 8035 (2024): 818–23.

'David France | Identity Protection with Deepfakes'. *MIT Open Documentary Lab*, n.d. Accessed 5 June 2026. https://opendoclab.mit.edu/presents/david-france-identity-protection-deepfakes/.

Deep Market Insights. *Virtual Concert Platform Market Research Report*. Deep Market Insights, 2025. https://deepmarketinsights.com/report/virtual-concert-platform-market-research-report.

Deezer. 'Deezer and Ipsos Study: AI Fools 97% of Listeners'. *Deezer Newsroom*, 12 November 2025. https://newsroom-deezer.com/2025/11/deezer-ipsos-survey-ai-music/.

Deloitte Insights. 'Earning Trust as Gen AI Takes Hold: 2024 Connected Consumer Survey'. Accessed 2 April 2026. https://www.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey/2024.html.

Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 'Arcface: Additive Angular Margin Loss for Deep Face Recognition'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4690–99. http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Doloriel, Chandler Timm C., Habib Ullah, Kristian Hovde Liland, Fadi Al Machot, and Ngai-Man Cheung. 'Towards Sustainable Universal Deepfake Detection with Frequency-Domain Masking'. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17 March 2026, 3797266. https://doi.org/10.1145/3797266.

Dutt, Deborshi, Beena Ammanath, Costi Perricos, and Brenna Sniderman. *The State of Generative AI in the Enterprise: Now Decides Next*. Deloitte AI Institute, 2024.

https://www.deloitte.com/gr/en/services/consulting/research/state-of-generative-ai-in-enterprise.html.

El Ali, Abdallah, Karthikeya Puttur Venkatraj, Sophie Morosoli, Laurens Naudts, Natali Helberger, and Pablo Cesar. 'Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How'. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2 May 2024, 1–11. https://doi.org/10.1145/3613905.3650750.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 'GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models'. arXiv:2303.10130. Preprint, arXiv, 21 August 2023. http://arxiv.org/abs/2303.10130.

Europol. 'Facing Reality? Law Enforcement and the Challenge of Deepfakes'. Accessed 1 June 2026. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

Europol. *Steal, Deal and Repeat: Internet Organised Crime Threat Assessment (IOCTA) 2025*. Europol, 2025. https://www.europol.europa.eu/cms/sites/default/files/documents/Steal-deal-repeat-IOCTA_2025.pdf.

Europol Innovation Lab. *Facing Reality? Law Enforcement and the Challenge of Deepfakes*. Europol, 2022. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

Fabien, Maël, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 'BertAA: BERT Fine-Tuning for Authorship Attribution'. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, 127–37. https://aclanthology.org/2020.icon-main.16/.

Fagot, Vincent. 'Drake and The Weeknd's fake AI song raises questions about respect for intellectual property'. Economy,Music. *Le Monde*, 26 April 2023. https://www.lemonde.fr/en/economy/article/2023/04/19/drake-and-the-weeknd-s-fake-ai-song-raises-questions-about-respect-for-intellectual-property_6023514_19.html.

Fairoze, Jaiden, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. 'Publicly-Detectable Watermarking for Language Models'. arXiv:2310.18491. Preprint, arXiv, 4 January 2025. https://doi.org/10.48550/arXiv.2310.18491.

Farchy, Joëlle, and Bastien Blain. *Rémunération Des Contenus Culturels Utilisés Par Les Systèmes d'IA*. 2025. https://strossburi.eu/wp-content/uploads/2025/05/CSPLA_Rapport-economique_Remuneration-des-contenus_Version-provisoire_Mai-2025.pdf.

Fernandez, Pierre, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models'. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 22466–77. http://openaccess.thecvf.com/content/ICCV2023/html/Fernandez_The_Stable_Signature_Rooting_Watermarks_in_Latent_Diffusion_Models_ICCV_2023_paper.html.

Fraissard, Guillaume. 'Reprise de « Papaoutai » Avec l'IA : « Savoir Qu'une Œuvre Est Fausse Change La Nature de l'émotion Qu'elle Nous Procure »'. *Le Monde*, 30 January 2026. https://www.lemonde.fr/idees/article/2026/01/30/reprise-de-papaoutai-avec-l-ia-savoir-qu-une-uvre-est-fausse-change-la-nature-de-l-emotion-qu-elle-nous-procure_6664754_3232.html.

Fritz, Mario. *Technical Solutions for Marking and Detecting AI-Generated Image Andf Video Content in the Context of Article 50(2) AI Act*. Publications Office of the European Union, 2026. https://data.europa.eu/doi/10.2759/9763153.

Future Market Insights. 'AI in Media and Entertainment Market | Global Market Analysis Report - 2036'. 2025. https://www.futuremarketinsights.com/reports/ai-in-media-and-entertainment-market.

GEMA. ‘Study: AI and Music’. GEMA News, 30 January 2024. https://www.gema.de/en/news/ai-study.

‘Generative AI in Movies and TV: How the 2023 SAG-AFTRA and WGA Contracts Address Generative AI | Perkins Coie’. Accessed 5 June 2026. https://perkinscoie.com/insights/blog/generative-ai-movies-and-tv-how-2023-sag-aftra-and-wga-contracts-address-generative?utm_source=chatgpt.com.

Geng, Mingmeng, and Thierry Poibeau. ‘On the Detectability of LLM-Generated Text: What Exactly Is LLM-Generated Text?’ arXiv:2510.20810. Preprint, arXiv, 23 October 2025. https://doi.org/10.48550/arXiv.2510.20810.

Goldman Sachs. ‘Generative AI Could Raise Global GDP by 7%’. 5 April 2023. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent.

Goldmedia. *AI and Music: A Study on the Impact of Artificial Intelligence on the Music Sector*. Goldmedia, 2024. https://www.goldmedia.com/fileadmin/goldmedia/Studie/2023/GEMA-SACEM_AI-and-Music/AI_and_Music_GEMA_SACEM_Goldmedia.pdf.

Google. ‘Règlement Sur Les Utilisations Interdites de l’IA Générative’. 2024. https://policies.google.com/terms/generative-ai/use-policy.

Gramigna, Remo. ‘Preserving Anonymity: Deep-Fake as an Identity-Protection Device and as a Digital Camouflage’. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 37, no. 3 (2024): 729–51. https://doi.org/10.1007/s11196-023-10079-y.

Grand View Research. *AI in Filmmaking Market Size, Share & Trends Analysis Report, 2033*. Grand View Research, 2025. https://www.grandviewresearch.com/industry-analysis/ai-filmmaking-market-report.

Grand View Research. *Europe Generative AI in Content Creation Market Outlook*. Grand View Research, 2025. https://www.grandviewresearch.com/horizon/outlook/generative-ai-in-content-creation-market/europe.

Grand View Research. *Generative AI in Music Market (2024-2030): Size, Share & Trends Analysis Report*. GVR-4-68040-418-2. Grand View Research, 2024. https://www.grandviewresearch.com/industry-analysis/generative-ai-in-music-market-report.

Guo, Biyang, Xin Zhang, Ziyuan Wang, et al. ‘How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection’. arXiv:2301.07597. Preprint, arXiv, 18 January 2023. https://doi.org/10.48550/arXiv.2301.07597.

Guo, Yinlin, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. ‘Audio Deepfake Detection with Self-Supervised Wavlm and Multi-Fusion Attentive Classifier’. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, 12702–6. https://ieeexplore.ieee.org/abstract/document/10447923/.

Hans, Abhimanyu, Avi Schwarzschild, Valeriia Cherepanova, et al. ‘Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text’. arXiv:2401.12070. Preprint, arXiv, 13 October 2024. https://doi.org/10.48550/arXiv.2401.12070.

Hashmi, Ammarah, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. ‘Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes’. arXiv:2405.04097. Preprint, arXiv, 11 November 2024. https://doi.org/10.48550/arXiv.2405.04097.

Hernandez-Ortega, Javier, Ruben Tolosana, Julian Fierrez, and Aythami Morales. ‘DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-Frame’. In *Handbook of Digital Face Manipulation and Detection*, edited by Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. Advances in Computer Vision and Pattern Recognition. Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-87664-7_12.

HeyGen. 'HeyGen AI Video Translator : Translate Videos into 175+ Languages'. Accessed 5 June 2026. https://www.heygen.com/translate.

Hinkle, Mallory. 'UNIVERSAL MUSIC GROUP AND UDIO ANNOUNCE UDIO'S FIRST STRATEGIC AGREEMENTS FOR NEW LICENSED AI MUSIC CREATION PLATFORM'. *UMG*, 30 October 2025. https://www.universalmusic.com/universal-music-group-and-udio-announce-udios-first-strategic-agreements-for-new-licensed-ai-music-creation-platform/.

Hoppner, Thomas, and Luke Streatfeild. 'ChatGPT, Bard & Co.: An Introduction to AI for Competition and Regulatory Lawyers'. *An Introduction to AI for Competition and Regulatory Lawyers (February 23, 2023)* 9 (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4371681.

'Hôtel du Temps | FranceTvPro.fr'. Accessed 5 June 2026. https://www.francetvpro.fr/contenu-de-presse/16034879.

'How Respeecher Restored Michael York's Voice for an Inspirational Healthcare Initiative'. Accessed 5 June 2026. https://www.respeecher.com/case-studies/respeecher-gives-voice-michael-york-healthcare-initiative.

Huang, Baixiang, Canyu Chen, and Kai Shu. 'Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges'. *ACM SIGKDD Explorations Newsletter* 26, no. 2 (2025): 21–43. https://doi.org/10.1145/3715073.3715076.

Huang, Baojin, Zhongyuan Wang, Jifan Yang, et al. 'Implicit Identity Driven Deepfake Face Swapping Detection'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 4490–99. http://openaccess.thecvf.com/content/CVPR2023/html/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.html.

Huertas-Tato, Javier, Alejandro Martin, and David Camacho. 'PART: Pre-Trained Authorship Representation Transformer'. arXiv:2209.15373. Preprint, arXiv, 9 May 2025. https://doi.org/10.48550/arXiv.2209.15373.

Huijstee, Mariëtte van, Pieter van Boheemen, Djurre Das, et al. *Tackling Deepfakes in European Policy*. EPRS_STU(2021)690039. STOA: Panel for the Future of Science and Technology. European Parliamentary Research Service, 2021. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039.

Hynek, Nik, Beata Gavurova, and Matus Kubak. 'Risks and Benefits of Artificial Intelligence Deepfakes: Systematic Review and Comparison of Public Attitudes in Seven European Countries'. *Journal of Innovation & Knowledge* 10, no. 5 (2025): 100782. https://doi.org/10.1016/j.jik.2025.100782.

Identity Music. 'Artificial Streaming: Fake Streams, Fraud & Fines'. 10 February 2026. https://identitymusic.com/blog/artificial-streaming-fake-streams-fraud-fines.

IMARC Group. *Visual Effects (VFX) Market Size, Share, Trends and Forecast by Component, Product, Technology, Application, and Region, 2025-2033*. IMARC Group, 2025. https://www.imarcgroup.com/visual-effects-market.

'Information Manipulation in the Age of Generative Artificial Intelligence | Think Tank | European Parliament'. Accessed 5 June 2026. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)779259.

'IPTC Photo Metadata Standard 2025.1'. 2025. https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata-2025.1.html#contributor.

Jia, Ye, Yu Zhang, Ron Weiss, et al. 'Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis'. *Advances in Neural Information Processing Systems* 31 (2018). https://proceedings.neurips.cc/paper/2018/hash/6832a7b24bc06775d02b7406880b93fc-Abstract.html.

Joint Research Centre. *Generative AI Outlook Report: Exploring the Intersection of Technology, Society and Policy*. JRC142598. Publications Office of the European Union, 2025. https://publications.jrc.ec.europa.eu/repository/handle/JRC142598.

Josan, Harnoorvir Singh. *AI and Deepfake Voice Cloning: Innovation, Copyright and Artists' Rights*. Digital Policy Hub Working Paper. Centre for International Governance Innovation, 2024. https://www.cigionline.org/publications/ai-and-deepfake-voice-cloning-innovation-copyright-and-artists-rights/.

'Journalism, Media, and Technology Trends and Predictions 2025 | Reuters Institute for the Study of Journalism'. 9 January 2025. https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2025.

Jovanović, Nikola, Robin Staab, and Martin Vechev. 'Watermark Stealing in Large Language Models'. arXiv:2402.19361. Preprint, arXiv, 24 June 2024. https://doi.org/10.48550/arXiv.2402.19361.

Kang, Seok, and Kayla Valadez. 'Deepfakes' Cognitive, Emotional, and Behavioral Impact: A Systematic Review and Meta-Analysis of Individual Responses'. *Journalism & Mass Communication Quarterly* 102, no. 4 (2025): 958–92. https://doi.org/10.1177/10776990251357294.

Kanter, Jake. 'Inside YouTube's Weird World Of Fake Movie Trailers — And How Studios Are Secretly Cashing In On The AI-Fueled Videos'. *Deadline*, 28 March 2025. https://deadline.com/2025/03/inside-youtube-fake-movie-trailers-1236352406/.

Kim, Zae Myung, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 'Threads of Subtlety: Detecting Machine-Generated Texts through Discourse Motifs'. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, 5449–74. https://aclanthology.org/2024.acl-long.298/.

Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 'A Watermark for Large Language Models'. *International Conference on Machine Learning*, 2023, 17061–84. https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Koltai, Kolina. 'Behind a Secretive Global Network of Non-Consensual Deepfake Pornography'. *Bellingcat*, 23 February 2024. https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/.

Krishna, Kalpesh, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 'Paraphrasing Evades Detectors of Ai-Generated Text, but Retrieval Is an Effective Defense'. *Advances in Neural Information Processing Systems* 36 (2023): 27469–500.

Kulyasov, Alexey. 'AI Dubbing 2025: How Technology Is Transforming the Video Localization Market'. Speeek.Io Blog, 31 July 2025. https://speeek.io/en/blog/ai-dubbing-2025.

Lavan, Nadine, Mairi Irvine, Victor Rosi, and Carolyn McGettigan. 'Voice Clones Sound Realistic but Not (yet) Hyperrealistic'. *PLOS One* 20, no. 9 (2025): e0332692. https://doi.org/10.1371/journal.pone.0332692.

Le Monde. 'Deepfakes pornographiques : l'un des principaux sites ferme, un pharmacien canadien identifié'. Pixels,Deepfakes. 7 May 2025. https://www.lemonde.fr/pixels/article/2025/05/07/deepfakes-pornographiques-l-un-des-principaux-sites-ferme-un-pharmacien-canadien-identifie_6603814_4408996.html.

Lees, Dominic. ‘Deepfakes in Documentary Film Production: Images of Deception in the Representation of the Real’. *Studies in Documentary Film* 18, no. 2 (2024): 108–29. https://doi.org/10.1080/17503280.2023.2284680.

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. ‘GPT Detectors Are Biased against Non-Native English Writers’. *Patterns* 4, no. 7 (2023). https://www.cell.com/patterns/fulltext/S2666-3899(23)00130-7?ref=maginative.com.

Lin, Kaiqing, Zhiyuan Yan, Ke-Yue Zhang, et al. ‘Guard Me If You Know Me: Protecting Specific Face-Identity from Deepfakes’. arXiv:2505.19582. Preprint, arXiv, 3 December 2025. https://doi.org/10.48550/arXiv.2505.19582.

Liu, Aiwei, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. ‘A Semantic Invariant Robust Watermark for Large Language Models’. arXiv:2310.06356. Preprint, arXiv, 19 May 2024. https://doi.org/10.48550/arXiv.2310.06356.

Liu, Chang, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. ‘Dear: A Deep-Learning-Based Audio Re-Recording Resilient Watermarking’. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 11 (2023): 13201–9. https://ojs.aaai.org/index.php/AAAI/article/view/26550.

Liu, Jiaxin, Jia Wang, Saihui Hou, et al. ‘Beyond Face Swapping: A Diffusion-Based Digital Human Benchmark for Multimodal Deepfake Detection’. arXiv:2505.16512. Preprint, arXiv, 28 January 2026. https://doi.org/10.48550/arXiv.2505.16512.

Liv, Nadine, and Dov Greenbaum. ‘Deep Fakes and Memory Malleability: False Memories in the Service of Fake News’. *AJOB Neuroscience* 11, no. 2 (2020): 96–104. https://doi.org/10.1080/21507740.2020.1740351.

Lu, Hang, and Shupei Yuan. ‘“I Know It’sa Deepfake”: The Role of AI Disclaimers and Comprehension in the Processing of Deepfake Parodies’. *Journal of Communication* 74, no. 5 (2024): 359–73.

Lundberg, Ebba, and Peter Mozelius. ‘The Potential Effects of Deepfakes on News Media and Entertainment’. *AI & SOCIETY* 40, no. 4 (2025): 2159–70. https://doi.org/10.1007/s00146-024-02072-1.

Majestya, Nayla, and Irwanto Irwanto. ‘Beyond Behavioral Change: Evaluating the Impact of Environmental Films on Audience Perceptions and Beliefs about Food System Issues’. *Frontiers in Communication* 10 (June 2025). https://doi.org/10.3389/fcomm.2025.1519348.

Market Research Future. *Generative AI in Music Market Growth Report 2035*. Market Research Future, 2025. https://www.marketresearchfuture.com/reports/generative-ai-in-music-market-31943.

Market.us. *AI in Film Market Size, Share, Trends*. Market.us, 2025. https://market.us/report/ai-in-film-market/.

Microsoft Customer Stories. ‘60 Days to Launch: Coca-Cola Reaches Millions with Immersive Campaign Built on Azure’. Microsoft Customer Stories, 15 April 2025. https://www.microsoft.com/en/customers/story/22668-coca-cola-company-azure-ai-and-machine-learning.

Milmo, Dan, and Dan Milmo Global technology editor. ‘AI Avatar Generator Synthesia Does Video Footage Deal with Shutterstock’. Technology. *The Guardian*, 10 April 2025. https://www.theguardian.com/technology/2025/apr/10/ai-avatar-generator-synthesia-video-footage-shutterstock.

Minsker, Evan. ‘Holly Herndon’s AI Deepfake “Twin” Holly+ Transforms Any Song Into a Holly Herndon Song’. Pitchfork, 14 July 2021. https://pitchfork.com/news/holly-herndons-ai-deepfake-twin-holly-transforms-any-song-into-a-holly-herndon-song/.

Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 'Detectgpt: Zero-Shot Machine-Generated Text Detection Using Probability Curvature'. *International Conference on Machine Learning*, 2023, 24950–62. https://proceedings.mlr.press/v202/mitchell23a.html.

Monroe, Jazz. 'Grimes Unveils Software to Mimic Her Voice, Offering 50-50 Royalties for Commercial Use'. Pitchfork, 2 May 2023. https://pitchfork.com/news/grimes-unveils-software-to-mimic-her-voice-and-announces-2-new-songs/.

Mordor Intelligence. *Deepfake AI Market Size & Share Analysis: Growth Trends & Forecasts*. Mordor Intelligence, 2025. https://www.mordorintelligence.com/industry-reports/deepfake-ai-market.

Morgan Stanley. 'GenAI's Leading Role in Entertainment'. Accessed 4 June 2026. https://www.morganstanley.com/insights/articles/ai-in-media-entertainment-benefits-and-risks.

Müller, Nicolas M., Piotr Kawa, Wei Herng Choong, et al. 'Mlaad: The Multi-Language Audio Anti-Spoofing Dataset'. *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, 1–7. https://ieeexplore.ieee.org/abstract/document/10650962/.

Murphy, Gillian, Didier Ching, John Twomey, and Conor Linehan. 'Face/Off: Changing the Face of Movies with Deepfakes'. *PLOS ONE* 18, no. 7 (2023): e0287503. https://doi.org/10.1371/journal.pone.0287503.

Murphy, Gillian, and Emma Flynn. 'Deepfake False Memories'. *Memory* 30, no. 4 (2022): 480–92. https://doi.org/10.1080/09658211.2021.1919715.

Naffi, Nadia. 'Deepfakes and the Crisis of Knowing'. UNESCO, 1 October 2025. https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing.

National Institute of Standards and Technology. *Reducing Risks Posed by Synthetic Content*. 2024. https://dpo-india.com/Resources/NIST/Reducing-Risks-Synthetic-Content-Overview-Technical-Digital-Content-Transparency.pdf.

Öhman, Carl. 'Who Are the Targets of Deepfakes? Evidence From Flagged Videos on TikTok and YouTube'. *Journal of Digital Social Research* 7, no. 2 (2025): 18–33. https://doi.org/10.33621/jdsr.v7i255529.

Oliullah, Md, and H. M. Murtuza. 'Exploring How Political Deepfakes Trigger Cognitive Processes and Downstream Psychological and Behavioral Outcomes: A Systematic Review'. *Computers in Human Behavior Reports* 22 (May 2026): 101066. https://doi.org/10.1016/j.chbr.2026.101066.

OpenAI. 'How the Voices for ChatGPT Were Chosen'. Accessed 5 June 2026. https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/.

OpenAI. 'New AI Classifier for Indicating AI-Written Text'. 13 March 2024. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/.

Pelly, Liz. 'As AI-Made Music Explodes, Deezer Lays out Strategy to Identify AI Tracks and Weed out Illegal and Fraudulent Content on Its Platform'. Music Business Worldwide, 6 June 2023. https://www.musicbusinessworldwide.com/as-ai-made-music-explodes-deezer-lays-out-strategy-to-identify-ai-tracks-and-weed-out-illegal-and-fraudulent-content-on-its-platform/.

People.Com. 'James Bond Trailer Featuring Henry Cavill, Margot Robbie Receives 2 Million Views on YouTube — Unfortunately, It's Fake'. Accessed 4 June 2026. https://people.com/fake-james-bond-trailer-henry-cavill-margot-robbie-receives-2-million-views-youtube-8634904.

Phukan, Orchid Chetia, Girish, Mohd Mujtaba Akhtar, Arun Balaji Buduru, and Rajesh Sharma. 'Towards Neural Audio Codec Source Parsing'. arXiv:2506.12627. Preprint, arXiv, 14 June 2025. https://doi.org/10.48550/arXiv.2506.12627.

Pindrop. *2024 Voice Intelligence & Security Report*. Pindrop, 2024. https://www.pindrop.com/press-release/pindrop-unveils-voice-intelligence-security-report/.

Powers, Greig, Jennifer P. Johnson, and Ginger Killian. 'To Tell or Not to Tell: The Effects of Disclosing Deepfake Video on US and Indian Consumers' Purchase Intention'. *Journal of Interactive Advertising* 23, no. 4 (2023): 339–55. https://doi.org/10.1080/15252019.2023.2260399.

Prada, Luis. '"Musician" Arrested for Using AI Songs and a Bot Army to Scam Spotify for Millions in Royalties'. VICE, 6 September 2024. https://www.vice.com/en/article/ai-songs-spotify-scam/.

Puccetti, Giovanni. *Technical Solutions for Marking and Detecting AI Generated Text Content in the Context of Article 50(2) AI Act*. Publications Office of the European Union, 2026. https://data.europa.eu/doi/10.2759/7579127.

Pulver, Andrew. 'The Brutalist and Emilia Perez's Voice-Cloning Controversies Make AI the New Awards Season Battleground'. Film. *The Guardian*, 20 January 2025. https://www.theguardian.com/film/2025/jan/20/the-brutalist-and-emilia-perezs-voice-cloning-controversies-make-ai-the-new-awards-season-battleground.

Rare Connections. 'AI Music Statistics and Facts: How Producers Use AI'. 2025. https://www.rareconnections.io/ai-music-statistics.

Regehr, Kaitlyn, Caitlin Shaughnessy, Minzhu Zhao, Idil Cambazoglu, Alfie Turner, and Nicola Shaughnessy. 'Normalizing Toxicity: The Role of Recommender Algorithms for Young People's Mental Health and Social Wellbeing'. *Frontiers in Psychology* 16 (November 2025). https://doi.org/10.3389/fpsyg.2025.1523649.

Rigole, Neil J., Zoroayka V. Sandoval, Shannon Beasley, and Kembley Lingelbach. 'Perceptions of AI-Generated and Co-Created Music among University Students and Social Media Users.' *Issues in Information Systems* 26, no. 4 (2025): 102–17.

Rijsbosch, Bram, Gijs van Dijck, and Konrad Kollnig. 'Adoption of Watermarking for Generative AI Systems in Practice and Implications under the New EU AI Act'. arXiv:2503.18156. Preprint, arXiv, 8 October 2025. https://doi.org/10.48550/arXiv.2503.18156.

Roman, Robin San, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. 'Proactive Detection of Voice Cloning with Localized Watermarking'. arXiv:2401.17264. Preprint, arXiv, 6 June 2024. https://doi.org/10.48550/arXiv.2401.17264.

Rosi, Victor, Emma Soopramanien, and Carolyn McGettigan. 'Perception and Social Evaluation of Cloned and Recorded Voices: Effects of Familiarity and Self-Relevance'. *Computers in Human Behavior: Artificial Humans* 4 (May 2025): 100143. https://doi.org/10.1016/j.chbah.2025.100143.

Safi, Michael, Alex Atack, Joshua Kelly, Philip McMahon, and Luke Hoyland. 'Revealed: The Names Linked to ClothOff, the Deepfake Pornography App'. Technology. *The Guardian*, 29 February 2024. https://www.theguardian.com/technology/2024/feb/29/clothoff-deepfake-ai-pornography-app-names-linked-revealed.

SAG-AFTRA. 'Tentative Agreement Reached'. 8 November 2023. https://www.sagaftra.org/tentative-agreement-reached.

San Roman, Robin, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. 'Latent Watermarking of Audio Generative Models'. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, 1–5. https://ieeexplore.ieee.org/abstract/document/10889782/.

Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natália S. Bueno. 'The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?' *American Political Science Review* 119, no. 1 (2025): 71–90. https://doi.org/10.1017/S0003055423001454.

Scribner, Herb. 'Fake Movie Trailers Were an Art Form. Then Came the AI Slop.' *The Washington Post*, 28 April 2025. https://www.washingtonpost.com/entertainment/movies/2025/04/28/fake-movie-trailers-ai-youtube/.

Segundo, Eugenia San, Aurora López-Jareño, Xin Wang, and Junichi Yamagishi. 'Human Perception of Audio Deepfakes: The Role of Language and Speaking Style'. arXiv.Org, 10 December 2025. https://arxiv.org/abs/2512.09221v1.

Serra, Xavier, Araz, Oguz, Batlle Roca, Roser, Juvela, Lauri, López, David, and Rocamora, Martín. *Technical Solutions for Marking and Detecting AI-Generated Audio Content in the Context of Article 50(2) AI Act*. Publications Office of the European Union, 2026. https://data.europa.eu/doi/10.2759/0784462.

Singla, Alex, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. *The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value*. McKinsey & Company, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024.

Soto-Sanfiel, María T., Ariadna Angulo-Brunet, and Sanjay Saha. 'Deepfakes as Narratives: Psychological Processes Explaining Their Reception'. *Computers in Human Behavior* 165 (2025): 108518.

Soto-Sanfiel, María T., and Gina Junhan Fu. 'Deepfaking the Past: Memory and Perceived Truth of Resurrected Historical Figures'. *Computers in Human Behavior*, 2026, 109008.

SPECTRA. *Analysis of the Stakeholder Consultation on Transparency Requirements for Certain AI Systems under Article 50 AI Act*. Publications Office of the European Union, 2026. https://data.europa.eu/doi/10.2759/0753439.

Stamatatos, Efstathios. 'A Survey of Modern Authorship Attribution Methods'. *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538–56. https://doi.org/10.1002/asi.21001.

Sumsub. *Identity Fraud Report 2025–2026*. Sumsub, 2025. https://sumsub.com/fraud-report-2025/.

Tancik, Matthew, Ben Mildenhall, and Ren Ng. 'Stegastamp: Invisible Hyperlinks in Physical Photographs'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 2117–26. http://openaccess.thecvf.com/content_CVPR_2020/html/Tancik_StegaStamp_Invisible_Hyperlinks_in_Physical_Photographs_CVPR_2020_paper.html.

Terrone, Enrico. 'Documentaries, Docudramas, and Perceptual Beliefs'. *The Journal of Aesthetics and Art Criticism* 78, no. 1 (2020): 43–56. https://doi.org/10.1111/jaac.12703.

The Business Research Company. *Artificial Intelligence (AI) Dubbing Tools Global Market Report 2025*. The Business Research Company, 2025. https://www.researchandmarkets.com/reports/6075298/artificial-intelligence-ai-dubbing-tools.

The YouTube team. 'How We're Helping Creators Disclose Altered or Synthetic Content'. Blog.Youtube, 2024. https://blog.youtube/news-and-events/disclosing-ai-generated-content/.

Twomey, John, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 'Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets That Discuss Deepfakes in the Russian Invasion of Ukraine'. *PLOS ONE* 18, no. 10 (2023): e0291668. https://doi.org/10.1371/journal.pone.0291668.

U.S. Attorney's Office, Southern District of New York. 'North Carolina Musician Charged With Music Streaming Fraud Aided By Artificial Intelligence'. 4 September 2024. https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence.

U.S. Government Accountability Office. *Science & Tech Spotlight: Combating Deepfakes*. GAO-24-107292. U.S. Government Accountability Office, 2024. https://www.gao.gov/products/gao-24-107292.

Vaccari, Cristian, and Andrew Chadwick. 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News'. *Social Media + Society* 6, no. 1 (2020): 2056305120903408. https://doi.org/10.1177/2056305120903408.

Visionary Data Reports. *Virtual Concert Platform Market*. Visionary Data Reports, 2025.

Wallace, Allison. 'Spotify and Universal Music Group Announce Landmark Licensing Agreements for Fan-Made Covers and Remixes'. *Spotify*, 21 May 2026. https://newsroom.spotify.com/2026-05-21/universal-music-group-spotify-licensing-agreements-fan-made-covers-remixes/.

Wang, Mei, and Weihong Deng. 'Deep Face Recognition: A Survey'. *Neurocomputing* 429 (March 2021): 215–44. https://doi.org/10.1016/j.neucom.2020.10.081.

Wang, Xin, Héctor Delgado, Hemlata Tak, et al. 'Asvspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech'. *Computer Speech & Language* 95 (2026): 101825.

Wang, Xin, and Junichi Yamagishi. 'Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders'. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5. https://ieeexplore.ieee.org/abstract/document/10094779/.

'Want Your Own "Celebrity Elevator Selfie"? Here's How TikTokers Are Doing It'. 17 October 2025. https://dailydot.com/celebrity-elevator-tiktok-trend/.

Weikmann, Teresa, Hannah Greber, and Alina Nikolaou. 'After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media'. *The International Journal of Press/Politics* 30, no. 1 (2025): 187–210. https://doi.org/10.1177/19401612241233539.

Wen, Yuxin, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 'Tree-Ring Watermarks: Fingerprints for Diffusion Images That Are Invisible and Robust'. arXiv:2305.20030. Preprint, arXiv, 4 July 2023. https://doi.org/10.48550/arXiv.2305.20030.

World Intellectual Property Organization. 'IFPI Looks at a Decade of Digital Transformation in the Music Industry'. *WIPO Magazine*, 23 April 2025. https://www.wipo.int/en/web/wipo-magazine/articles/ifpi-looks-at-a-decade-of-digital-transformation-in-the-music-industry-73661.

Yeo, Sara K., Andrew R. Binder, Michael F. Dahlstrom, and Dominique Brossard. 'An Inconvenient Source? Attributes of Science Documentaries and Their Effects on Information-Related Behavioral Intentions'. *Journal of Science Communication* 17, no. 2 (2018): A07. https://doi.org/10.22323/2.17020207.

Yi, Jiangyan, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 'Audio Deepfake Detection: A Survey'. *arXiv Preprint arXiv:2308.14970*, 2023. https://arxiv.org/abs/2308.14970.

You, Bangjian. 'Explore the Impact of the Application of Artificial Intelligence Facial Image Processing and Synthesis Technology in Documentaries on Social Acceptance'. *Proceedings of the 2nd International Conference on Applied Psychology and Marketing Management*, 2025. https://www.scitepress.org/Papers/2025/141122/141122.pdf.

Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 'Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data'. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14448–57. http://openaccess.thecvf.com/content/ICCV2021/html/Yu_Artificial_Fingerprinting_for_Generative_Models_Rooting_Deepfake_Attribution_in_Training_ICCV_2021_paper.html.

Zhang, Kaichen, Zixuan Yuan, and Hui Xiong. 'The Impact of Generative Artificial Intelligence on Market Equilibrium: Evidence from a Natural Experiment'. arXiv:2311.07071. Preprint, arXiv, 10 October 2024. https://doi.org/10.48550/arXiv.2311.07071.

Zhang, Xinyi, Chenshuo Sun, Renyu Zhang, and Khim-Yong Goh. 'The Value of AI-Generated Metadata for UGC Platforms: Evidence from a Large-Scale Field Experiment'. arXiv.Org, 24 December 2024. https://arxiv.org/abs/2412.18337v1.

Zhao, Xuandong, Sam Gunn, Miranda Christ, et al. 'Sok: Watermarking for Ai-Generated Content'. *2025 IEEE Symposium on Security and Privacy (SP)*, 2025, 2621–39. https://ieeexplore.ieee.org/abstract/document/11023391/.

Zhao, Yuan, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, and Xin Yu. 'Proactive Deepfake Defence via Identity Watermarking'. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, 4602–11. https://openaccess.thecvf.com/content/WACV2023/html/Zhao_Proactive_Deepfake_Defence_via_Identity_Watermarking_WACV_2023_paper.html.