Roudy KARAM
Jane AZIZ
Sarra MEJRI
Kevin ANDRIANASOLO LALA

Synthèse:

IA Génératives et Deepfakes

2025-2026

1) Résumé du sujet

L'**IA générative** s'est imposée comme un outil puissant, bien loin du simple gadget : en quelques secondes, elle permet désormais de produire des textes, images, voix ou vidéos d'un réalisme saisissant. Cette révolution soulève une question centrale : celle de la **confiance**. Peut-on encore croire ce que l'on voit et entend en ligne ? Les **deepfakes** (contenus fabriqués ou retouchés par IA) sont au cœur de cette question. Les premières générations utilisaient des GANs, aujourd'hui les modèles de diffusion ont encore augmenté la qualité, baissé les coûts et facilité la production.

En quelques années, on est passé de l'émergence, (premiers exemples qui circulent) à la démocratisation (outils accessibles à tous), puis à la maturité (créations rapides et virales). Les risques sont réels : confusion entre le plausible et le vrai, remise en cause de la parole publique, fraudes et usurpations d'identité, perte de repères dans l'information scientifique. Nos biais jouent contre nous : on croit plus facilement ce qui confirme nos idées, et on se pense souvent "moins influençable" que les autres. D'où notre question : comment limiter les dérives sans bloquer l'innovation ?

2) Résumé des approches présentées

Nous avons montré que la réponse ne peut pas être seulement technique. Il faut aussi des règles, de la gouvernance et de l'éducation. Côté règles, **l'AI Act** en Europe impose notamment d'indiquer clairement quand un contenu est généré ou modifié par IA, de réduire la production de contenus illégaux et d'être plus transparent sur les données protégées utilisées pour entraîner les modèles. Il prévoit aussi des "bacs à sable" pour tester dans un cadre sécurisé. **L'UNESCO** met l'accent sur la gouvernance (outils d'auto-évaluation), la transparence des plateformes, des audits indépendants, et surtout l'éducation aux médias pour tous. **Le NIST** (aux États-Unis) propose une méthode simple à retenir pour les organisations : gouverner (rôles et règles), cartographier les risques (usages et abus possibles), mesurer (fiabilité, biais, remontées d'incidents) puis gérer dans le temps (corriger, informer, améliorer). **Le Partnership on AI** pousse des bonnes pratiques communes : expliquer clairement quand un contenu vient d'une IA, fixer des règles d'usage, et mettre en place des canaux de signalement et de recours.

Côté techniques, l'objectif est de savoir d'où vient un contenu et s'il a été généré par IA. Le **C2PA** attache à chaque fichier un "passeport" de provenance signé (qui a créé quoi, avec quel outil, quelles modifications). L'intérêt du C2PA, c'est qu'il assure l'intégrité et la traçabilité des médias, mais ne dit pas si le contenu est vrai ou faux : il dit seulement d'où ils viennent et s'ils ont été altérés. Une autre solution technique est le **filigranage invisible** (watermarking) qui insère un signal dans le contenu lui-même. Pour le texte, des approches de soft watermarking (comme **SynthID-Text**) introduisent des signatures statistiques détectables sans altérer la lisibilité. Pour l'image, les méthodes de type **Tree-Ring**, incorporent le filigrane dès le bruit initial des modèles de diffusion, améliorant la robustesse sans sacrifier la qualité visuelle. La combinaison C2PA + filigranage crée une base solide pour l'attribution, la responsabilité et la dissuasion.

La détection des deepfakes repose donc sur deux approches principales : active, qui authentifie la source et vérifie l'intégrité du média via le tatouage numérique, et passive, qui

utilise les réseaux de neurones profonds pour distinguer le vrai du faux. Les méthodes passives analysent les images et vidéos pour repérer des anomalies visuelles ou temporelles, comme un clignement d'œil irrégulier, des ombres incohérentes ou des problèmes de synchronisation labiale. L'évaluation ne doit pas se limiter à l'accuracy ou à l'AUC, mais inclure la robustesse aux perturbations réelles (bruit, flou, compression), l'empreinte de calcul, la consommation mémoire et la latence d'inférence. Aucun modèle ne domine sur tous les critères, surtout les performances chutent significativement sur des données réelles ("in the wild") par rapport aux jeux de données de référence, ce qui montre que la généralisation et la robustesse restent des défis majeurs. De là découlent nos recommandations : garantir la transparence, assurer une chaîne de provenance complète, privilégier la détection hybride, mettre en place une gouvernance continue et, transversalement, promouvoir l'éducation aux biais et aux réflexes de vérification pour tous, avec des mécanismes clairs de signalement, de retrait et de réparation.

3) Résumé du débat

Nous avons commencé par un mini-sondage et un test visuel. Résultat : beaucoup se sont trompés entre une image réelle et une image IA. Cela montre qu'"à l'œil nu", ce n'est plus si simple, même en regardant de près les détails de la photo, d'identifier les images générées par IA.

On a commencé par la question : les filigranes suffisent-ils contre les fake news ? On a trouvé comme conclusion que les filigranes ne suffisent pas : ils sont faciles à contourner et pas toujours présents. Pour la non-altération, on peut ajouter un tatouage invisible "ne pas altérer" : en *mode actif*, les outils/plateformes lisent ce marqueur et peuvent refuser d'éditer ou publier, alerter ou déréférencer les copies où le tatouage est cassé. Ça n'arrête pas complètement les fake news puisque le filigrane et même le tatouage peuvent être supprimés par une simple capture d'écran, mais ça dissuade les actions malveillantes.

On a enchaîné avec l'idée de rendre les filigranes visibles au public. Ça présente un intérêt évident et facilite la détection rapide par le grand public sans recourir à des outils spécialisés. Cependant, maintenir des filigranes invisibles évite de stigmatiser les créateurs, protège parfois la vie privée et complique leur retrait pour un acteur malveillant.

La discussion a alors glissé vers la responsabilité. Un consensus s'est formé autour d'une responsabilité partagée : l'auteur malveillant pour l'intention et la diffusion, les plateformes pour la modération et le retrait rapide. Nous avons également abordé l'idée suivante : malgré des garde-fous contre les contenus haineux, des contournements subsistent côté texte via la reformulation ou l'insistance, tandis que les pipelines image et vidéo sont mieux verrouillés mais pas infaillibles éventuellement dans le futur.

En conclusion, les IA génératives peuvent être bénéfiques mais posent un vrai défi pour notre société : celui de la confiance. Il existe aujourd'hui plusieurs solutions techniques et légales, comme le C2PA, le filigranage ou encore l'AI Act, mais elles ne sont pas forcément suffisantes. Alors la question reste ouverte : faut-il limiter l'accès à ces outils, mieux les encadrer, ou tout simplement miser sur l'éducation et l'esprit critique ? C'est sans doute dans l'équilibre entre ces trois approches que se joue notre capacité à vivre avec l'IA sans perdre confiance dans le réel.

Sources

Risques

Livre blanc IA générative et hypertrucages, avril 2024 (CNPEN – CNIL, France).
https://www.minalogic.com/livre-blanc-ia-generative-hypertrucages-deepfake/

Twomey J., Ching D., Aylett M. P., Quayle M., Linehan C., Murphy G. (2023) Do deepfake videos undermine our epistemic trust?
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291668

Mesures d'encadrement

Union européenne. Regulation (EU) 2024/1689 … (Artificial Intelligence Act). EUR-Lex / Official Journal of the European Union, 2024
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

UNESCO. Recommendation on the Ethics of Artificial Intelligence. UNESCO, 2021
https://unesdoc.unesco.org/ark:/48223/pf0000381137

National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST, 2023
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Adeptiv AI. 5 Powerful Ways the NIST Generative AI Framework Protects You. Adeptiv, 2024
https://adeptiv.ai/nist-generative-ai-framework/

Partnership on AI. Responsible Practices for Synthetic Media: A Framework for Collective Action. Partnership on AI, 2023
https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf

Solutions techniques (C2PA+filigranage)

Coalition for Content Provenance and Authenticity (C2PA). Content Credentials: C2PA Technical Specification, v2.2. C2PA, 2025.
https://spec.c2pa.org/specifications/specifications/2.2/specs/_attachments/C2PA_Specification.pdf

Coalition for Content Provenance and Authenticity (C2PA). C2PA and Content Credentials Explainer, v2.2. C2PA, 2025.
https://spec.c2pa.org/specifications/specifications/2.2/explainer/_attachments/Explainer.pdf

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. A Watermark for Large Language Models. arXiv / ICML 2023, 2023. https://arxiv.org/abs/2301.10226

Dathathri, S., See, A., Ghaisas, S., et al. Scalable watermarking for identifying large language model outputs. Nature, 2024. https://www.nature.com/articles/s41586-024-08025-4

Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. arXiv, 2023. https://arxiv.org/abs/2305.20030

Huang, H., Wu, Y., & Wang, Q. ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization. arXiv (accepted NeurIPS 2024), 2024. https://arxiv.org/abs/2411.03862

Google DeepMind. SynthID. Google DeepMind, s. d. https://deepmind.google/science/synthid/

Alan Turing Institute, Behind the Deepfake: 8% Create; 90% Concerned (2024, OA) https://www.turing.ac.uk/sites/default/files/2024-07/behind_the_deepfake_full_publication.pdf

Pawelec, Deepfakes and Democracy (Theory) (PMC 2022, OA) https://pmc.ncbi.nlm.nih.gov/articles/PMC9453721/

Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024 https://arxiv.org/html/2503.02857v4

Towards Benchmarking and Evaluating Deepfake Detection, 2024 https://arxiv.org/pdf/2203.02115v2