RAPPORT DE SYNTHÈSE SCIENTIFIQUE ET RÉGLEMENTAIRE

Frontières du Deepfake & Marquage de l'IA : Technologies de Génération, Hallucinations et Détection Proactive

Synthèse thématique et comparative des avancées scientifiques, réglementaires (AI Act 2026) et cryptographiques

Auteur / Éditeur : *Gemini Notebook (Analyseur de sources)*	***Domaine :*** *Intelligence Artificielle & Sécurité Numérique*
Corpus d'étude : *10 sources scientifiques et réglementaires*	***Date de publication :*** *Septembre 2026*

SYNTHÈSE EXÉCUTIVE

Résumé Exécutif : L'essor des modèles génératifs (GANs, Diffusion, LLMs) a profondément transformé la création de médias synthétiques, mais pose des défis majeurs en matière de désinformation, d'intégrité de la preuve numérique et d'hallucinations factuelles. Face aux limites de la détection passive a posteriori, la communauté scientifique et les législateurs convergent vers la **forensique proactive** basée sur le filigranage numérique (*watermarking*). Ce rapport offre une synthèse rigoureuse articulée autour de 5 axes : (1) les mécanismes de génération de Deepfakes, (2) la typologie des hallucinations des LLM, (3) les fondements du watermarking multi-modal (texte, visuel, audio), (4) les avancées cryptographiques du filigranage basé sur les attributs, et (5) l'encadrement juridique européen (Article 50 de l'AI Act) et les standards industriels (C2PA, SynthID).

1. Introduction & Paysage des Médias Synthétiques (Deepfakes)

L'IA générative (GenAI) a franchi un cap décisif grâce à l'émergence d'architectures d'apprentissage profond hautement performantes. Initialement limitées à des trucages rudimentaires ou des méthodes graphiques traditionnelles (copier-coller/splicing), les technologies actuelles reposent sur des Autoencodeurs Variationnels (VAEs), des Réseaux Adversaires Générateurs (GANs tels que StyleGAN2, DCGAN, PGGAN) et plus récemment des Modèles de Diffusion (DALL-E 2, Stable Diffusion) ainsi que des architectures Transformers (Dehghani & Saberi, 2025 ; Lai et al., 2025 ; Cao, 2025).

Typologie des manipulations visuelles et auditives :

- **Échange de visage (Face Swapping) :** Transfert de l'identité d'un visage source vers une vidéo destination en préservant la pose et l'expression (ex. Roop, DeepFaceLab, LivePortrait) (Dehghani & Saberi, 2025 ; Lai et al., 2025).
- **Reconstitution faciale (Face Reenactment) :** Transfert des mouvements et expressions d'un conducteur sur la personne cible tout en préservant son identité (ex. Face2Face) (Dehghani & Saberi, 2025 ; Lai et al., 2025).
- **Édition d'attributs (Attribute Editing) :** Modification ciblée de traits spécifiques (âge, genre, cheveux, peau, expressions) via des cGANs ou STGAN (Dehghani & Saberi, 2025 ; Lai et al., 2025).
- **Synchronisation labiale (Lip Sync) :** Alignement temporel des mouvements de la bouche d'une cible sur une bande audio arbitraire (ex. Wav2Lip, LIP-INCONS) (Dehghani & Saberi, 2025).
- **Clonage vocal & Conversion de voix (Voice Conversion - VC) :** Transformation du timbre vocal d'un locuteur vers un autre sans modifier le contenu sémantique, via des approches parallèles ou non-parallèles (ex. kNN-VC, Seed-VC, HiFi-GAN) (Dehghani & Saberi, 2025).

Si ces avancées offrent des opportunités remarquables dans l'industrie cinématographique, les effets visuels, la création d'avatars ou le doublage vidéo (Dehghani & Saberi, 2025 ; Lai et al., 2025), elles véhiculent des risques sociétaux, politiques et sécuritaires majeurs. L'accessibilité accrue d'outils open-source abaisse la barrière technique pour des acteurs malveillants, favorisant l'extorsion, le chantage, la désinformation électorale et l'affaiblissement de la confiance dans la preuve numérique juridique (Dehghani & Saberi, 2025 ; Lai et al., 2025).

2. Le Phénomène des Hallucinations dans les Modèles de Langage

Parallèlement aux falsifications visuelles et auditives, la génération de texte par les Grands Modèles de Langage (LLM) est confrontée à une vulnérabilité critique : l'**hallucination**. Une hallucination désigne la production par un modèle d'un contenu fluide, grammaticalement correct et convaincant, mais factuellement inexact, fictif ou non étayé par le contexte source (Alansari & Luqman, 2025 ; Ehrenström, 2025).

Comme le souligne Me Philippe Ehrenström (2025), les hallucinations ne sont pas un « bug » accidentel, mais la conséquence de la nature fondamentalement probabiliste des LLM. Les architectures Transformers prédisent le token suivant selon des motifs statistiques sans disposer d'un accès intrinsèque à la réalité ou d'un référentiel de vérité vérifiable. Il s'agit en quelque sorte de la « logique de l'idée » poussée à son terme : l'optimisation de la plausibilité linguistique au détriment de la véracité factuelle (Ehrenström, 2025).

Taxonomie rigoureuse des hallucinations (Alansari & Luqman, 2025) :

1. **Inexactitudes Factuelles (Factuality Hallucinations) :** Divergence par rapport aux faits du monde réel.

- *Contradiction factuelle :* Génération d'une affirmation en contradiction directe avec un fait établi (ex. situer la capitale d'un pays dans une autre ville).
- *Fabrication factuelle :* Invention totale d'entités, d'événements, d'articles juridiques ou de références scientifiques inexistantes.

2. **Inexactitudes de Fidélité (Faithfulness Hallucinations) :** Dérive par rapport au contexte fourni par l'utilisateur.

- *Incohérence d'instruction :* Non-respect des contraintes imposées par le prompt (format, longueur, directives).
- *Incohérence de contexte :* Altération ou omission des faits essentiels contenus dans le document source.
- *Incohérence logique :* Rupture de la chaîne de raisonnement interne au sein du texte produit.

Causes des hallucinations sur le cycle de vie du LLM (Alansari & Luqman, 2025) :

Étape du Cycle	Facteurs Déclencheurs d'Hallucination	Impacts Constatés
Curage des Données	Données bruitées, préjugés sociaux, fausses croyances imitative (*imitative falsehoods*), connaissances rares (*long-tail*) ou obsolètes.	Reproduction de préjugés et d'erreurs historiques répétées sur le web.
Architecture & Pré-entraînement	Mécanismes d'attention diffuse sur longues séquences, prédiction unidirectionnelle, stratégie *teacher forcing*, absence d'exemples négatifs.	L'erreur initiale s'amplifie en chaîne (effet *snowball*).
Alignement & RLHF	Sur-ajustement (*overfitting*), sycophancie (tendance à valider les biais de l'utilisateur) et décalage capacités/attentes.	Priorité accordée à la confiance de la réponse plutôt qu'à l'exactitude.
Inférence & Décodage	Prompts ambigus, échantillonnage stochastique à haute température, goulot d'étranglement Softmax, limites de raisonnement multi-étapes.	Sélection de tokens à faible probabilité factuelle.

3. Forensique Proactive vs Détection Passive des Deepfakes

Face à la prolifération des contenus synthétiques, deux paradigmes d'analyse forensique s'affrontent (Lai et al., 2025 ; Cao, 2025) :

- **La Détection Passive (A posteriori) :** Elle cherche à extraire des artefacts visuels, spectraux ou temporels à partir des images ou vidéos déjà diffusées (incohérences de posture, clignotement des yeux, flux optique, analyse fréquentielle). Bien qu'efficace sur certains benchmarks fermés, la détection passive souffre de graves limites : faible généralisation face à des techniques de falsification inconnues, lourdeur computationnelle, incapacité à prévenir la diffusion virale et extrême vulnérabilité aux bruits d'attaque anti-forensique (compression JPEG, flou, rognage) (Lai et al., 2025).
- **La Forensique Proactive (Watermarking) :** Elle consiste à intégrer une marque ou signature numérique imperceptible mais traçable au sein du média *au moment même de sa génération* ou avant sa mise en ligne. Le problème de la détection en monde ouvert est ainsi transformé en un problème de vérification de marque et de correspondance de clés (Lai et al., 2025 ; Cao, 2025).

AVANTAGES DU WATERMARKING PROACTIF

Les 4 Piliers Stratégiques de la Forensique Proactive (Lai et al., 2025) :
1. **Détection en temps réel :** Alerte immédiate du propriétaire du contenu avant ou dès la diffusion virale.
2. **Défense embarquée :** Protection passive continue intégrée directement au fichier média.
3. **Résistance au frelatage :** Conception cryptographique rendant le retrait du filigrane extrêmement difficile sans dégrader la qualité du média.
4. **Force probante juridique :** Fourniture d'éléments de preuve techniques irréfutables lors de litiges de propriété intellectuelle ou de poursuites pour diffamation.

4. Taxonomie et Mécanismes Techniques du Filigranage Multi-Modal

Le filigranage numérique varie considérablement selon la modalité du contenu (texte, image, vidéo, audio) et les objectifs d'inviolabilité recherchés (Lalai et al., 2024 ; Lai et al., 2025 ; Cao, 2025 ; GrosNoob, 2026).

A. Filigranage du Texte (LLMs) :

Contrairement aux images représentées par des valeurs de pixels continues, le texte est composé de tokens discrets. Les principales approches de watermarking textuel comprennent (Lalai et al., 2024 ; Cao, 2025 ; GrosNoob, 2026) :
• *Filigranage probabiliste (Listes Vertes/Rouges) :* Introduit par Kirchenbauer et al. (2023), cet algorithme divise le vocabulaire du modèle en une « liste verte » et une « liste rouge » à chaque étape de décodage, en utilisant une clé pseudo-aléatoire basée sur le token précédent. Le modèle privilégie les tokens verts. La détection s'effectue en analysant la surreprésentation statistique des tokens verts sans altérer la qualité de lecture.
• *Substitutions lexicales et syntaxiques :* Remplacement contrôlé de mots par des synonymes équivalents ou modification de la structure grammaticale (ex. transformations d'arbres syntaxiques) (Lalai et al., 2024).
• *Filigranage sémantique :* Invariant sémantique modifiant les représentations latentes des phrases pour résister à la reformulation (Lalai et al., 2024).

B. Filigranage Visuel (Images et Vidéos) :

• *Domaine Spatial :* Modification directe des bits de poids faible (LSB) ou des motifs de texture LBP. Rapide mais très sensible à la recompression (Lai et al., 2025 ; Cao, 2025).
• *Domaine Fréquentiel :* Insertion dans les coefficients transformés (DCT, DWT, SVD), offrant une excellente résistance à la compression JPEG et au flou (Lai et al., 2025 ; Cao, 2025).
• *Filigranage Réseau Profond (Deep Learning) :* Insertion de filigranes directement dans l'espace latent des modèles de diffusion ou GANs. Exemples majeurs : **SynthID** de Google DeepMind (intégré dans Imagen, Gemini, Veo) et **Stable Signature** de Meta (Lai et al., 2025 ; Cao, 2025 ; GrosNoob, 2026).

C. Filigranage Audio (Parole et Musique) :

Modifications inaudibles des composantes spectrales (DFT, DCT) ou temporelles (amplitude, phase, prosodie). Les modèles récents basés sur des réseaux profonds tels qu'**AudioSeal** (Roman et al., 2024 ; Cao, 2025) permettent de localiser exactement les fragments de voix synthétisés au sein d'un enregistrement sonore.

Modalité	Technique Principale	Mécanisme d'Embedding	Niveau de Robustesse
Texte	Probabiliste (Listes Vertes/Rouges)	Biais de sélection des logits lors du décodage des tokens	Moyenne (Vulnérable au paraphrasage lourd)
Image / Vidéo	Fréquentiel (DWT/SVD) & Latent Deep	Altération de coefficients fréquentiels ou vecteurs latents (SynthID)	Élevée (Résilient au rognage, flou, compression)
Audio / Voix	Modulation spectrale & AudioSeal	Masquage psychoacoustique et signatures latentes profondes	Élevée (Résilient au re-codage MP3 et au bruit)

5. Contrôle d'Accès Cryptographique : Filigranage Basé sur les Attributs

Jusqu'à récemment, les systèmes de filigranage indétectables souffraient d'une faille majeure lors de la délégation des capacités de détection : l'attribution d'une clé de détection globale donne à un contrôleur le pouvoir d'inspecter n'importe quelle production du modèle, sans restriction de périmètre. Cela engendre trois vulnérabilités critiques (Huang et al., 2026) :

1. **Assainissement du filigrane (Watermark Sanitization) :** Un attaquant disposant d'une clé publique ou globale peut interroger le détecteur en boucle jusqu'à effacer le filigrane sans altérer visuellement le média.
2. **Abus de périmètre (Scope Abuse) :** Une autorité mandatée pour vérifier des rapports publics peut utiliser sa clé pour scanner des documents confidentiels internes.
3. **Profilage d'utilisateurs (User Profiling) :** La détection systématique permet de tracer les habitudes d'usage de l'IA d'un individu à son insu.

Pour résoudre ce défi, Huang, Lee, Raffel et Tang (2026) ont introduit le premier schéma de **Filigranage Indétectable Basé sur les Attributs (Attribute-based Undetectable Watermarking)**. Dans ce cadre, chaque sortie générée est associée à des attributs (ex. domaine médical, finance, cours) via un classifieur de sorties robuste. Les clés de détection sont contraintes par des politiques logiques (*Constraint Circuits*). Une clé ne permet de détecter que les filigranes dont les attributs satisfont la politique définie, tandis que les contenus hors périmètre demeurent cryptographiquement indifférenciables de contenus non filigranés (Huang et al., 2026).

INNOVATION CRYPTOGRAPHIQUE (HUANG ET AL., 2026)

Architecture Cryptographique du Filigranage Basé sur les Attributs (Huang et al., 2026) :
• **CPRF (Constrained Pseudorandom Functions) :** Dérivation des clés de filigranage dépendant des attributs à partir d'une clé maître.
• **PRC (Pseudorandom Error-Correcting Codes) :** Codage de mots de code indétectables résistant aux bruits d'altération.
• **RandRecover :** Algorithme déterministe de récupération de la pseudo-aléatoire du modèle à partir du texte/média généré.
• **Garanties Sécuritaires Prouvées :**
- *Consistance :* Équivalence exacte entre la détection contrainte et la détection maître sur le périmètre autorisé.
- *Robustesse adaptative :* Maintien du signal même en cas de corruption d'une fraction δ des tokens.
- *Indétectabilité :* Indistinguabilité computationnelle absolue pour tout observateur non autorisé.
- *Justesse (Soundness) :* Impossibilité pour un détenteur de clé de détection d'accuser à tort un contenu sain.

6. Gouvernance, AI Act Européen et Normalisation Industrielle

Le passage du filigranage du laboratoire au monde réel est fortement accelerated par l'entrée en vigueur de cadres réglementaires stricts, en particulier en Europe avec l'**AI Act** (Leto.legal, 2026 ; GrosNoob, 2026).

A. Les Obligations de l'Article 50 de l'AI Act (2026) :

L'Article 50 de l'AI Act impose des exigences légales de transparence concernant les contenus générés par IA à compter du **2 août 2026** (avec une période de transition jusqu'au 2 décembre 2026 pour le marquage lisible par machine des systèmes existants) (Leto.legal, 2026 ; GrosNoob, 2026) :
• **Régime des Fournisseurs :** Obligation de marquer les contenus générés (texte, image, audio, vidéo) dans un format lisible par machine, détectable et résistant aux manipulations usuelles (via filigrane, métadonnées ou signatures de provenance) (Leto.legal, 2026).
• **Régime des Déployeurs (Entreprises/ETI) :** Obligation d'étiqueter de manière visible les *Deepfakes* (images/audios/vidéos simulant des personnes réelles) et d'indiquer l'usage de l'IA pour les textes informant le public sur des sujets d'intérêt général (sauf contrôle éditorial humain assumé) (Leto.legal, 2026).
• **Code de Bonne Conduite (10 juin 2026) :** Document opérationnel publié par la Commission Européenne établissant des standards techniques communs et une taxonomie partagée pour valider la conformité des entreprises (Leto.legal, 2026 ; GrosNoob, 2026).

B. Standards Industriels et Déploiement à Grande Échelle :

• **Coalition C2PA (Content Credentials) :** Standard ouvert soutenu par Adobe, Google, Microsoft, Meta, OpenAI et Amazon, permettant d'attacher des métadonnées de provenance cryptographiques directement aux fichiers numériques (GrosNoob, 2026 ; Cao, 2025).
• **Standard Inter-Industrie SynthID :** Initialement conçu par Google DeepMind, SynthID s'est imposé en 2026 comme le standard de filigranage multi-acteurs, adopté par OpenAI (ChatGPT, DALL-E, API), NVIDIA (Cosmos), ElevenLabs et Kakao, comptabilisant plus de 100 milliards de médias filigranés (GrosNoob, 2026).

C. Limites, Attaques Adaptatives et Perspectives :

Bien que le filigranage constitue une avancée majeure, plusieurs obstacles persistent :
1. *Attaques de paraphrasage & Réécriture :* Pour le texte, quelques reformulations légères par un autre LLM suffisent souvent à effacer la signature statistique (OpenReview, 2026 ; Lalai et al., 2024 ; GrosNoob, 2026).
2. *Modèles Open-Source :* Les modèles dont les poids sont ouverts (ex. Llama, Mistral) peuvent être modifiés pour désactiver le module de filigranage lors de la génération (GrosNoob, 2026).
3. *Risques de Faux Positifs & Vie Privée :* Risque d'accusation à tort pour des textes formels ou écrits par des non-natifs, et interrogations sur la traçabilité permanente des écrits (Lalai et al., 2024 ; GrosNoob, 2026).

7. Références Bibliographiques

Références complètes des sources scientifiques, réglementaires et techniques constituant le corpus de ce rapport d'analyse :

Ref / Clé	Auteurs & Titre de la Publication	Éditeur / Revue / URL
[1] Dehghani & Saberi (2025)	Dehghani, A., & Saberi, H. *Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools.*	*arXiv preprint / Faculty of AI, IHU.* https://arxiv.org
[2] Leto.legal / Schlumberger (2026)	Schlumberger, É. / Leto.legal. *AI Act : marquage des contenus IA (article 50, 2026) - Code de bonne conduite sur les contenus générés par IA.*	*Guide juridique Leto.legal (10 juin 2026).* https://www.leto.legal/guides/code-bonne-conduite-contenu-genere-ia
[3] Huang et al. (2026)	Huang, M. M.-Y., Lee, C.-W., Raffel, M., & Tang, E.-C. *Attribute-based Undetectable Watermarking for Generative AI Models.*	*Carnegie Mellon University, USC, UW / GitHub.* https://github.com/maxraffel/attribute-based-watermarking
[4] Lai et al. (2025)	Lai, Z., Arif, S., Feng, C., Liao, G., & Wang, C. *Enhancing Deepfake Detection: Proactive Forensics Techniques Using Digital Watermarking.*	*Computers, Materials & Continua, Vol. 82, No. 1, Tech Science Press.* DOI: 10.32604/cmc.2024.059370
[5] Lalai et al. (2024)	Lalai, H. N., Ramakrishnan, A. A., Shah, R. S., & Lee, D. *From Intentions to Techniques: A Comprehensive Taxonomy and Challenges in Text Watermarking for LLMs.*	*ACL Anthology (BITS Pilani, Penn State, Georgia Tech).* https://aclanthology.org
[6] Alansari & Luqman (2025)	Alansari, A., & Luqman, H. *Large Language Models Hallucination: A Comprehensive Survey.*	*arXiv preprint / KFUPM.* https://arxiv.org
[7] Ehrenström (2025)	Ehrenström, Ph. *Les hallucinations de l'IA générative : dérèglement des sens ou logique de l'idée ?*	*Blog Droit du travail & Protection des données (4 mai 2025).* https://droitdutravailensuisse.com/2025/05/04/les-hallucinations-de-lia-generative-dereglement-des-sens-ou-logique-de-lidee/
[8] OpenReview (2026)	Anonymous Authors. *Optimizing Adaptive Attacks Against Content Watermarks for Language Models.*	*OpenReview Forum.* https://openreview.net/forum?id=RKQcJ1IXNT
[9] GrosNoob.com (2026)	L'équipe GrosNoob. *Watermarking de l'IA : comment détecter qu'un texte ou une image a été généré par une machine.*	*Guide GrosNoob.com (30 juin 2026).* https://grosnoob.com/watermarking-ia-detecter-contenu-genere-explication/
[10] Cao (2025)	Cao, L. *Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities.*	*1st Workshop on GenAI Watermarking, ICLR 2025 / Amazon KDP.* https://www.amazon.com/dp/B0F2ZKH2R4