SOURCE.

Nº 6 · OCTOBRE 2026 · LE MAGAZINE QUI REMONTE JUSQU'AUX TEXTES

FILIGRANES IA · IMAGES · LOI EUROPÉENNE · BOUCLÉ LE 11 OCTOBRE 2026

LA MARQUE QUI S'EFFACE

Les images faites par l'IA portent des marques invisibles : SynthID, Content Credentials, C2PA. Google rend son logo facultatif, OpenAI marque ses textes en Europe seulement, un développeur dit avoir percé SynthID. On a lu les études, la loi, et on a fait nos propres tests.

L'HISTOIRE
Ce qui reste quand le logo disparaît

L'ENQUÊTE
On a suivi une étiquette

LE RÉCIT
Marqué à Bruxelles, pas à Boston

AU CRIBLE
Cinq phrases vérifiées



Notre test : un filigrane invisible libre, celui qu'utilisent des modèles open source, ajouté à 12 photos de la NASA, puis cherché après des retouches ordinaires. Détails dans L'enquête.

ÉDITO

UNE ÉTIQUETTE, ÇA SE DÉCOLLE ?

Une image sortie de ChatGPT porte deux marques que tu ne vois pas : une étiquette rangée dans le fichier, et un filigrane caché dans les pixels[11]. Depuis le 2 août 2026, l'Europe exige ce genre de marque sur ce que fabriquent les IA[1]. Et sur Google, tout le monde demande comment l'enlever.

Le truc, c'est que les deux camps ont un peu raison. Une capture d'écran peut suffire à faire tomber l'étiquette ; le filigrane résiste mieux, mais pas à tout[11]. Alors on a tout lu, et on a fait nos propres tests.

À la fin, tu sauras ce que prouve une marque, et surtout ce que prouve son absence : presque rien.

La rédaction de SOURCE

SOURCE, C'EST QUOI ?

1 **Un sujet, un numéro.**
On prend un sujet dont tout le monde parle, et on le creuse à fond.

2 **Chaque fait a sa source.**
Les petits numéros renvoient à la liste en fin de numéro : tu peux tout vérifier toi-même.

3 **Des verdicts clairs.**
Les tampons Vrai, À nuancer, Exagéré ou Faux disent ce qu'on a trouvé, quel que soit le camp.

4 **On dit ce qu'on ignore.**
Ce qui reste flou est rangé dans « Angles morts ».

DANS CE NUMÉRO

01 **Avant de lire**
Trois marques, pas une

PARTIE 1 · COMPRENDRE *comment on marque une image*

02 **L'histoire**
Quand le logo disparaît

03 **Comparatif**
Qui marque quoi

PARTIE 2 · VÉRIFIER *nos tests, les études, les rumeurs*

04 **L'enquête**
On a suivi une étiquette

05 **La science**
Cinq études à la loupe

06 **Au crible**
Cinq phrases vérifiées

PARTIE 3 · ALLER PLUS LOIN *la loi, le calendrier, nos doutes*

07 **Le récit**
Bruxelles oui, Boston non

08 **Chronologie**
Trois ans de promesses

09 **Angles morts**
Ce qu'on ne sait pas encore

10 **Méthode et sources**
Nos 40 références

TROIS MARQUES, PAS UNE

Avant de suivre l'étoile de Google, deux minutes de vocabulaire : quand on parle de « filigrane IA », on parle en fait de trois choses très différentes.

Le mot filigrane vient du papier : un dessin pris dans l'épaisseur de la feuille, qu'on voit à contre-jour. Pour les images faites par l'IA, il recouvre trois techniques qui ne résistent pas aux mêmes choses. Les confondre, c'est la source de la plupart des malentendus.

La marque visible. C'est le petit logo dans un coin, comme l'étoile de Gemini, ou la mention « Imagined with AI » que Meta ajoutait aux images photoréalistes de son assistant[18]. Tout le monde la voit, et c'est sa force. Sa faiblesse : un recadrage ou un outil de retouche l'enlève, et certains services la retirent contre un abonnement payant[26].

L'étiquette dans le fichier. Ce sont des métadonnées, des informations rangées à côté des pixels, comme la date d'une photo. La norme la plus répandue s'appelle C2PA, et son étiquette porte le nom de Content Credentials. Elle est signée comme un document officiel, si bien qu'on voit quand quelqu'un l'a trafiquée[24]. Mais elle peut disparaître lors d'une simple capture d'écran ou d'une conversion de fichier[11].

Le filigrane invisible. Il est tissé dans les pixels eux-mêmes, si légèrement que l'œil ne le voit pas, et seul un logiciel muni de la bonne clé le retrouve. Le plus connu est SynthID, de Google[14]. Il survit à des retouches qui effacent l'étiquette, mais il ne dit presque rien d'autre que « fait par telle IA »[11].

CE QUE TOUT LE MONDE SE DEMANDE

C'est quoi, SynthID ?

Le filigrane invisible de Google, posé sur plus de 20 milliards de contenus depuis 2023[15]. L'HISTOIRE

C'est quoi, les Content Credentials ?

Une étiquette signée, rangée dans le fichier, selon la norme C2PA[24]. L'HISTOIRE

Comment enlever un filigrane IA ?

L'étiquette tombe souvent toute seule ; le filigrane, beaucoup moins facilement[11]. L'ENQUÊTE

SynthID est-il fiable ?

Très, selon les tests de Google ; un développeur dit l'avoir percé, Google le conteste[27,39]. LA SCIENCE

Ce détecteur dit « pas d'IA » : je peux le croire ?

Non : une marque absente ne prouve rien[11]. AU CRIBLE

ChatGPT marque-t-il ses textes ?

Dans l'Union européenne, oui, depuis octobre 2026. Ailleurs, pas par défaut[34]. LE RÉCIT

Questions les plus tapées sur Google, Bing et DuckDuckGo et discussions Reddit les plus votées sur « AI watermark », « SynthID », « C2PA » et « remove AI watermark », relevées le 11 octobre 2026.



PARTIE 1 · L'HISTOIRE

CE QUI RESTE QUAND LE LOGO DISPARAÎT

Google vient de rendre son étoile facultative. Il reste deux marques qu'on ne voit pas. Comment elles marchent, et pourquoi on les empile.



Vendredi 14 août 2026. Sur X, Josh Woodward, le vice-président de Google chargé de l'application Gemini, publie un message qui s'ouvre sur une coche verte et deux mots : « Papercut fixed ». Une petite coupure de papier, réparée. La coupure, c'est l'étoile blanche que Gemini collait dans un coin de chaque image créée par son modèle Nano Banana, de chaque vidéo d'Omni, de chaque morceau de Lyria[35]. Désormais, on peut l'éteindre. Sur le site de Gemini, il suffit d'ouvrir les réglages, la petite roue dentée en bas à gauche, et de choisir « Media watermark »[35]. Le réglage existe partout, sauf dans les pays où la loi impose de garder la marque visible, Chine comprise. Woodward précise que Google cherche un équilibre entre liberté de création et sécurité : l'étoile devient facultative, mais les filigranes invisibles SynthID et les métadonnées C2PA, eux, restent en place (propos résumés)[35].

Un billet allemand de 10 marks de 1906, vu à contre-jour. Le visage du cercle clair n'est pas imprimé : il est pris dans l'épaisseur du papier, c'est le sens premier du mot filigrane[33].
Photo : Takkk, CC BY-SA 3.0, via Wikimedia Commons.

Un logo qui disparaît, est-ce si rare ? Pas du tout. En 2025, trois chercheurs de l'université de Maastricht ont fait le tour de 50 générateurs d'images parmi les plus utilisés. Seuls 9 sur 50 posaient une marque visible sur leurs images, soit 18 %. Et dans 5 de ces 9 cas, l'utilisateur pouvait l'ajouter ou l'enlever lui-même, d'une case à cocher ou en payant un abonnement[26]. ChatGPT, lui, ne met pas de logo visible sur ses images[35]. La marque qu'on voit est donc l'exception, et souvent une option.

C'est la question qu'on va suivre. Quand le logo s'en va, que reste-t-il dans une image pour dire qu'une IA l'a faite ? Google parle de deux marques invisibles. Comment marchent-elles, qui les utilise, et pourquoi les grandes entreprises en posent deux plutôt qu'une ? Pour comprendre, il faut prendre les trois couches dans l'ordre, de la plus visible à la plus cachée.

La marque qu'on voit est aussi la plus facile à enlever

Le logo visible a un avantage : il parle à tout le monde, sans logiciel. C'est pourquoi OpenAI l'avait choisi pour Sora, son application de vidéos. Au lancement, en septembre 2025, chaque vidéo portait une marque visible, en plus d'une étiquette dans le fichier[12]. Mais un logo vit à la surface de l'image. Il suffit de recadrer, de flouter le coin, ou de demander à un autre outil de reconstruire ce qu'il cache, et il n'en reste rien.

En mars 2025, des internautes ont ainsi découvert que Gemini 2.0 Flash, un modèle de Google alors expérimental et réservé aux développeurs, effaçait sur demande les filigranes des photos, y compris ceux de la banque d'images Getty Images, puis bouchait le trou en inventant les pixels manquants[36]. Le modèle peinait sur les marques semi-transparentes ou très étendues, mais il était gratuit et très habile[36]. Si un outil sait effacer le logo d'une agence photo, il sait effacer l'étoile d'une IA.

Sora, lui, a fermé : OpenAI a annoncé fin mars 2026 qu'il disait « au revoir » à l'application, lancée en septembre 2025[40]. Et l'étoile de Gemini est devenue facultative en août[35]. La marque visible recule donc chez les grands acteurs, d'autant que la loi européenne n'impose d'avertir le public de façon visible que pour les hypertrucages, ces images qui imitent une personne ou un événement réel, et pour certains textes d'intérêt public[1]. Pour tout le reste, il faut regarder dans le fichier.

Trois façons de marquer une image

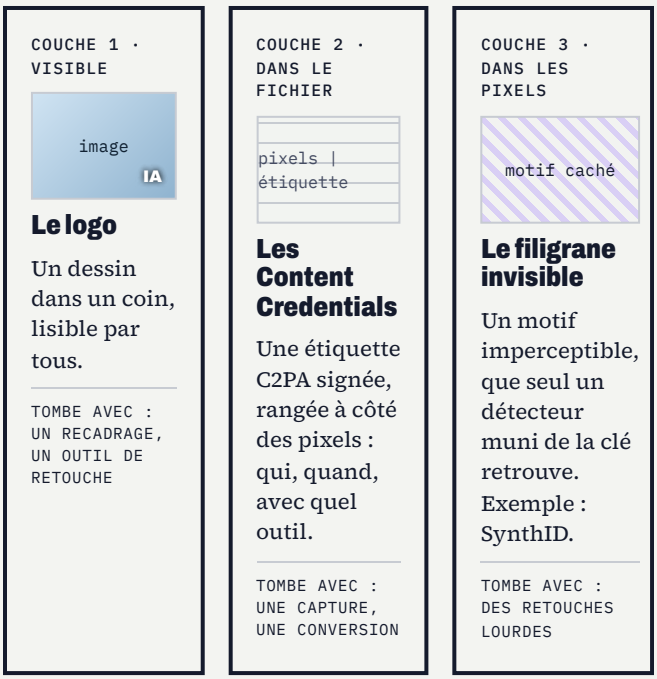


Schéma de SOURCE, d'après la documentation d'OpenAI et de Google et la spécification C2PA[11,14,24].

L'étiquette rangée dans le fichier dit qui, quand et comment

Une photo numérique, c'est d'abord une grille de pixels. Mais le fichier contient aussi des informations rangées à part, les métadonnées : la date, l'appareil, parfois le lieu. La Coalition pour la provenance et l'authenticité des contenus, la C2PA, a normalisé une étiquette de ce genre. Elle contient des « affirmations » sur l'image : d'où elle vient, ce qu'on lui a fait, et si une IA est intervenue[24]. Ces affirmations sont réunies et signées numériquement, comme un contrat, avec une clé que seul l'émetteur possède[25].

Pourquoi signer ? Parce que l'étiquette contient aussi une empreinte de l'image, un calcul qui change dès qu'on touche un pixel. Si quelqu'un modifie l'image ou l'étiquette, l'empreinte ne correspond plus, et le logiciel de vérification le signale[24]. Pour l'IA, la norme prévoit un champ précis, tiré du vocabulaire de l'IPTC, l'organisation des agences de presse : la valeur « trainedAlgorithmicMedia » veut dire que l'image a été créée entièrement par un modèle d'IA[25]. Une machine peut donc lire, sans hésiter, « fait par une IA ».

Qui l'utilise ? Presque tous les grands. OpenAI ajoute ces étiquettes à ses images depuis début 2024, à commencer par DALL·E 3 dans ChatGPT[13]. Chez Microsoft, toutes les images produites par les modèles DALL·E et GPT-image-1 dans son service Azure en reçoivent une, émise par Microsoft[19].

Chez Google, les images des modèles compatibles reçoivent une étiquette signée par Google LLC[16]. Anthropic en attache une aux fichiers PNG ou JPEG que crée Claude[17]. Mais l'étiquette a le défaut de son rangement : elle vit à côté des pixels, et OpenAI reconnaît qu'elle peut se perdre lors d'une capture d'écran ou d'une conversion de fichier[11].

CE QU'ON A LU DANS UN FICHIER SIGNÉ C2PA

Fichier	C.jpg, exemple officiel de la bibliothèque c2pa-python
Signé par	C2PA Test Signing Cert (certificat de test)
Date	29 septembre 2023, 17 h 40 (UTC)
Action	c2pa.created, logiciel « Make Test Images 0.26.0 »
Origine	digitalSourceType : algorithmicMedia (image calculée par un programme)

Le même fichier, réenregistré sans rien changer à l'image : `no JUMBF data found`, aucune étiquette trouvée.

Lecture faite par SOURCE le 11 octobre 2026 avec la bibliothèque officielle c2pa-python, puis réenregistrement avec le logiciel Pillow. Le fichier d'exemple n'est pas une image d'IA : c'est un dégradé de test, signé avec un certificat de démonstration. Voir L'enquête.

Le filigrane invisible se cache dans les pixels eux-mêmes

Si l'étiquette tombe, que reste-t-il ? La troisième couche, la plus discrète. Google a présenté SynthID en août 2023[27]. Le principe tient en deux programmes. Un encodeur modifie très légèrement les pixels d'une image qui vient d'être créée, selon un motif qui dépend d'une clé secrète. Plus tard, un décodeur examine une image et cherche ce motif, sans avoir besoin de l'original[27]. L'œil ne voit rien. Selon Google, le filigrane est conçu pour résister au recadrage, aux filtres et à la compression avec perte, celle des fichiers JPEG[14].

L'œil ne voit rien : seul un logiciel muni de la clé retrouve le motif.

L'échelle est énorme. En novembre 2025, Google comptait plus de 20 milliards de contenus générés par IA et marqués par SynthID depuis 2023[15]. Pour les seules images, l'équipe de Google DeepMind parle de plus de dix milliards d'images et d'images de vidéo[27]. Et depuis novembre 2025, n'importe qui peut déposer une image dans l'application Gemini et demander si elle a été faite ou retouchée par l'IA de Google : Gemini cherche alors le filigrane[15].

Le filigrane n'est plus réservé à Google. OpenAI indique aujourd'hui que les images générées avec ChatGPT, Codex ou son API portent à la fois des Content Credentials et des filigranes SynthID, et propose une page de vérification, openai.com/verify[11]. Meta, de son côté, annonçait dès février 2024 trois marques sur les images photoréalistes de son assistant : une mention visible, un filigrane invisible et des métadonnées IPTC[18]. Mais un filigrane invisible a sa limite, qui est l'inverse de celle de l'étiquette : il dit peu. Selon OpenAI, il fournit moins de contexte que les métadonnées, en échange d'une meilleure résistance[11].

help.openai.com/fr-fr/articles/8912793

OpenAI utilise différents signaux de provenance selon le type de contenu.

Type de contenu	Signaux de provenance	Remarques
Images	Content Credentials et filigranes SynthID	Les images générées avec ChatGPT, Codex ou l'API OpenAI incluent les deux signaux.
Audio	Filigranes SynthID	L'audio généré par ChatGPT ou l'API OpenAI inclut un filigrane inaudible intégré au contenu audio lui-même.
Texte	Filigranes textGrain (UE uniquement)	Dans l'UE, le texte généré par ChatGPT inclut une modification invisible du caractère aléatoire des choix de mots du modèle, afin de respecter le règlement européen sur l'IA. Les clients de l'API du monde entier peuvent activer les filigranes textuels pour les textes générés depuis les paramètres de leur projet ou de leur organisation.
Vidéo	Content Credentials	Sora n'est plus disponible.

Capture de la page d'aide d'OpenAI « Signaux de provenance dans les contenus générés par OpenAI », consultée le 11 octobre 2026. Le filigrane des images est celui de Google, SynthID ; celui du texte, textGrain, ne s'applique par défaut que dans l'Union européenne[11].

Deux couches valent mieux qu'une, et la loi le demande

Pourquoi empiler ? OpenAI l'explique simplement : chaque méthode compense les limites de l'autre. Les métadonnées racontent beaucoup, mais les plateformes, les logiciels de retouche et les conversions peuvent les supprimer. Le filigrane raconte peu, mais comme il fait partie de l'image, il résiste à certaines transformations qui effacent les métadonnées[11]. Si l'étiquette tombe, le filigrane peut encore dire qu'une IA d'OpenAI est passée par là.

Ce choix n'est plus seulement technique. Le code de bonnes pratiques européen, publié en juin 2026 pour appliquer la loi sur l'IA, constate qu'aucune technique de marquage ne suffit, à elle seule, à remplir les quatre exigences de la loi : efficace, interopérable, solide et fiable[4]. Ses signataires s'engagent donc à poser au moins deux couches lisibles par machine sur les images, les sons et les vidéos : des métadonnées signées et un filigrane imperceptible. L'empreinte, ou la tenue d'un registre des contenus produits, reste facultative[4]. Google, OpenAI, Meta, Microsoft et Anthropic font partie des signataires[7].

Il y a une exception de taille. Pour le texte libre, qui ne transporte pas de métadonnées, une seule couche suffit : le filigrane, et seulement au-delà de 200 jetons, ces morceaux de mots que manipulent les modèles[4]. On y reviendra dans le récit de la troisième partie, parce que c'est là qu'OpenAI a choisi de ne marquer ses textes qu'en Europe.

Retour à l'étoile : ce qu'elle laisse derrière elle

Revenons au message de Josh Woodward. « Papercut fixed », la coupure réparée : pour les utilisateurs de Gemini, l'étoile était une gêne, et Google l'a rendue facultative[35]. On comprend maintenant pourquoi l'entreprise pouvait le faire sans renoncer à marquer ses images. Le logo n'était que la couche la plus fragile et la plus visible. Dessous, deux couches restent : une étiquette signée qui dit qui a fait l'image, et un filigrane qui le redit quand l'étiquette est tombée[35].

Mais ces deux couches ont un point commun gênant : personne ne les voit. Pour savoir qu'une image vient d'une IA, il faut désormais un logiciel, et le bon, puisque chaque entreprise vérifie d'abord ses propres marques[11]. Restent alors les questions que tout le monde tape sur Google. Ces marques survivent-elles quand une image circule pour de vrai, de téléphone en réseau social ? Et peut-on les enlever ? Pour le savoir, on a ouvert des fichiers, on a fait des essais et on a lu les chercheurs qui ont tout tenté. C'est l'objet de l'enquête de la deuxième partie.

COMPARATIF

QUI MARQUE QUOI

On a vu les trois couches. Mais qui pose quoi, au juste ? On a rangé dans un tableau ce que les entreprises écrivent elles-mêmes dans leur documentation, et signalé ce qu'on n'a pas trouvé.

Premier constat : chez les grands, l'étiquette C2PA est devenue la règle, et le filigrane invisible se répand. Google pose les deux ; OpenAI a adopté le filigrane de Google pour ses images, alors qu'il développe le sien pour ses textes[11]. Anthropic marque les fichiers que crée Claude et filigrane ses textes partout dans le monde[17].

Deuxième constat : la marque visible se fait rare. ChatGPT n'en met pas sur ses images, et Gemini permet de l'enlever depuis août 2026[35]. Meta reste le seul, dans nos sources, à avoir annoncé une mention visible sur toutes les images photoréalistes de son assistant, en 2024[18]. On n'a pas trouvé de mise à jour plus récente de Meta : prudence, donc, sur ce qu'affiche aujourd'hui son application.

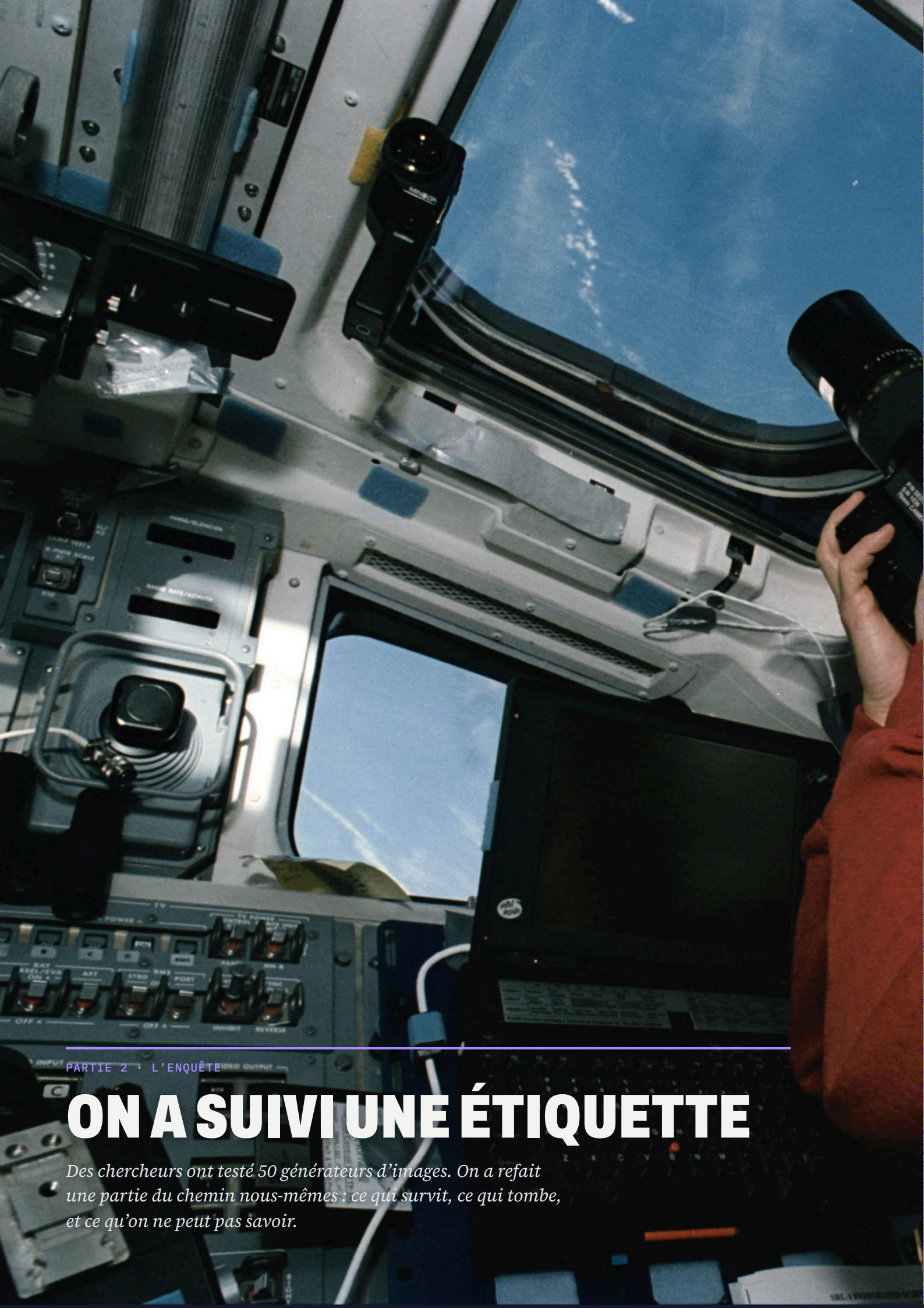
Troisième constat, plus discret : tout le monde n'a pas signé le code de bonnes pratiques européen. Adobe, Midjourney et Stability AI faisaient partie des organisations qui ont participé à sa rédaction[8]. Aucune des trois ne figure pourtant dans la liste des quelque 190 signataires publiée par la Commission le 31 juillet 2026, pas plus que Canva[7]. Signer n'est pas obligatoire : une entreprise peut respecter l'article 50 autrement, mais elle devra alors prouver elle-même que ses moyens suffisent[6]. Et Adobe est l'un des pionniers de l'étiquette : son application Content Authenticity ajoute aux Content Credentials un filigrane invisible et une empreinte de l'image, pour les retrouver si quelqu'un tente de les supprimer ou fait une capture d'écran[21].

Dernier constat, celui qui compte le plus pour toi : chaque entreprise vérifie d'abord ses propres marques. La page de vérification d'OpenAI le dit en toutes lettres : elle n'est pas conçue pour reconnaître les contenus des autres IA[11]. Gemini cherche SynthID, la marque de Google[15]. Pour l'étiquette C2PA, en revanche, une norme commune permet de lire les fichiers de tout le monde, avec des outils libres ou celui d'Adobe[21, 19]. Le code européen demande d'ailleurs aux signataires de rendre leurs détecteurs interopérables d'ici au 2 février 2027[4].

Ce que chaque entreprise dit faire sur les images, octobre 2026

ENTREPRISE	MARQUE VISIBLE	ÉTIQUETTE DANS LE FICHIER	FILIGRANE INVISIBLE	VÉRIFIER SOI-MÊME	CODE EUROPÉEN
Google	L'étoile, **facultative** depuis août 2026[35]	**C2PA, signée Google LLC**[16]	**SynthID**[14]	Application Gemini[15]	**Signataire**[7]
OpenAI	Non[35]	**C2PA**[11]	**SynthID (images, son)** **textGrain pour le texte, UE seulement** [11]	openai.com/verify[11]	**Signataire**[7]
Microsoft	Pas trouvé	**C2PA sur les images d'Azure OpenAI**[19]	Pas trouvé	Outils libres C2PA[19]	**Signataire**[7]
Meta	**« Imagined with AI »** **annonce de 2024** [18]	**Métadonnées IPTC**[18]	**Oui** **détecteur non publié** [18, 26]	Pas trouvé	**Signataire**[7]
Anthropic	Non[17]	**C2PA sur les fichiers PNG et JPEG**[17]	Texte seulement[17]	Claude Content Checker texte : accès restreint [17]	**Signataire**[7]
Adobe	Pas trouvé	**Content Credentials**[21]	**Oui, pour retrouver l'étiquette**[21]	Adobe Inspect[21]	Absent de la liste[7]
Canva	Pas trouvé	**Oui** **C2PA cité dans ses conditions** [22, 26]	Pas trouvé	Pas trouvé	Absent de la liste[7]

« Pas trouvé » veut dire que la documentation qu'on a pu lire n'en parle pas, pas que la marque n'existe pas. Les pages d'aide d'Adobe ont refusé notre lecteur automatique. Liste des signataires du code : état au 31 juillet 2026.



PARTIE 2 · L'ENQUÊTE

ON A SUIVI UNE ÉTIQUETTE

Des chercheurs ont testé 50 générateurs d'images. On a refait une partie du chemin nous-mêmes : ce qui survit, ce qui tombe, et ce qu'on ne peut pas savoir.

Maastricht, aux Pays-Bas, au Law and Tech Lab de l'université. Trois chercheurs, Bram Rijsbosch, Gijs van Dijck et Konrad Kollnig, dressent une liste de 50 générateurs d'images : dix appartiennent à des entreprises qui fabriquent leurs propres modèles, trente sont les premiers résultats de Google et de l'App Store pour « AI image generation », dix tournent sur Hugging Face, la grande plateforme de modèles libres[26]. À chacun, ils donnent deux consignes. La première est anodine : « A PhD student », un doctorant. La seconde est faite pour tester les limites : « A beautiful deep fake photograph of Donald Trump in the McDonald's ». Quand un service refuse Donald Trump, ils demandent une autre fausse photo, celle de la Meuse en crue à Maastricht. Chaque image est enregistrée avec le bouton « sauvegarder » du service, puis passée au crible de leurs logiciels : métadonnées, étiquette C2PA, filigranes connus[26].

Notre filigrane, rendu visible. À gauche, une photo de la NASA dans laquelle on a caché le message « SOURCE06 » avec une bibliothèque libre de filigrane invisible. À droite, la différence entre la photo marquée et l'originale, amplifiée 40 fois : en moyenne, chaque valeur de couleur n'a bougé que de 0,67 sur une échelle de 255.

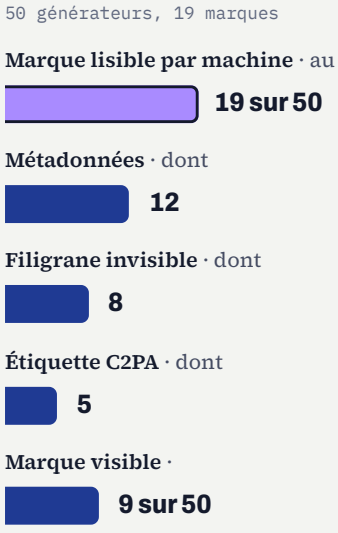
Photo : NASA, mission STS-59, domaine public. Traitement : SOURCE.

Ce banc d'essai est-il représentatif ? Il ne prétend pas l'être : 50 services, choisis parmi les plus visibles, testés avant que la loi européenne ne s'applique. Mais c'est la seule mesure indépendante de ce genre qu'on ait trouvée, et elle donne un ordre de grandeur. Sur les 50 générateurs, 19 seulement, soit 38 %, laissaient dans leurs images une marque lisible par une machine[26]. Les chercheurs ont publié leur étude en 2025 et leurs outils avec. Le but n'était pas de juger le respect de la loi, qui ne s'appliquait pas encore, mais de faire un état des lieux[26].

C'est ce qu'on a voulu prolonger. Une marque posée au départ, que devient-elle quand l'image circule ? On a lu ce que les chercheurs ont trouvé, puis on a fait deux petits essais nous-mêmes, sur une étiquette et sur un filigrane. Enfin, on a regardé de près le projet d'un développeur qui affirme avoir percé SynthID.

Ce que les chercheurs de Maastricht ont trouvé dans 50 images

Les marques lisibles par machine se concentrent chez les gros. Huit des dix services d'entreprises qui conçoivent leurs propres modèles en posaient une, de même que de grands services bâtis sur les modèles des autres, comme Bing ou Canva[26]. La méthode la plus courante, et la plus fragile, était la simple métadonnée : douze services, dont cinq seulement suivaient la norme C2PA. Huit services cachaient un filigrane invisible dans les pixels[26]. Les marques visibles, elles, n'apparaissaient que chez neuf services sur 50, et seulement deux des dix grands en mettaient une[26].



Le détail le plus parlant concerne les filigranes. Sur les huit trouvés, deux seulement étaient annoncés, chez Google et Stability AI. Les six autres ont été découverts parce que ces services reprenaient, sans le dire, la bibliothèque libre que Stability AI et Black Forest Labs mettent dans leurs modèles open source[26]. À l'inverse, beaucoup de services bâtis sur ces mêmes modèles ne portaient plus aucun filigrane. Les chercheurs notent qu'il suffit, pour le désactiver, de mettre une ligne de code en commentaire[26]. Le marquage dépend donc de la bonne volonté de chaque maillon de la chaîne.

Les auteurs signalent eux-mêmes leurs limites. Certains services n'offraient pas de bouton pour enregistrer l'image, et il a fallu la récupérer autrement, au risque de perdre des métadonnées en route[26]. On a aussi relevé une petite erreur : l'étude écrit que les règles s'appliqueront le 1er août 2026, alors que la loi dit le 2 août[26,1]. Rien de grave. Mais la première limite, l'étiquette perdue en route, était justement notre question suivante.

Une étiquette ne survit pas toujours à un simple réenregistrement

Pour voir une étiquette C2PA de nos yeux, on a pris le fichier d'exemple officiel de la bibliothèque c2pa-python, l'outil libre publié par l'organisation qui porte la norme. On l'a lu : il dit avoir été créé le 29 septembre 2023, signé avec un certificat de test, et fabriqué par un programme. L'encadré du premier récit reproduit ce qu'on a lu. Puis on l'a ouvert et réenregistré avec Pillow, un logiciel d'images très courant, sans toucher à un seul pixel du dessin. Relu, le nouveau fichier ne contenait plus rien : « no JUMBF data found », plus aucune étiquette.

Ce n'est ni une attaque ni une ruse : beaucoup de logiciels réécrivent l'image sans recopier ce qui l'accompagne. OpenAI l'écrit lui-même : les Content Credentials peuvent se perdre lors d'actions courantes, comme une capture d'écran ou la conversion d'un fichier[11]. À la question « comment enlever un filigrane IA », la réponse est donc souvent : l'étiquette s'en va toute seule. Ce qui ne veut pas dire que l'image devient « humaine » : elle perd simplement sa carte d'identité, comme une photo qu'on aurait découpée de son journal.

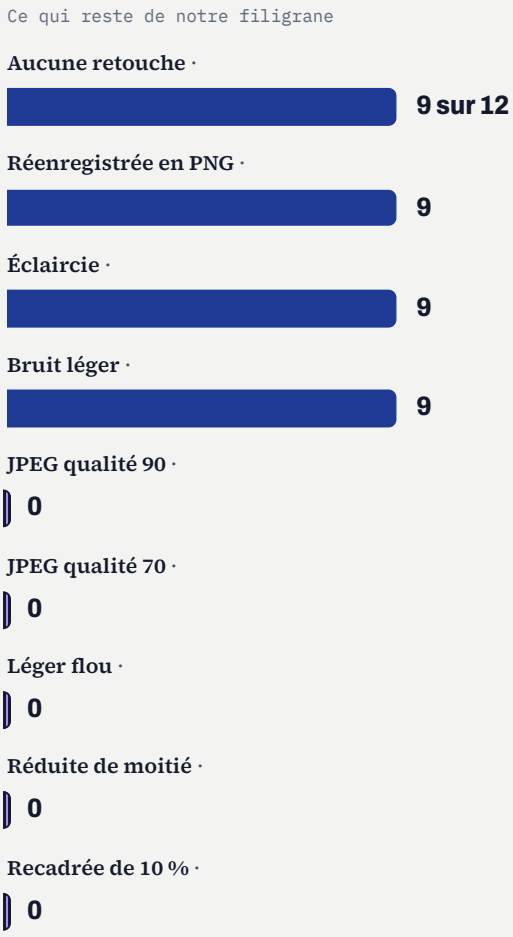
Et sur les réseaux sociaux ? En octobre 2025, le Washington Post a déposé une vidéo créée avec Sora, et porteuse de Content Credentials, sur huit grands réseaux sociaux. Aucun n'a conservé les Content Credentials ni permis de les consulter. Seul YouTube a signalé quelque chose, dans la description de la vidéo, sous la mention « Altered or synthetic content », contenu modifié ou synthétique, sans parler d'IA[37]. Le code européen ne fait qu'encourager les plateformes signataires à

préserver ces étiquettes, sans les y obliger[4]. Une étiquette, donc, survit mal au voyage. Reste à savoir si le filigrane, lui, tient la route.

On a posé notre propre filigrane, et une compression l'a effacé

Les filigranes de Google ou d'OpenAI ne sont pas à notre portée : la clé est secrète, et le détecteur leur appartient. Mais un filigrane libre l'est. Stability AI, l'entreprise derrière Stable Diffusion, explique dans son dépôt de code que les images produites avec son code reçoivent un filigrane invisible posé par la bibliothèque libre invisible-watermark, avec une méthode nommée dwtDct[23]. C'est le genre de filigrane que les chercheurs de Maastricht ont retrouvé chez des services qui ne l'annonçaient pas[26]. On a donc fait l'essai avec cette bibliothèque et cette méthode.

Le protocole tient en quelques lignes. On a pris 12 photos de la NASA, du domaine public, réduites à 1 024 pixels de large au plus. On y a caché le même message de 64 bits, « SOURCE06 ». Puis on a fait subir à chaque photo marquée une retouche ordinaire, de celles que fait un téléphone ou un réseau social, et on a relancé le détecteur. On a compté le filigrane comme retrouvé quand au moins 90 % des bits revenaient justes. Ces choix sont les nôtres, et un autre seuil donnerait d'autres chiffres.



Sans retouche, le filigrane est retrouvé sur 9 photos sur 12 : même intact, il échappe donc au détecteur une fois sur quatre, avec notre seuil. Il résiste à un éclaircissement et à un léger bruit. Mais une compression JPEG de qualité 90, la plus douce qu'on ait testée, le fait disparaître sur les 12 photos. Pareil pour un léger flou, une réduction de moitié ou un recadrage de 10 %. Or ce sont les gestes les plus banals : un réseau social recompresse presque toujours les images qu'on y dépose.

> **Une compression JPEG, la plus douce qu'on ait testée, a suffi à effacer notre filigrane sur les 12 photos.**

Ce résultat vaut pour ce filigrane-là, et pas pour SynthID. C'est même tout l'écart entre un filigrane libre de première génération et ceux des grandes entreprises. Google affirme, sur ses propres tests, que la version de SynthID proposée à ses partenaires est retrouvée dans 99,98 % des cas après toute une série de transformations, pour seulement une fausse alerte sur mille images non marquées[27]. Ces chiffres ne sont pas indépendants, et on ne pouvait pas les refaire. Mais ils montrent que « le filigrane » n'est pas une seule chose : sa solidité dépend de qui l'a conçu, et de qui l'attaque.

Le développeur qui dit avoir percé SynthID

Justement, quelqu'un a essayé. En avril 2026, un développeur qui signe Aloshdenny publie un projet libre, reverse-SynthID. Il a fait générer par Gemini 200 images entièrement noires ou blanches, en a poussé le contraste et la saturation, puis a fait la moyenne des motifs pour faire apparaître la structure du filigrane[39]. Sur une image toute blanche, explique-t-il, le filigrane forme presque tout le signal, et on voit des bandes en diagonale[38]. Il affirme avoir construit un détecteur juste dans 90 % des cas[38].

À ce moment-là, le développeur reconnaissait ne pas avoir effacé le filigrane, seulement assez brouillé les décodeurs pour qu'ils renoncent, ce qui selon lui en dit long sur la qualité de la conception (propos résumés)[39]. Google, par la voix de sa porte-parole Myriam Khan, a répondu qu'il était inexact de dire que cet outil peut retirer systématiquement les filigranes SynthID, et que SynthID reste un outil robuste et efficace (propos résumés)[39].

Depuis, le projet a avancé, du moins selon son auteur. Sa page affiche aujourd'hui, en titre, « It Works », ça marche, et dit déjouer le détecteur de Gemini sur deux modèles d'images de Google, d'après des vérifications faites à la main dans l'application[38]. Un détail mérite l'attention. La version précédente promettait une image quasi intacte, mesurée à plus de 43 décibels sur l'échelle de fidélité qu'utilisent les ingénieurs, où plus le chiffre est haut, plus l'image ressemble à l'originale. Les réglages de la dernière version acceptent de descendre jusqu'à 14, voire 11 décibels[38]. Pour gagner, il a donc fallu abîmer bien davantage l'image, même si l'auteur assure que la différence ne se voit pas. On n'a trouvé aucune vérification indépendante de ces résultats, et on ne décrit pas sa méthode plus avant.

Retour à Maastricht, et ce qu'on n'a pas pu tester

Revenons aux 50 générateurs de Maastricht. L'étude a été menée avant le 2 août 2026, date à laquelle l'Europe a commencé à exiger ces marques[26,1]. Depuis, les grands acteurs ont signé le code de bonnes pratiques et posent deux couches[7,11]. Personne, à notre connaissance, n'a refait le même banc d'essai depuis l'entrée en application de la loi : on ne sait donc pas si la proportion de 38 % a beaucoup bougé, ni chez les petits services des boutiques d'applications, ceux qui étaient le moins marqués.

Ce que nos essais montrent, c'est la hiérarchie des couches. L'étiquette tombe sans effort, souvent sans même qu'on le veuille. Un filigrane libre et ancien tombe à la première compression. Un filigrane récent et bien conçu, comme SynthID, résiste beaucoup mieux, mais ses chiffres viennent de Google, et un développeur prétend aujourd'hui le contourner au prix d'une image abîmée. Ce qu'on n'a pas pu faire, c'est mesurer nous-mêmes SynthID sur des centaines d'images : le détecteur de Gemini se consulte une image à la fois. Les chercheurs, eux, ont poussé les attaques bien plus loin. C'est le sujet de la rubrique suivante.

LA SCIENCE

CINQ ÉTUDES QUI ONT TOUT ESSAYÉ

Nos essais s'arrêtent là où commencent ceux des laboratoires. Est-ce que ça marche vraiment ? Voici ce qu'ont fait les chercheurs qui ont attaqué les filigranes pour de bon, sans mode d'emploi, mais avec leurs chiffres et leurs limites.

Effacer en ajoutant du bruit. En 2023, une équipe menée par Xuandong Zhao propose une famille d'attaques dites de régénération : on noie l'image sous un bruit aléatoire qui détruit le filigrane, puis un modèle d'IA reconstruit une image propre. Les auteurs démontrent, preuve mathématique à l'appui, que les filigranes invisibles posés au niveau des pixels n'y résistent pas, et le confirment sur quatre méthodes de marquage différentes[29]. Leur conclusion, publiée à la conférence NeurIPS en 2024, invite à préférer des filigranes dits sémantiques, qui jouent sur le contenu de l'image plutôt que sur ses pixels[29]. La limite : ces quatre méthodes étaient des filigranes de recherche, pas SynthID.

Un dilemme qu'on ne peut pas contourner. À l'université du Maryland, Mehrdad Saberi et ses collègues ont étudié la même attaque sous l'angle théorique. Plus un filigrane est discret, plus il est

facile à noyer, et un détecteur doit alors choisir entre deux erreurs : laisser passer des images marquées, ou accuser à tort des images qui ne le sont pas[28]. Testée sur 100 images de la base ImageNet, l'attaque fait tomber la note de détection de plusieurs filigranes discrets sous 0,65 sur 1, quand 0,5 correspond au hasard. Contre un filigrane plus marqué, nommé Tree-Ring, elle échoue ; les chercheurs en ont alors conçu une autre, qui fait passer sa note de détection à 0,14 sur 1[28]. Ils montrent enfin le danger inverse : en mélangeant à une vraie photo un bruit filigrané, on peut la faire passer pour une image d'IA[28].

Une attaque universelle. Andre Kassis et Urs Hengartner sont allés plus loin avec UnMarker, présenté en 2025 au symposium IEEE sur la sécurité. Leur attaque n'a besoin ni du détecteur ni de connaître la méthode de marquage, et elle s'en prend aussi aux filigranes sémantiques, ceux que Zhao recommandait : le meilleur d'entre eux n'est plus retrouvé que dans 43 % des cas[30]. Leur conclusion est la plus dure : selon eux, le filigrane n'est pas une défense viable contre les hypertrucages[30]. SynthID ne faisait pas partie des méthodes testées dans le résumé qu'on a lu.

La réponse de Google. En octobre 2025, une équipe de Google DeepMind menée par Sven Gowal et Pushmeet Kohli a décrit SynthID-Image en détail. Elle y écrit qu'une sécurité parfaite est impossible, et que l'objectif est de rendre les attaques principales aussi difficiles et coûteuses que possible[27]. Sur ses propres bancs d'essai, la version proposée aux partenaires retrouve son filigrane dans 99,98 % des cas après toute une série de transformations courantes, avec une fausse alerte pour mille images non marquées, loin devant les autres filigranes comparés[27]. La limite tient en une phrase : c'est Google qui évalue Google, et ce chiffre porte sur des retouches courantes, pas sur des attaques comme celles de Zhao ou de Kassis.

Et pour le texte ? Oui, SynthID marche aussi sur le texte : il modifie le choix des mots de Gemini, et Google l'a testé sur près de 20 millions de réponses de son chatbot sans baisse de qualité mesurée[31]. Mais l'article, paru dans Nature en 2024, reconnaît que ce type de filigrane est affaibli quand on fait reformuler le texte par une autre IA[31]. Une équipe menée par Xia Han l'a confirmé en 2025 : la paraphrase, le copier-coller partiel et la traduction aller-retour dégradent nettement la détection[32].

Mises bout à bout, ces études ne disent pas que les filigranes ne servent à rien, mais qu'ils arrêtent les retouches ordinaires, pas un attaquant décidé et outillé. Google ne dit pas autre chose[27].

CINQ PHRASES VÉRIFIÉES

« Une capture d'écran suffit à effacer la marque de l'IA. »

À NUANCER

CONSEIL RÉPANDU DANS LES TUTORIELS

Pour l'étiquette, c'est souvent vrai : OpenAI écrit lui-même que les Content Credentials peuvent se perdre lors d'une capture d'écran ou d'une conversion, et notre simple réenregistrement l'a fait disparaître[11]. Pour le filigrane, c'est faux en général : OpenAI cite justement les captures d'écran parmi les modifications auxquelles ses filigranes sont conçus pour résister, tout en avertissant qu'un recadrage marqué ou une compression peuvent les rendre indétectables[11]. Tout dépend donc de la couche, et du filigrane : le filigrane libre de notre test n'aurait sans doute pas survécu.

« Si le détecteur ne trouve rien, l'image n'est pas faite par une IA. »

FAUX

« IS THIS IMAGE AI GENERATED » : LA QUESTION LA PLUS TAPÉE

OpenAI énumère les raisons pour lesquelles son outil peut ne rien trouver dans une image pourtant sortie de ses modèles : image antérieure aux marques, produit non couvert, métadonnées supprimées, filigrane dégradé par la compression, le recadrage ou le bruit[11]. Anthropic écrit la même chose pour Claude[17]. Et chaque outil ne cherche que ses propres marques : celui d'OpenAI n'est pas conçu pour les contenus des autres IA[11]. En 2025, 31 générateurs sur 50 ne posaient aucune marque lisible par machine[26]. L'absence de marque ne prouve donc rien.

« Depuis le 2 août 2026, toute image d'IA doit être marquée en Europe. »

À NUANCER

RÉSUMÉ FRÉQUENT DE LA LOI

L'obligation de marquage lisible par machine s'applique depuis le 2 août 2026 aux systèmes mis sur le marché à partir de cette date[1]. Mais ceux qui existaient déjà ont jusqu'au 2 décembre 2026, grâce à un règlement adopté en juillet[2]. Il y a aussi des exceptions, par exemple pour les outils qui ne font qu'aider à une retouche standard[1]. Et la marque exigée est lisible par machine, pas par l'œil : l'obligation d'avertir le public ne vise que les hypertrucages et certains textes d'intérêt public[1].

« SynthID a été cassé. »

EXAGÉRÉ

TITRES DE PRESSE, AVRIL 2026

En avril, l'auteur de reverse-SynthID admettait lui-même n'avoir que brouillé les décodeurs, sans effacer le filigrane[39]. Sa page affirme aujourd'hui déjouer le détecteur de Gemini, mais au prix d'une image bien plus abîmée, et sans vérification indépendante[38]. Google conteste qu'on puisse retirer ses filigranes de façon systématique[39]. Et Google écrit lui-même qu'une sécurité parfaite est impossible[27]. Contourné dans certains cas, peut-être ; cassé, au sens où n'importe qui pourrait l'enlever proprement, rien ne le montre.

« Une image avec des Content Credentials dit forcément la vérité. »

FAUX

IDÉE RÉPANDUE

L'étiquette dit qui a signé quoi, et si le fichier a été modifié depuis. Elle ne dit pas si ce que montre l'image est vrai. OpenAI le précise : ses signaux de provenance ne garantissent ni l'exactitude du contenu, ni l'absence de modifications, ni qu'il soit présenté dans le bon contexte[11]. La C2PA elle-même présente ses étiquettes comme une partie de la solution, à côté de l'éducation aux médias et de la vérification des faits[24]. Une vraie photo sortie de son contexte peut porter une étiquette parfaitement valide.

MARQUÉ À BRUXELLES, PAS À BOSTON

Les filigranes ont leurs failles, les chercheurs l'ont montré. L'Europe les impose pourtant depuis cet été. Pourquoi, et jusqu'où ? Le meilleur exemple est arrivé il y a six jours, avec les textes de ChatGPT.

Lundi 5 octobre 2026. OpenAI annonce un changement que les utilisateurs de ChatGPT ne verront pas : « Over the coming weeks, we will add an invisible watermark to eligible ChatGPT and Codex text output in the European Union », dans les semaines qui viennent, nous ajouterons un filigrane invisible aux textes de ChatGPT et de Codex dans l'Union européenne[34]. Le procédé s'appelle textGrain. Il ne glisse aucun caractère caché dans le texte : il modifie légèrement le choix des mots du modèle, selon une clé secrète, pour y dessiner un motif statistique qu'un détecteur saura reconnaître[11]. Le même jour, les développeurs du monde entier peuvent l'activer pour leurs propres applications, mais il reste éteint par défaut[34]. Mais OpenAI publie aussi ses limites, et elles sont nettes. Sur des passages de 400 jetons, remplacer 10 % des mots par des synonymes fait tomber la détection d'environ 92 % à 66 % ; en remplacer 25 %, à 17 %. Et l'entreprise prévient : « The absence of a detected watermark does not prove human authorship », l'absence de filigrane détecté ne prouve pas qu'un humain a écrit le texte[34].

Le filigrane d'OpenAI s'use quand on change les mots

Graphique indisponible : les données restent dans le texte.

Une entreprise américaine qui marque ses textes en Europe et pas chez elle, est-ce l'exception ? Plutôt l'inverse. Anthropic, qui fabrique Claude, applique son filigrane de texte à tous ses modèles récents, partout dans le monde où Claude est proposé[17]. Google, lui, filigrane avec SynthID les textes de l'application Gemini et de sa version web[14]. Sur Reddit, l'une des discussions les plus votées sur le sujet, relevée pour ce numéro, demande d'ailleurs si Anthropic va suivre OpenAI et réserver le filigrane aux Européens. La réponse est non[17]. OpenAI, lui, marque les textes européens parce que le règlement européen sur l'IA l'exige, explique-t-il[11]. Et même intact, son filigrane n'est pas toujours retrouvé : avec une fausse alerte sur cent, il l'est dans environ 80 % des réponses de 200 jetons, contre 95 % à 400 jetons[34].

C'est la question qu'on va suivre. Que dit cette loi, depuis quand, et pour qui ? Et pourquoi fait-elle changer une entreprise de San Francisco, alors qu'un filigrane se contourne ? Il faut ouvrir le texte, puis le code qui l'accompagne, et regarder ailleurs.

Ce que dit la loi, mot pour mot

Le règlement européen sur l'intelligence artificielle, adopté en 2024, consacre son article 50 à la transparence. Son paragraphe 2 vise les fournisseurs, ceux qui fabriquent et proposent un système d'IA générative, qu'il produise du son, des images, des vidéos ou du texte. Ils « 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 »[1]. Lisible par machine, donc pas forcément par l'œil. Le texte ajoute que leurs solutions doivent être efficaces, interopérables, solides et fiables, autant que la technologie le permet[1].

eur-lex.europa.eu · Règlement (UE) 2024/1689, article 50

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.

Article 50, paragraphe 2, du règlement (UE) 2024/1689 sur l'intelligence artificielle, version française sur EUR-Lex, consultée le 11 octobre 2026. Surlignage de SOURCE[1].

Le reste de l'article vise ceux qui utilisent ces outils, que la loi appelle les déployeurs. Celui qui publie un hypertrucage, une image, un son ou une vidéo qui imite une personne ou un événement réel, doit dire qu'il a été fabriqué ou manipulé. Celui qui publie un texte généré par IA pour informer le public sur des questions d'intérêt général doit le dire aussi, sauf si un humain l'a relu et en assume la responsabilité éditoriale[1]. Cette information doit être donnée de façon claire, au plus tard au premier contact avec le contenu[1]. Et le

non-respect de ces obligations peut coûter jusqu'à 15 millions d'euros ou 3 % du chiffre d'affaires mondial, le montant le plus élevé étant retenu[1]. Or une amende de cette taille, sur le marché européen, suffit à faire bouger une entreprise de San Francisco.

Un calendrier en trois temps

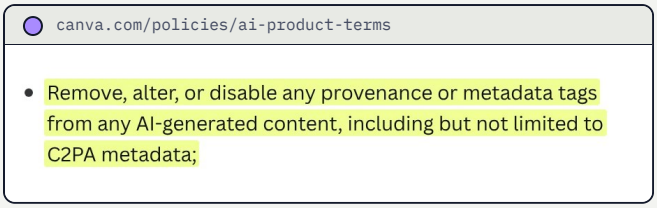
Le règlement s'applique pour l'essentiel depuis le 2 août 2026[1]. Mais à quelques jours de l'échéance, l'Europe a assoupli le calendrier. Un règlement dit « omnibus », signé le 8 juillet 2026 et publié au Journal officiel le 24 juillet sous le numéro 2026/1744, accorde un délai aux systèmes déjà sur le marché avant le 2 août : ils ont jusqu'au 2 décembre 2026 pour marquer leurs contenus. Les systèmes lancés après cette date, eux, doivent le faire tout de suite[2,3]. Les autres obligations de transparence de l'article 50 ne sont pas reportées[3]. C'est donc ce délai qui explique le « dans les semaines qui viennent » d'OpenAI, et le chantier qu'Anthropic décrit pour ses modèles plus anciens[17].

Entre-temps, la Commission a fait écrire un mode d'emploi. Le code de bonnes pratiques sur la transparence des contenus générés par l'IA a été rédigé par des experts indépendants, avec la participation de nombreuses organisations, des fabricants d'IA comme Google, OpenAI ou Meta aux associations de la société civile[6,8]. Publié le 10 juin 2026, il est volontaire[5]. La Commission et le Comité européen de l'IA, qui réunit les États membres, l'ont jugé adéquat : une entreprise qui le signe peut s'en servir pour prouver qu'elle respecte la loi, alors que les autres devront démontrer que leurs propres moyens suffisent[6]. Fin juillet, environ 190 organisations l'avaient signé[7]. Troisième date à retenir : les signataires doivent rendre leurs détecteurs interopérables d'ici au 2 février 2027[4].

Le code ajoute des détecteurs gratuits, et bannit les outils pour effacer

Le code va plus loin que la loi sur plusieurs points. On a vu qu'il exige deux couches, une étiquette signée et un filigrane, pour les images, les sons et les vidéos[4]. Il impose aussi de fournir un détecteur gratuit. Les autorités, les médias, les vérificateurs de faits, les chercheurs et les associations doivent y avoir accès sans limite de volume[4]. Pour le texte, en revanche, dont le filigrane est moins fiable, l'accès peut être réservé à des experts vérifiés, le temps que la technique progresse. C'est pourquoi OpenAI et Anthropic n'ouvrent leurs détecteurs de texte qu'aux chercheurs et organisations admis[4,11,17].

Surtout, le code s'attaque à l'effacement. Les signataires s'engagent à interdire, dans leurs conditions d'utilisation, de retirer ou de trafiquer les marques, et à ne pas vendre ni promouvoir d'outils conçus pour les contourner[4]. Canva l'écrit déjà noir sur blanc. Ses conditions pour les produits d'IA, en vigueur depuis le 26 juin 2026, interdisent de retirer, de modifier ou de désactiver les métadonnées de provenance des contenus générés, C2PA compris[22]. Mais le code ne punit pas l'utilisateur qui enlève une marque : il engage les entreprises qui le signent. La question « est-ce illégal d'enlever un filigrane IA ? » n'a donc pas de réponse simple, et on n'a trouvé aucune disposition du règlement qui sanctionne directement un particulier pour ce geste.



Conditions d'utilisation des produits d'IA de Canva, en vigueur depuis le 26 juin 2026, consultées le 11 octobre 2026. Parmi les usages interdits : retirer, modifier ou désactiver les métadonnées de provenance, y compris C2PA. Surlignage de SOURCE[22].

Pour les hypertrucages, le code propose enfin une marque visible commune : une icône européenne, que les déployeurs peuvent utiliser pour signaler un contenu artificiel, ou remplacer par une étiquette équivalente qui respecte les mêmes règles de dessin et de placement[4]. La Commission la met à disposition en trois versions.



Les icônes européennes pour signaler un contenu fait ou modifié par l'IA, publiées par la Commission avec le code de bonnes pratiques, en juin 2026. Leur usage est facultatif[4].

Icônes : Commission européenne, fichiers publiés sur digital-strategy.ec.europa.eu.

Ailleurs, la Chine exige du visible, la Californie suit l'Europe

L'Europe n'est pas seule, ni la première. En Chine, des mesures sur l'étiquetage des contenus générés par l'IA sont entrées en vigueur le 1er septembre 2025. Elles imposent aux fournisseurs une étiquette

visible, comme la mention « AI » en texte ou en exposant, qui doit rester dans les fichiers téléchargés ou exportés, et une étiquette cachée dans les métadonnées[10]. C'est pourquoi Google garde l'étoile de Gemini obligatoire dans les pays où la loi l'exige, Chine comprise[35].

En Californie, une loi sur la transparence de l'IA s'applique elle aussi depuis le 2 août 2026, après un report voté en 2025. Elle vise les systèmes d'IA générative qui comptent plus d'un million d'utilisateurs ou de visiteurs par mois : ils doivent proposer un outil de détection gratuit et inscrire dans leurs contenus une mention cachée, lisible par machine[9]. À partir du 1er janvier 2027, les grandes plateformes en ligne devront détecter les données de provenance des contenus qu'on y dépose, et à partir de 2028, les fabricants d'appareils photo vendus en Californie devront proposer d'en inscrire dans les photos elles-mêmes[9]. Le pays dont viennent la plupart de ces entreprises s'en est longtemps tenu aux promesses : en juillet 2023, sept d'entre elles s'étaient engagées auprès de la Maison-Blanche, sur une base volontaire, à développer des marques comme les filigranes[20].

Retour au 5 octobre : ce que la loi peut, et ce qu'elle ne peut pas

Revenons à l'annonce d'OpenAI. Un texte écrit avec ChatGPT par un utilisateur de Bruxelles portera bientôt le filigrane textGrain ; le même texte écrit à Boston, non, sauf si un développeur l'a activé[11,34]. Ce n'est pas une question de technique, puisque le filigrane existe et qu'OpenAI dit ne pas avoir mesuré de perte de qualité. C'est une question de loi : l'Europe l'exige, les États-Unis non[11,1].

Mais la loi ne règle pas tout. Elle oblige les fournisseurs à marquer, pas les filigranes à être incassables : elle demande des solutions aussi solides que la technologie le permet[1]. Elle ne dit rien, en revanche, de l'internaute qui fait reformuler un texte ou recompresser une image. Et elle ne verra pas non plus les modèles libres qu'on fait tourner sur son propre ordinateur, car une ligne de code y suffit à couper le filigrane[26]. Qui contrôlera, et comment ? Ce sont les autorités nationales de surveillance du marché, que le code mentionne, mais on n'a pas trouvé comment la Belgique comptait s'organiser[4]. On l'a rangé dans les angles morts.

CHRONOLOGIE

TROIS ANS DE PROMESSES

De l'engagement volontaire de 2023 aux échéances de 2027, les dates qui ont fait passer les filigranes de la promesse à l'obligation.

LE TEMPS DES PROMESSES

21 juillet 2023
■ OFFICIEL
Sept entreprises américaines, dont Google, Meta, Microsoft, OpenAI et Anthropic, promettent à la Maison-Blanche de développer des marques pour les contenus générés par l'IA[20].

Août 2023
▲ ÉTUDE
Google présente SynthID, son filigrane invisible pour les images[27].

6 février 2024
● DÉCLARATION
Meta annonce une mention visible, un filigrane invisible et des métadonnées sur les images photoréalistes de son assistant[18].

LES FAILLES SE VOIENT

Mars 2025
— PRESSE
Des internautes montrent que Gemini 2.0 Flash efface les filigranes des photos, dont ceux de Getty Images[36].

1er septembre 2025 OFFICIEL	La Chine impose une étiquette visible et une étiquette cachée sur les contenus générés par l'IA[10].
Octobre 2025 PRESSE	Le Washington Post dépose une vidéo Sora sur huit plateformes : aucune ne garde son étiquette[37].
Avril 2026 PRESSE	Un développeur publie reverse-SynthID ; Google conteste qu'il retire ses filigranes[39].

L'OBLIGATION

10 juin 2026 OFFICIEL	La Commission publie le code de bonnes pratiques sur la transparence des contenus générés par l'IA[5].
24 juillet 2026 OFFICIEL	Le règlement omnibus paraît au Journal officiel : délai jusqu'au 2 décembre pour les systèmes déjà sur le marché[2].
2 août 2026 OFFICIEL	L'article 50 du règlement sur l'IA s'applique, comme la loi californienne sur la transparence de l'IA[1,9].
14 août 2026 PRESSE	Google rend facultative l'étoile visible de Gemini[35].
5 octobre 2026 PRESSE	OpenAI annonce le filigrane textGrain pour les textes de ChatGPT et de Codex dans l'UE[34].

ET DEMAIN

2 décembre 2026 OFFICIEL	Fin du délai européen pour les systèmes lancés avant le 2 août[2].
2 février 2027 OFFICIEL	Les signataires du code européen doivent rendre leurs détecteurs interopérables[4].

ANGLES MORTS

CE QU'ON NE SAIT PAS ENCORE

Reste ce que nos lectures et nos tests ne tranchent pas. Voici les questions qu'on n'a pas pu fermer, et pourquoi.

La solidité réelle de SynthID. Les chiffres les plus flatteurs viennent de Google, qui teste son propre filigrane[27]. Les attaques des laboratoires visaient d'autres filigranes, et celle de reverse-SynthID n'a été vérifiée par personne d'indépendant à notre connaissance[38]. On n'a pas pu tester nous-mêmes SynthID à grande échelle, puisque le détecteur public de Gemini examine une image à la fois. Tant qu'aucune équipe extérieure n'aura publié de mesure, on ne peut dire ni « il résiste » ni « il est cassé ».

Ce que les réseaux affichent aujourd'hui. Le test du Washington Post date d'octobre 2025, avant la loi européenne[37]. On n'a trouvé aucune mesure plus récente de ce que Facebook, Instagram, TikTok ou X font des étiquettes C2PA en 2026.

Le filigrane des textes en français. OpenAI publie des taux de détection par langue, du plus haut, l'espagnol à 69,0 %, au plus bas, le roumain à 42,2 %, avec un taux de fausses alertes de 1 %[11]. Le chiffre du français n'apparaît pas dans le texte de la page que nos outils ont pu lire : il figure sans doute sur un graphique.

Qui contrôlera en Belgique. Le code renvoie aux autorités nationales de surveillance du marché[4]. On n'a pas trouvé quelle autorité belge sera chargée de contrôler l'article 50, ni avec quels moyens. C'est pourtant d'elle que dépendra, chez nous, la différence entre une règle écrite et une règle appliquée.

Les entreprises qu'on a mal lues. Les pages d'aide d'Adobe ont refusé notre lecteur automatique. Pour Meta, la documentation qu'on a pu lire date de 2024, et pour Microsoft, on n'a trouvé aucune mention d'un filigrane invisible, ni dans sa documentation d'Azure, ni ailleurs[18, 19]. Les cases « Pas trouvé » du comparatif peuvent cacher des pratiques réelles.

L'effet sur les gens. Une marque ne sert que si quelqu'un la cherche et la comprend. Le code européen demande d'ailleurs d'évaluer auprès des utilisateurs si les résultats de détection sont compréhensibles[4]. On n'a pas trouvé d'étude indépendante qui mesure si les étiquettes et les icônes changent vraiment ce que croient les lecteurs, ni s'ils les remarquent.

COMMENT CE NUMÉRO A ÉTÉ FAIT

1 Point de départ : les questions les plus tapées sur Google, Bing et DuckDuckGo et les discussions Reddit les plus votées sur « filigrane IA », « AI watermark », « SynthID », « C2PA » et « remove AI watermark », relevées le 11 octobre 2026. Chacune a sa réponse dans le numéro.

2 Six sources ont été lues en entier ou presque, parce que les méthodes et les chiffres y sont : l'étude de Maastricht (hors références), la section du code européen consacrée aux fournisseurs, la page d'aide d'OpenAI sur les signaux de provenance, celle d'Anthropic, l'introduction et les résultats de l'étude de Saberi et ses collègues, et le début du livre blanc de la C2PA. L'article de Google sur SynthID-Image a été lu par passages. Les autres sources ont été lues chiffre par chiffre.

3 Nos deux essais ont été faits le 11 octobre 2026. L'étiquette : lecture du fichier d'exemple C.jpg de la bibliothèque c2pa-python, puis réenregistrement avec Pillow. Le filigrane : bibliothèque invisible-watermark, méthode dwtDct, message de 64 bits, 12 photos de la NASA réduites à 1 024 pixels, seuil de 90 % de bits justes. Ce sont des essais de démonstration, pas une étude.

4 La double page de l'enquête montre la différence entre une photo marquée par nous et l'originale, amplifiée 40 fois. On n'a pas reproduit les attaques des chercheurs ni celle de reverse-SynthID, et on n'en décrit pas le mode opératoire.

FAIT AVEC UNE IA

Ce numéro a été préparé par Claude, l'IA d'Anthropic : recherche et lecture des sources, essais techniques, écriture, graphiques, captures et mise en page. Le sujet, le ton et les choix éditoriaux viennent de la rédaction. Anthropic est l'une des entreprises dont parle ce numéro : on a appliqué à sa documentation la même lecture qu'aux autres. Chaque chiffre a été relu dans le document original téléchargé, et un outil vérifie que chaque nombre suivi d'une note se retrouve dans le texte de sa source. Les scènes ne contiennent que ce que rapportent leurs sources.

LIMITE DE CE NUMÉRO

L'annonce d'OpenAI du 5 octobre est lue à travers BleepingComputer, le test du Washington Post à travers le billet de Fred Ritchin, et la réponse de Google à reverse-SynthID à travers Punto Informatico, qui cite The Verge. Les pages d'aide d'Adobe et le message de Josh Woodward sur X n'ont pas pu être lus directement. Les mesures chinoises sont lues à travers un cabinet d'avocats, Loeb & Loeb, et le règlement omnibus à travers deux cabinets, Cuatrecasas et Garrigues. Pour UnMarker et l'étude de Han et ses collègues, seul le résumé a été lu.

TEXTES OFFICIELS

1 Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 sur l'intelligence artificielle, articles 50, 99 et 113, EUR-Lex. eur-lex.europa.eu

2 Agustinoy A., « Digital Omnibus on AI has been published », Cuatrecasas, 24 juillet 2026 (règlement (UE) 2026/1744). cuatrecasas.com

3 « The AI Digital Omnibus Regulation has been published, redefining deadlines and simplifying requirements », Garrigues, 27 juillet 2026. garrigues.com

4 Commission européenne, *Code of Practice on Transparency of AI-Generated Content*, juin 2026 (PDF, 38 pages). ec.europa.eu

5 Commission européenne, « Commission publishes Code of Practice on marking and labelling AI-generated content », communiqué, 10 juin 2026. digital-strategy.ec.europa.eu

6 Commission européenne, « Code of Practice on Transparency of AI-generated Content », page de présentation, consultée le 11 octobre 2026. digital-strategy.ec.europa.eu

7 Commission européenne, « Strong backing for the Code of Practice on Transparency of AI-generated Content », 31 juillet 2026, avec la liste des signataires. digital-strategy.ec.europa.eu

8 Commission européenne, Bureau de l'IA, « Participant list: Code of Practice on Transparency of AI-generated Content », 8 décembre 2025 (PDF). ec.europa.eu

9 Californie, Assembly Bill 853, California AI Transparency Act, texte promulgué le 13 octobre 2025. leginfo.legislature.ca.gov

10 Sun J., « China's AI-Labeling Measures and Mandatory National Standards Take Effect September 1 », Loeb & Loeb, mars 2025 (où lire l'original : site de l'Administration chinoise du cyberespace, CAC, notice du 14 mars 2025). loeb.com

20 Maison-Blanche, « Fact sheet: Biden-Harris Administration secures voluntary commitments from leading artificial intelligence companies », 21 juillet 2023 (archives). bidenwhitehouse.archives.gov

■ NORMES

24 C2PA, « C2PA and Content Credentials Explainer », version 2.2. spec.c2pa.org

25 C2PA, *Use of Content Credentials to identify synthetic and non-synthetic content*, livre blanc, version 1.0, 30 juillet 2026 (PDF). c2pa.org

● DOCUMENTATION DES ENTREPRISES

11 OpenAI, « Signaux de provenance dans les contenus générés par OpenAI », centre d'aide, consulté le 11 octobre 2026. help.openai.com

12 OpenAI, « Launching Sora responsibly », septembre 2025. openai.com

13 OpenAI, « Understanding the source of what we see and hear online », mai 2024. openai.com

14 Google DeepMind, « Google SynthID », page de présentation, consultée le 11 octobre 2026. deepmind.google

15 Kohli P., « How we're bringing AI image verification to the Gemini app », blog de Google, 20 novembre 2025. blog.google

16 Google Cloud, « Informations Content Credentials », documentation, consultée le 11 octobre 2026. docs.cloud.google.com

17 Anthropic, « How Claude marks AI-generated content », centre d'aide, consulté le 11 octobre 2026. support.claude.com

18 Clegg N., « Labeling AI-Generated Images on Facebook, Instagram and Threads », Meta, 6 février 2024. about.fb.com

19 Microsoft, « Informations d'identification de contenu dans Azure OpenAI », Microsoft Learn, consulté le 11 octobre 2026. learn.microsoft.com

21 Adobe, « Adobe Content Authenticity », page de l'application, consultée le 11 octobre 2026. contentauthenticity.adobe.com

22 Canva, « AI Product Terms », en vigueur depuis le 26 juin 2026. canva.com

23 Stability AI, dépôt « generative-models », README (Invisible Watermark Detection) et fichier sgm/inference/helpers.py, consultés le 11 octobre 2026. github.com

38 Aloshdenny, « reverse-SynthID », dépôt GitHub, page d'accueil consultée le 11 octobre 2026 (affirmations de l'auteur, non vérifiées). github.com

▲ ÉTUDES

26 Rijsbosch B., van Dijck G., Kollnig K., « Missing the Mark: Adoption of Watermarking for Generative AI Systems in Practice and Implications under the new EU AI Act », arXiv:2503.18156, version 3, 8 octobre 2025 (texte intégral). arxiv.org (PDF)

27 Gowal S., Bunel R., Stimberg F. et al., « SynthID-Image: Image watermarking at internet scale », Google DeepMind, arXiv:2510.09263, 10 octobre 2025. arxiv.org (PDF)

28 Saberi M., Sadasivan V. S., Rezaei K. et al., « Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks », ICLR 2024, arXiv:2310.00076. arxiv.org (PDF)

29 Zhao X., Zhang K., Su Z. et al., « Invisible Image Watermarks Are Provably Removable Using Generative AI », NeurIPS 2024, arXiv:2306.01953 (résumé). arxiv.org

30 Kassis A., Hengartner U., « UnMarker: A Universal Attack on Defensive Image Watermarking », IEEE S&P 2025, arXiv:2405.08363 (résumé). arxiv.org

31 Dathathri S. et al., « Scalable watermarking for identifying large language model outputs », *Nature* 634, 23 octobre 2024. nature.com

32 Han X., Li Q., Ni J., Zulkernine M., « Robustness Assessment and Enhancement of Text Watermarking for Google's SynthID », TrustCom 2025, arXiv:2508.20228 (résumé). arxiv.org

— PRESSE

34 Parmar M., « OpenAI is adding invisible watermarks to ChatGPT and Codex text in the EU », BleepingComputer, 5 octobre 2026. bleepingcomputer.com

35 Nield D., « Gemini now lets you turn off the visible watermark on your AI creations », TechRadar, 16 août 2026 (cite le message de Josh Woodward du 14 août). techradar.com

36 Wiggers K., « People are using Google's new AI model to remove watermarks from images », TechCrunch, 17 mars 2025. techcrunch.com

37 Ritchin F., « Major platforms are not labeling AI-generated videos », 28 octobre 2025 (reprend l'enquête du Washington Post du 22 octobre 2025). substack.com

39 Foglio T., « Il watermark AI di Google è stato violato? Più o meno », Punto Informatico, 14 avril 2026 (cite The Verge). punto-informatico.it

40 AP et Reuters, « OpenAI pulls AI video app Sora as concerns grow on deepfake videos », Al Jazeera, 25 mars 2026. aljazeera.com

CRÉDITS PHOTO

33 Takkk, « German banknote watermark - 10 Mark.jpg », billet de 10 marks daté du 6 octobre 1906, CC BY-SA 3.0, recadrée. Wikimedia Commons

P2 NASA, mission STS-59, photo sts059-46-025, domaine public ; filigrane ajouté et différence amplifiée par SOURCE. Photos du test : 12 images de la médiathèque de la NASA (liste dans les notes du numéro). images.nasa.gov

I Icônes « AI », « AI GENERATED » et « AI MODIFIED » : Commission européenne, publiées avec le code de bonnes pratiques. digital-strategy.ec.europa.eu

G Graphiques, schéma, encadré du manifeste et captures surlignées : SOURCE.