

## La blockchain : protection du droit d'auteur à l'ère de l'IA générative

### ➤ Résumé

L'intelligence artificielle générative repose sur l'exploitation à grande échelle d'œuvres protégées par le droit d'auteur. Les principaux fournisseurs d'IA ont entraîné leurs modèles sur des millions d'œuvres (textes et images) souvent sans autorisation, entraînant l'ouverture de nombreux contentieux.

Le cadre européen (TDM et IA Act) instaure un *opt-out* pour les détenteurs de droit. Ce dispositif se heurte toutefois à plusieurs limites : opacité persistante des corpus d'entraînement, non-respect généralisé des directives et absence de moyens de contrôle pour les créateurs.

Dans ce contexte, la technologie blockchain apporte une réponse technique complémentaire en permettant l'horodatage certifié des créations, la preuve d'antériorité et la traçabilité des usages. Contrairement aux mécanismes actuels, un registre blockchain permettrait ainsi aux créateurs de tracer l'utilisation de leurs œuvres et aux fournisseurs d'IA de démontrer leur conformité.

*Cet article de l'Adan analyse comment la blockchain peut contribuer à concilier innovation technologique et protection du droit d'auteur.*

### ➤ Sommaire

<b>1. Contexte et définition.....</b>	<b>2</b>
<b>2. Les contentieux entre artistes et fournisseurs de modèles d'IA.....</b>	<b>5</b>
<b>2.1. Affaires aux États-Unis (L'Authors Guild v. OpenAI, Bartz v. Anthropic, Kadrey v. Meta).....</b>	<b>5</b>
❖ Authors Guild v. OpenAI.....	6
❖ Bartz v. Anthropic.....	7
❖ Kadrey v. Meta.....	8
<b>2.2. Affaires Européennes (Syndicats français du livre v. Meta, Like Company v. Google Ireland).....</b>	<b>8</b>
❖ SNE / SGDL / SNAC v. Meta.....	8
❖ Like Company v. Google Ireland.....	9
<b>3. Les réponses juridiques européennes (TDM et IA act).....</b>	<b>9</b>
3.1. L'exception de fouille de textes et de données (TDM).....	9
3.2. Règlement européen sur l'intelligence artificielle (AI Act).....	10
<b>4. Blockchain : registre de confiance pour les acteurs de l'IA, opt-out pour les ayants droits.....</b>	<b>11</b>
<b>5. Conclusion.....</b>	<b>14</b>
<b>Annexe - Sources.....</b>	<b>16</b>

## 1. Contexte et définition

L'intelligence artificielle générative a connu une adoption sans précédent depuis le lancement de ChatGPT par OpenAI en novembre 2022, atteignant plus de 100 millions d'utilisateurs en deux mois. Cette démocratisation s'appuie sur plusieurs décennies de recherches scientifiques, et plus particulièrement sur la conjonction de plusieurs facteurs et percées technologiques décisives dont notamment la découverte en 2017 par Google de l'architecture Transformer, l'explosion de la masse de données disponibles en ligne (big data) et la croissance exponentielle des capacités de calcul dédiées (GPU/TPU).

Ces avancées ont rendu possible le développement d'une diversité de modèles génératifs, dont les capacités et usages varient selon la nature des données d'entraînement et le type de contenu produit (texte, image, vidéo, audio).

Type de modèle	Capacités principales	Exemples
Modèles de langage LLM - Large Language Models	Rédiger, traduire, résumer, répondre à des questions, générer du code. Grande fluidité et cohérence dans la production textuelle.	GPT, Claude, LLaMA, Gemini
Modèles de génération d'images T2I - Text-to-Image	Transformer une description en image réaliste ou artistique. Création d'illustrations allant du croquis au photoréalisme.	DALL·E, Stable Diffusion, Midjourney
Modèles de génération audio T2A - Text-to-Audio, T2M - Text-to-Music	Générer des sons, voix ou musiques à partir de texte ou d'exemples sonores. Imiter des instruments, créer des compositions inédites, synthétiser des voix naturelles.	Jukebox (OpenAI), Suno, MusicLM
Modèles de génération vidéo T2V - Text-to-Video	Produire des séquences vidéo à partir de texte ou d'images. Combiner génération d'images et animation temporelle pour créer des scènes cohérentes.	Sora (OpenAI), Runway Gen-2, Pika Labs
Modèles multimodaux VLM MLLM	Traiter simultanément plusieurs types de données (texte, image, son, vidéo). Analyser une image et la décrire, combiner texte + image pour des réponses contextuelles.	Gemini, GPT-4 (vision), LLaVA

Modèles de génération 3D T23D – Text-to-3D	Créer des objets ou environnements 3D à partir d'instructions textuelles. Modéliser formes, textures et volumes pour design, jeux vidéo, VR.	Point-E, Shap-E, GET3D
---	--	------------------------

Ces modèles génératifs suivent, pour la plupart, un processus d'entraînement comparable, lequel consiste à apprendre la sémantique, les styles et les connaissances présentes dans de vastes ensembles de données (œuvres du domaine public, sous licences ouvertes ou protégées par le droit d'auteur). Ils permettent ensuite de générer à la demande de nouveaux contenus, en s'inspirant, voire en reproduisant, certains styles, structures ou connaissances présents dans leur corpus d'entraînement.

La qualification "d'originalité" de ces contenus générés demeure sujette à débat, notamment au regard du droit d'auteur. Dans le domaine de l'intelligence artificielle, les données constituent le principal matériau d'entraînement des modèles. Souvent difficiles à identifier, collecter et qualifier juridiquement, elles représentent une ressource stratégique pour les entreprises du secteur. Pour acquérir des données de haute qualité et exploitables, les principaux acteurs s'appuient sur des ensembles de données (datasets) issus d'un mélange parfois peu transparent de sources publiques, privées ou partenariales.

Plusieurs analyses ont montré que ces corpus peuvent inclure, dans certains cas, des œuvres protégées sans autorisation expresse, soulevant ainsi des enjeux majeurs en matière de respect du droit d'auteur et de rémunération des ayants droit.

#### ❖ **Les données d'entraînement : un manque de transparence**

Pour illustrer cette problématique, prenons l'exemple du modèle GPT-3. En 2020, les chercheurs d'OpenAI ont publié un article scientifique intitulé "*Language Models are Few-Shot Learners*". Ils y démontrent empiriquement qu'en augmentant considérablement la quantité de données d'entraînement ingérées par les modèles de langage, ceux-ci acquièrent la capacité d'exécuter de nouvelles tâches sans réentraînement spécifique à partir de seulement quelques exemples fournis par l'utilisateur (principe du *few-shot learning*).

L'article décrit la composition du corpus d'entraînement ayant servi à l'élaboration de GPT-3 :

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

**Figure 1** : Répartition des corpus utilisés pour l'entraînement de GPT-3.

**Source** : OpenAI.

Comme le montre la **Figure 1**, une large part du corpus d'entraînement de GPT-3 provient de milliards de pages web rendues accessibles en ligne, principalement via des collectes massives (*scraping*<sup>1</sup>).

Common Crawl en est un parfait exemple. Ce projet à but non lucratif, actif depuis 2008, archive en continu une grande partie du web mondial et représente aujourd'hui près d'une dizaine de pétaoctets de données. On y trouve des contenus très variés dont des articles de presse, des billets de blog, des pages de produits en e-commerce, des documents institutionnels, des forums de discussion et même certaines publications issues de réseaux sociaux.

Pour WebText2, les chercheurs d'OpenAI ont constitué ce corpus en collectant les pages web les plus citées et partagées sur Reddit, considérant que ce signal social constituait un proxy de qualité rédactionnelle et de pertinence thématique.

Plus controversés, les corpus Books1 et Books2 regroupent des ensembles de livres numériques utilisés pour l'entraînement de GPT-3. Leur finalité était d'apporter des textes longs et mieux structurés, de renforcer la cohérence contextuelle et d'enrichir la diversité stylistique. Représentant environ 350 000 ouvrages, ces corpus illustrent le manque de transparence de l'industrie. OpenAI n'ayant jamais divulgué la composition exacte de ces corpus, il est difficile pour les ayants droits d'en connaître le contenu précis.

Des chercheurs indépendants ont tenté de percer le mystère autour de la composition de ces *datasets*. Différentes techniques permettent de déceler des indices d'entraînement à partir d'œuvres protégées, notamment des attaques d'extraction.

- Cette méthode consiste à interroger méthodiquement un modèle au sujet d'un texte protégé en lui soumettant, par exemple, un passage incomplet (texte à trous)

<sup>1</sup> Technique d'extraction automatisée de données d'un site web, réalisée par des robots (*bots*) qui parcourent les pages, lisent leur contenu et collectent les informations ciblées pour les réutiliser.

ou une amorce (fin ouverte), on observe si le modèle est capable de compléter fidèlement le contenu.

Si le modèle reproduit des fragments exacts ou des passages caractéristiques d'une œuvre protégée, cela suggère fortement que le texte figurait dans son corpus d'entraînement<sup>2</sup>. D'autres techniques d'audit plus avancées existent comme la détection d'appartenance ou les tests de contamination.

Les recoupements académiques et journalistiques étayent l'hypothèse la plus répandue :

- **Books1** correspondrait à des œuvres du domaine public (Project Gutenberg)
- **Books2** correspondrait à des œuvres protégées issues de bibliothèques pirates (LibGen/Bibliotik/Z-Library)

Par ailleurs, plusieurs contentieux sont actuellement en cours aux États-Unis comme en Europe. Ils contribuent à l'émergence d'une jurisprudence concernant l'utilisation d'œuvres protégées à des fins d'entraînement de modèles, en cherchant à évaluer le préjudice économique subi par les ayants droit et à définir les modalités d'une rémunération équitable.

## 2. Les contentieux entre artistes et fournisseurs de modèles d'IA

### 2.1. Affaires aux États-Unis (L'Authors Guild v. OpenAI, Bartz v. Anthropic, Kadrey v. Meta)

Les craintes des auteurs sont devenues vives devant l'habileté de ChatGPT non seulement à imiter leur style mais également de reproduire partiellement certains extraits et de proposer des suites fictives de leurs œuvres, sur simple demande des utilisateurs.

Ces inquiétudes se sont traduites par plusieurs actions en justice intentées par des groupements d'auteurs à l'encontre des principaux fournisseurs de modèles d'IA. L'issue de ces procédures pourrait s'avérer déterminante.

 Le *fair use*, principale ligne de défense des entreprises technologiques

Les entreprises du secteur soutiennent que l'utilisation d'œuvres protégées à des fins d'entraînement de modèles relève du *fair use* (usage équitable). Il s'agit d'une doctrine juridique américaine (*Copyright Act*) qui autorise sous certaines conditions l'utilisation d'œuvres protégées sans l'autorisation préalable des ayants droit.

---

<sup>2</sup> Les chatbots conversationnels (ChatGPT, Claude) refusent désormais de compléter des textes à trous issus de corpus protégés, ce qui limite l'efficacité de ce type d'attaques.

L'exception du *fair use* repose sur l'examen, au cas par cas, de quatre facteurs :

1. **Utilisation transformative** : emploi de l'œuvre à des fins nouvelles, différentes et non concurrentielles.
2. **Nature de l'œuvre** : les œuvres fictionnelles ou non publiées bénéficient d'une protection renforcée.
3. **Quantité utilisée** : l'utilisation de l'œuvre dans son intégralité est moins susceptible d'être considérée comme équitable.
4. **Effet sur le marché** : l'usage ne doit pas porter préjudice au marché existant ou potentiel de l'œuvre originale.

Les entreprises défendent l'idée que l'entraînement constitue un usage transformatif, dans la mesure où il ne reproduit pas directement les œuvres. Elles invoquent également l'existence d'un bénéfice sociétal, notamment pour la recherche scientifique. Enfin, elles estiment que les modèles génératifs n'entrent pas en concurrence directe avec les artistes, mais qu'ils ouvrent de nouveaux marchés en leur offrant des moyens supplémentaires d'expression.

Dans un domaine où l'entraînement des LLM exige déjà des investissements en capital colossaux, l'industrie de la tech redoute une remise en cause de ses pratiques de collecte. Les données exploitables juridiquement étant rares et extrêmement coûteuses, cette remise en question pourrait engendrer des centaines de millions de dollars supplémentaires de dépense. À cela s'ajoute la menace de potentielles amendes pouvant atteindre plusieurs milliards de dollars.

#### ❖ **Authors Guild v. OpenAI**

L'affaire *Authors Guild v. OpenAI* est la plus emblématique. Cette action collective fédérale a été déposée le 19 septembre 2023 devant le United States District Court for the Southern District of New York (SDNY). Elle oppose la plus ancienne organisation professionnelle d'écrivains américains, l'Authors Guild (comptant plus de 14 000 membres), ainsi que 17 auteurs de renom<sup>3</sup>, à OpenAI et Microsoft<sup>4</sup>, en raison de leur partenariat stratégique et d'infrastructure.

Les plaignants reprochent à OpenAI d'avoir reproduit et intégré des œuvres protégées sans autorisation ni compensation, en les collectant à partir de sites pirates, avant de les exploiter pour l'entraînement de ses modèles de langage GPT-3, GPT-3.5 et GPT-4. Ils affirment qu' "au cœur de ces algorithmes se trouve un vol systématique à grande échelle".

---

<sup>3</sup> David Baldacci ; Mary Bly ; Michael Connelly ; Sylvia Day ; Jonathan Franzen ; John Grisham ; Elin Hilderbrand ; Christina Baker Kline ; Maya Shanbhag Lang ; Victor LaValle ; George R.R. Martin ; Jodi Picoult ; Douglas Preston ; Roxana Robinson ; George Saunders ; Scott Turow ; Rachel Vail

<sup>4</sup> Cette affaire fut consolidée avec l'action *Alter v. OpenAI*, créant une mega class action.

Pour les auteurs, les systèmes comme ChatGPT menacent le marché du livre puisqu'ils sont capables de générer, à la demande des utilisateurs, de nouveaux textes reprenant le style des auteurs. Dès lors, les consommateurs risqueraient de se détourner des œuvres originales.

L'affaire demeure pendante. Des documents judiciaires rendus publics révèlent qu'OpenAI aurait cessé d'utiliser, puis supprimé à la mi-2022, les corpus au cœur du litige<sup>5</sup> (*Books1* et *Books2*). Les chercheurs ayant conçu ces jeux de données ne travaillent désormais plus au sein de l'entreprise, qui refuse d'en divulguer publiquement les identités.

Dans d'autres affaires portées devant des tribunaux fédéraux américains, des groupes d'auteurs ont également poursuivi des entreprises concurrentes d'OpenAI, telles qu'Anthropic et Meta, et des jugements ont déjà été rendus. Les jugements rendus jusqu'à présent apparaissent plutôt contrastés. Ils ne valident ni pleinement la défense fondée sur le *fair use*, ni les thèses de contrefaçon ou de concurrence déloyale.

Dans certains cas, les auteurs obtiennent partiellement gain de cause. Ce fut notamment le cas dans l'affaire *Bartz v. Anthropic*. La justice ayant rendu une décision délimitant clairement la portée du *fair use* appliqué à l'entraînement de modèles.

#### ❖ **Bartz v. Anthropic**

Pour entraîner ses modèles Claude, Anthropic a recouru à diverses méthodes dont le téléchargement d'environ 7 millions d'ouvrages obtenus illégalement via des bibliothèques en ligne non autorisées, des achats puis numérisation d'œuvres protégées, ainsi que la constitution d'une bibliothèque interne décrite dans les documents judiciaires comme destinée à contenir "*tous les livres du monde*" et à les conserver "*pour toujours*".

Le 23 juin 2025, le juge dans cette affaire a estimé que l'entraînement d'un LLM à partir d'œuvres protégées relève bien du *fair use*, qualifiant cette pratique d'"*exceptionnellement transformative*" et "*parmi les plus transformatives que beaucoup d'entre nous verront dans leur vie*". En revanche, la collecte de livres obtenus illégalement constitue une violation du Copyright Act, indépendamment de leur utilisation ultérieure. En somme, cette décision souligne l'importance d'un accord préalable et d'une juste rémunération des auteurs, comme une condition sine qua non à l'utilisation d'œuvres protégées à des fins d'entraînement.

Anthropic a été jugée de violation du droit d'auteur. Face au risque de plusieurs milliards de dollars de dommages-intérêts à reverser aux ayants droit (jusqu'à 150 000 dollars par ouvrage selon le Copyright Act), un règlement transactionnel a été conclu. Anthropic

---

<sup>5</sup> Sans être nommé par les plaignants.

versera 1,5 milliard de dollars, soit environ 3 000 dollars par livre. Il s'agit du plus important règlement pour violation du droit d'auteur jamais enregistré aux États-Unis.

#### ❖ **Kadrey v. Meta**

Dans l'affaire Kadrey v. Meta, intentée par treize auteurs, la défense fondée sur le *fair use* a été retenue. Les plaignants soutenaient que Meta se servait de corpus contenant des œuvres protégées pour entraîner son modèle LLaMA, et que cette utilisation constituait une menace pour leur marché. Meta, à l'instar de ses concurrents, a invoqué le caractère équitable et transformatif de cet usage.

Les plaignants, n'ayant pas apporté d'éléments probants démontrant un préjudice de marché significatif, ont été déboutés. Néanmoins, la question d'une éventuelle violation du droit d'auteur demeure en suspens.

En reconnaissant l'utilisation ainsi que l'obtention du corpus *Books3* via un système de torrent, Meta aurait pu en redistribuer des fragments au sein du réseau pair-à-pair. La question centrale des prochains débats sera donc de déterminer si Meta a, ou non, partagé à son tour des parties de *Books3* avec d'autres utilisateurs du réseau torrent.

**Ces affaires mettent en lumière une certaine tendance jurisprudentielle qui distingue l'usage d'œuvres protégées à des fins d'entraînement de la manière dont ces œuvres sont obtenues. Finalement, la jurisprudence émergente suggère que les principaux fournisseurs d'IA générative pourraient exploiter des œuvres protégées, à condition de prévoir une rémunération équitable des ayants droit, proportionnée à la contribution de ces œuvres à la performance des modèles.**

## 2.2. **Affaires Européennes (Syndicats français du livre v. Meta, Like Company v. Google Ireland)**

En Europe, les inquiétudes des ayants droits et des artistes s'intensifient. Les contentieux relatifs à l'usage d'œuvres protégées se multiplient parallèlement aux actions intentées aux États-Unis.

#### ❖ **SNE / SGDL / SNAC v. Meta**

Le 6 mars 2025, une coalition réunissant les trois principaux syndicats de l'industrie littéraire française - le Syndicat national de l'édition (SNE), la Société des gens de lettres (SGDL) et le Syndicat national des auteurs et des compositeurs (SNAC) - a assigné Meta en justice.

Au cœur des allégations, l'utilisation massive et non autorisée d'œuvres protégées, dont plusieurs œuvres françaises, provenant du corpus *Books3*, dont Meta a reconnu avoir exploité dans le cadre de l'affaire Kadrey v. Meta.

À l'instar de leurs homologues américains, les syndicats demandent la reconnaissance d'une violation du droit d'auteur et d'actes de parasitisme économique, ainsi qu'une compensation financière au bénéfice des auteurs concernés. L'affaire suit actuellement son cours et fera, sans aucun doute, jurisprudence en France.

### ❖ **Like Company v. Google Ireland**

Dans une autre affaire européenne majeure, le 3 avril 2025, le Tribunal de Budapest a renvoyé devant la Cour de justice de l'Union européenne (CJUE) l'affaire Like Company v. Google Ireland, marquant ainsi le premier dossier soumis à la plus haute juridiction européenne relatif à l'intelligence artificielle et au droit d'auteur. Like Company, éditeur de presse hongrois, reproche à Google d'avoir reproduit - via Gemini - et mis à disposition du public des portions substantielles de ses articles de presse, dépassant ainsi les "extraits très courts" autorisés par l'article 15 de la directive sur le droit d'auteur dans le marché unique numérique.

La juridiction hongroise a saisi la CJUE de quatre questions préjudicielles portant notamment sur la légalité de l'entraînement de modèles d'IA sur des contenus protégés, sur l'applicabilité de l'exception de fouille de textes et de données (TDM), ainsi que sur la qualification juridique, au regard du droit d'auteur, des réponses générées par les chatbots.

Google conteste ces allégations, arguant que Gemini ne stocke pas de copies d'articles mais génère du nouveau contenu à partir de prédictions probabilistes issues de son entraînement. L'entreprise affirme que toute similarité constatée résulte d'un processus d'apprentissage légitime et couvert par les exceptions légales ou soit le résultat de simples hallucinations du modèle.

La décision de la CJUE, attendue en 2026, établira un arrêt de principe pour l'ensemble de l'Union européenne et déterminera l'équilibre entre protection du droit d'auteur et innovation en matière d'IA générative.

## **3. Les réponses juridiques européennes (TDM et IA act)**

### **3.1. L'exception de fouille de textes et de données (TDM)**

À l'ère du numérique, la protection des droits d'auteur doit être encadrée par des instruments juridiques clairs et transparents. En ce sens, l'Union européenne a développé une réglementation scindée en deux textes : le *Text and Data Mining* (TDM) et l'*IA Act*.

Adoptée en 2019, la directive *Copyright in the Digital Single Market* (DSM) introduit deux exceptions de TDM (articles 3 et 4). L'article 3 autorise les organismes de recherche et de

conservation à effectuer de la fouille de textes et de données sur des œuvres licitement accessibles, sans possibilité d'opposition des ayants droit. L'article 4, plus large, permet à toute personne physique ou morale d'y recourir, sous réserve que les ayants droit n'aient pas exprimé leur refus par un mécanisme d'*opt-out*.

### **Les limites de l'*opt-out* à l'ère de l'IA générative**

Ce mécanisme d'*opt-out* permet aux ayants droit de signaler explicitement leur refus d'extraction ou de reproduction de leurs œuvres à des fins de TDM. Ce refus doit être exprimé par des moyens techniques lisibles par machine, tels que des métadonnées, conditions générales d'utilisation.

Ce mécanisme présente plusieurs lacunes structurelles, dont notamment :

- **Inversion de la charge** : les ayants droit supportent l'intégralité de la charge de protection : ils doivent repérer l'utilisation de leurs œuvres et exprimer leur refus au moyen d'un procédé technique, souvent méconnu.
- **Non-respect par les collecteurs** : plusieurs études ont montré que les robots collecteurs de données respectent sélectivement les directives robots.txt, certains les ignorant totalement lors de la constitution de corpus d'entraînement.
- **Absence de contrôle effectif** : en l'absence de transparence obligatoire sur les données réellement utilisées, le mécanisme d'*opt-out* reste largement inefficace.

Comme le souligne David El Sayegh, directeur général adjoint de la Sacem, il existe "*une absence totale de transparence sur les données utilisées pour entraîner les intelligences artificielles*", rendant "*presque impossible la connaissance des œuvres utilisées*".

## **3.2. Règlement européen sur l'intelligence artificielle (AI Act)**

Le Règlement européen sur l'intelligence artificielle (AI Act), entré en vigueur le 1er août 2024, constitue le premier cadre réglementaire complet au monde en la matière. Les modèles de langage (GPT-5, Claude Opus ou encore Stable Diffusion) y sont classés parmi les "*modèles d'IA à usage général*" (General-Purpose AI, GPAI).

Positionnant l'Europe comme pionnière dans l'instauration d'un cadre pour une IA éthique, l'IA act établit un dispositif garantissant la sûreté, la gouvernance et la transparence des systèmes d'intelligence artificielle mis sur le marché européen. Toute entreprise proposant de tels systèmes au sein de l'UE est soumise à ce règlement. En cas d'infraction constatée, des amendes graduées sont prévues : selon la gravité et le type d'IA concerné, elles peuvent atteindre jusqu'à 7 % du chiffre d'affaires annuel mondial de l'entreprise.

Les obligations pour les fournisseurs de GPAI comprennent plusieurs exigences :

- Constituer et maintenir une documentation technique complète retraçant les caractéristiques, usages et limites du modèle.
- Publier un résumé standardisé des données d'entraînement, selon le format défini par l'AI Office, l'autorité de supervision créée par le règlement.
- Adopter une politique claire de conformité au droit d'auteur, incluant des mécanismes de détection et le respect effectif des opt-out.

De plus, un code de bonnes pratiques en matière d'IA responsable a été publié le 10 juillet 2025 par la Commission européenne, élaboré en collaboration avec les fournisseurs, chercheurs et représentants de la société civile. Ce document reprend les principes de l'AI Act et en facilite la mise en œuvre à travers des méthodes et outils concrets.

Non contraignant, il encourage fortement les acteurs de l'industrie à y apposer leur signature. Les non-signataires, tels que Meta ou xAI, devront fournir des preuves supplémentaires de conformité et s'attendre à une surveillance renforcée.

**L'AI Act fait l'objet de vives critiques de par les ayants droit et de certains décideurs politiques. En renvoyant à l'exception TDM pour l'acquisition de données d'entraînement, il pénalise les créateurs et favorise les intérêts commerciaux des grands fournisseurs de GPAI. Les artistes n'ont aujourd'hui aucun moyen concret de savoir si leurs œuvres sont utilisées à des fins d'entraînement. La seule obligation imposée aux fournisseurs consiste à publier un résumé standardisé des types d'œuvres exploitées et de leur répartition, document jugé trop imprécis par les représentants des créateurs..**

Comme le montrent les études, le mécanisme d'opt-out est imparfaitement respecté dans sa mise en œuvre actuelle. De ce fait, les ayants droit se trouvent aujourd'hui dans l'incapacité, tant technique que juridique, de défendre efficacement leurs droits. Afin de concilier protection des créateurs, innovation et compétitivité européenne, plusieurs pistes pourraient être envisagées : remplacer l'opt-out par un véritable opt-in, instaurer à l'échelle de l'UE un registre basé sur la technologie blockchain permettant aux créateurs de contrôler l'usage de leurs œuvres.

#### **4. Blockchain : registre de confiance pour les acteurs de l'IA, opt-out pour les ayants droits**

La technologie blockchain apporte aux ayants droits une solution technique supplémentaire complémentaire au droit positif. En tant que registre électronique distribué, garantissant l'intégrité temporelle des informations qui y sont inscrites, elle constitue un système d'horodatage et de conservation des preuves infalsifiable.

Lors de l'ancrage au sein du registre, une date et une heure précises sont associées à l'empreinte numérique<sup>6</sup> de l'œuvre (hash), garantissant l'existence d'un contenu à un instant T. En droit français et européen, un tel horodatage peut constituer une preuve d'antériorité et d'existence de l'œuvre. Toutefois, il revient au juge d'en apprécier la valeur, au même titre que d'autres moyens de preuve (règlement eIDAS).

Contrairement à l'enveloppe Soleau, l'ancrage sur blockchain est moins coûteux, ne nécessite aucun envoi physique et permet une conservation *ad vitam aeternam*. De plus, la blockchain n'impose pas de limite de format de fichier : un développeur peut y déposer le hash de son code source, un musicien le hash d'un fichier audio.

**Attention** : cet horodatage ne crée pas le droit d'auteur - qui naît automatiquement du seul fait de la création (article L111-1 du Code de la propriété intellectuelle) - et ne prouve pas non plus l'identité du créateur.

💡 En France, la jurisprudence a franchi une étape importante avec l'*affaire AZ Factory c. Valeria Moda*, jugée en mars 2025 par le Tribunal judiciaire de Marseille. Pour la première fois, des horodatages blockchain ont été admis comme preuve recevable afin d'établir l'antériorité et la titularité de droits d'auteur sur des croquis de mode.

Le juge a considéré ces constats comme des éléments probants, sans leur accorder une valeur absolue, mais en les intégrant à un faisceau d'indices comprenant également la divulgation publique et la notoriété de l'auteur. Cette reconnaissance ouvre la voie à une utilisation plus fréquente de la blockchain comme outil de preuve dans les litiges français relatifs au droit d'auteur, confirmant que la technologie peut compléter efficacement les méthodes probatoires traditionnelles.

Face aux limites structurelles du système d'opt-out actuel, plusieurs études institutionnelles plaident pour la **création d'un registre européen unifié basé sur la blockchain**. Une étude commandée par le Parlement européen en mai 2025 sur l'IA générative et le droit d'auteur propose explicitement *"la mise en place d'un registre unifié et lisible par machine, potentiellement supervisé par l'EUIPO, permettant aux titulaires de droits d'exprimer leurs préférences en matière d'utilisation de leurs œuvres à des fins d'entraînement"*.

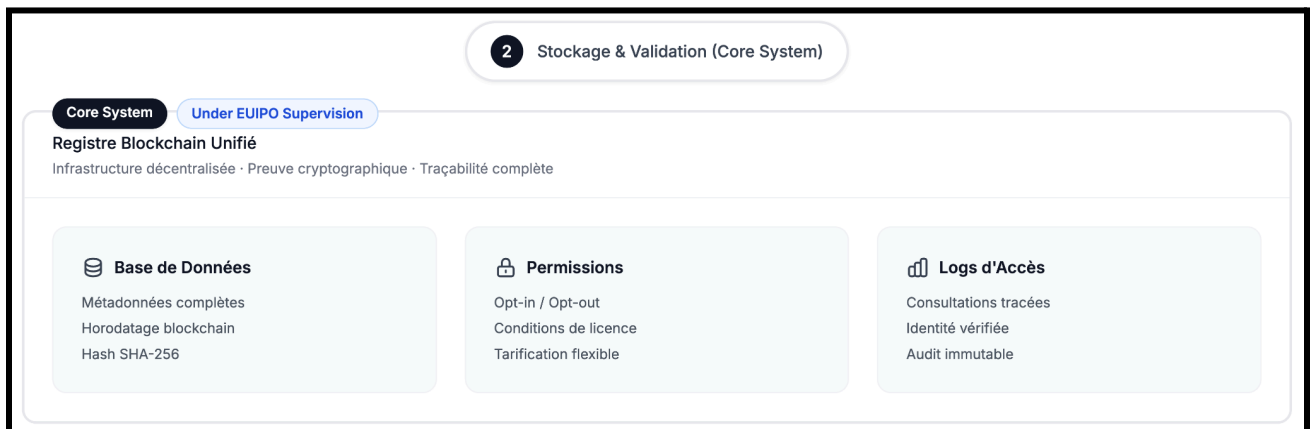
- Ce registre fonctionnerait sur un modèle de *"permissions centralisées"*. Les créateurs enregistreraient leurs œuvres et leurs préférences d'utilisation (autorisation totale, interdiction totale, ou licence conditionnelle) sur une plateforme blockchain administrée par une autorité publique de confiance. Les fournisseurs de modèles d'IA seraient tenus de consulter ce registre avant d'utiliser toute œuvre protégée, avec des mécanismes de vérification automatisés

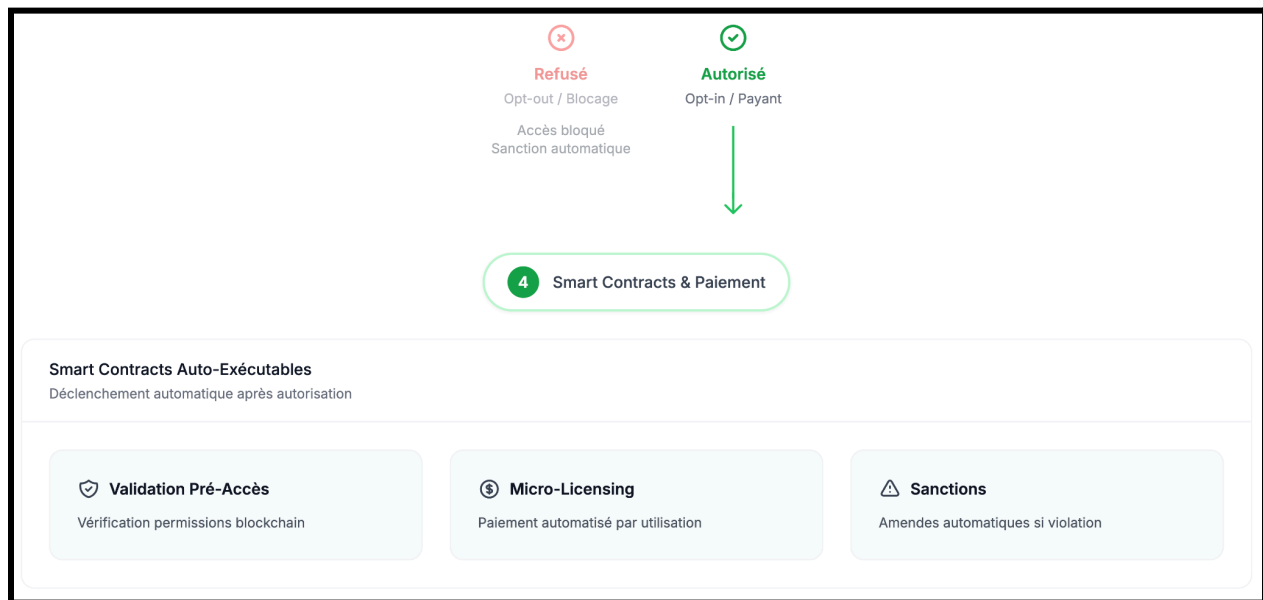
---

<sup>6</sup> Empreinte numérique calculée par une fonction cryptographique : permet d'identifier de façon unique et fiable un fichier.

et des sanctions en cas de non-respect. L'horodatage blockchain garantirait la preuve de chaque déclaration d'opt-out ainsi que de chaque consultation par un acteur de l'IA.

- Ce dernier pourrait également intégrer des fonctionnalités de *micro-licensing* automatisées *via* des *smart contracts* (contrats intelligents auto-exécutables). Un créateur pourrait, par exemple, autoriser l'utilisation de ses œuvres moyennant une rémunération fixe par utilisation, le smart contract déclenchant automatiquement le paiement dès qu'un fournisseur d'IA accède à l'œuvre dans le cadre de son entraînement. Cette approche résoudrait partiellement la question de la juste rémunération des artistes tout en simplifiant les processus de licensing.





**Figure 2 :** Schéma d'implémentation d'un possible registre blockchain d'opt-out

## 5. Conclusion

Chaque semaine, de nouveaux modèles de génération de textes, d'images, de vidéos et de sons sont mis sur le marché. Une industrie de l'intelligence artificielle, valorisée à plusieurs milliards de dollars, s'est construite - et continue de se développer - sur l'exploitation d'un volume considérable d'œuvres protégées.

Face à des coûts d'infrastructure élevés, la collecte de données et la conquête d'utilisateurs sont devenues des priorités pour les entreprises de l'IA. L'usage d'œuvres protégées sans autorisation alimente un modèle économique qui fragilise la création artistique et réduit les revenus des créateurs.

La sortie récente de l'application de génération et de partage de vidéos Sora2 illustre cette évolution. Comme GPT-3 en son temps, la reproduction d'œuvres protégées est désormais accessible à un large public, sans transparence ni garde-fous suffisants. Cette situation expose aussi bien les auteurs indépendants que les grandes entreprises culturelles, dans un contexte économique où la valeur se déplace vers les plateformes d'IA.

Cette situation révèle un déséquilibre profond entre les grandes plateformes d'IA et les titulaires de droits qui ne disposent d'aucun moyen de contrôle effectif. Le manque de transparence sur les données d'entraînement reste l'un des points faibles du cadre européen, que l'AI Act ne corrige que partiellement.

Dans l'attente de son application complète, il est nécessaire de renforcer les outils techniques et réglementaires permettant aux ayants droit d'exprimer leurs conditions d'utilisation, en s'appuyant sur des balises standardisées, des registres fondés sur la blockchain et des protocoles de traçabilité.

Ces initiatives, soutenues au niveau européen, pourraient devenir un standard international à l'image du RGPD.

## Annexe - Sources

### Sites officiels :

Union européenne. (2024, 12 juillet). *Règlement (UE) 2024/1689 du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (AI Act)*. Disponible à : [https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=OJ:L_202401689)

Union européenne / AI Act Explorer. (s.d.). *Article 99 – Pénalités*. Disponible à : <https://artificialintelligenceact.eu/article/99/>

Commission européenne (DG CNECT). (2025, 24 juillet). *Explanatory Notice and Template for the Public Summary of Training Content for GPAI*. Disponible à : <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models>

Union européenne. (2019, 17 mai). *Directive (UE) 2019/790 (DSM)*. Disponible à : <https://eur-lex.europa.eu/eli/dir/2019/790/oj>

Union européenne. (2014). *Règlement (UE) n° 910/2014 (eIDAS) – Article 41 : effet juridique de l'horodatage électronique*. Disponible à : <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32014R0910>

INPI. (2025, 14 janvier). *Déposer une e-Soleau ou un entiercement*. Disponible à : <https://www.inpi.fr/realiser-demarches/propriete-intellectuelle/deposer-une-e-soleau-ou-un-entiercement>

Syndicat national de l'édition (SNE). (2025, 18 mars). *Press Release – Authors and publishers unite in lawsuit against Meta...* Disponible à : <https://www.sne.fr/press-release-authors-and-publishers-unite-in-lawsuit-against-meta-to-protect-copyright-from-infringement-by-generative-ai-developers/>

U.S. District Court, N.D. Cal. (2025, 23 juin). *Order on Fair Use (Filing 231)*. Disponible à : <https://docs.justia.com/cases/federal/district-courts/california/candce/3%3A2024cv05417/434709/231>

U.S. District Court, N.D. Cal. (2025, 25 juin). *Order... (Doc. 598)*. Disponible à : <https://law.justia.com/cases/federal/district-courts/california/candce/3%3A2023cv03417/415175/598/>

Common Crawl Foundation. (s.d.). *About / Overview*. Disponibles à : <https://commoncrawl.org> et <https://commoncrawl.org/overview>

OpenAI. (2025, 30 sept). *Sora 2 is here*. Disponibles à : <https://openai.com/index/sora-2>

## Articles :

Vaswani, A. et al. (2017). *Attention Is All You Need*. (Transformer). Disponible à : <https://arxiv.org/abs/1706.03762>

Brown, T. et al. (2020). *Language Models are Few-Shot Learners (GPT-3)*. PDF : <https://arxiv.org/pdf/2005.14165>

Radford, A. et al. (2019). *Language Models are Unsupervised Multitask Learners (GPT-2 / WebText)*. Disponible à : [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)

Mozilla Foundation – Records, P. (2024/2025). *Common Crawl (9,5+ PB depuis 2008) & rôle dans l'IA générative*. Disponible à : <https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/>

Carlini, N. et al. (2021). *Extracting Training Data from Large Language Models*. Disponible à : <https://arxiv.org/abs/2012.07805>

Sacem. (2023, 17 nov.). *Intelligence artificielle : la transparence, condition fondamentale d'un modèle européen éthique*. Disponible à : <https://societe.sacem.fr/actualites/droit-dauteur/intelligence-artificielle-la-transparence-condition-fondamentale-dun-modele-europeen-ethique>

Meta-Média / France Télévisions (citation D. El Sayegh). (2025, 22 mars). *Défendre le droit d'auteur à l'ère de l'IA générative*. Disponible à : <https://www.meta-media.fr/2025/03/22/defendre-le-droit-dauteur-a-lere-de-lia-generative.html>

University of California. (2025, 14 août). *How can visual artists protect their work from AI crawlers?*. Disponible à : <https://www.universityofcalifornia.edu/news/how-can-visual-artists-protect-their-work-ai-crawlers-its-complicated>

WIRED. (2024, 7 oct.). *The Race to Block OpenAI's Scraping Bots Is Slowing Down* (enquête Perplexity/robots.txt). Disponible à : <https://www.wired.com/story/open-ai-publisher-deals-scraping-bots>

The Atlantic. Reisner, A. (2025, 20 mars). *The Unbelievable Scale of AI's Pirated-Books Problem*. Disponible à : <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/>

WIRED. (2025, 9 janv.). *Meta Secretly Trained Its AI on a Notorious Piracy Database, Newly Unredacted Court Docs Reveal.* Disponible à : <https://www.wired.com/story/new-documents-unredacted-meta-copyright-ai-lawsuit>

Business Insider. Chowdhury, H. (2024, 14 nov.). *OpenAI... unsealed documents showed "books1" and "books2" were deleted (GPT-3).* Disponible à : <https://www.businessinsider.com/openai-share-files-copyright-case-authors-guild-ilya-sutskever-2024-11>

Le Monde. (2025, 12 mars). *Auteurs et éditeurs attaquent Meta... (Books3, LLaMA).* Disponible à : [https://www.lemonde.fr/economie/article/2025/03/12/auteurs-et-editeurs-attaquent-meta-pour-violation-du-droit-d-auteur\\_6579662\\_3234.html](https://www.lemonde.fr/economie/article/2025/03/12/auteurs-et-editeurs-attaquent-meta-pour-violation-du-droit-d-auteur_6579662_3234.html)

Reuters. Hals, T. (2025, 25 sept.). *US judge preliminarily approves \$1.5B Anthropic settlement.* Disponible à : <https://www.reuters.com/sustainability/boards-policy-regulation/us-judge-approves-15-billion-anthropic-copyright-settlement-with-authors-2025-09-25/>

Law firm / doctrine (Meta – Kadrey). Skadden. (2025, 8 juill.). *Fair Use and AI Training: Two Recent Decisions (Kadrey v. Meta, 25 juin 2025).* Disponible à : <https://www.skadden.com/insights/publications/2025/07/fair-use-and-ai-training>

\*

## **À propos de l'Adan**

L'Adan rassemble plus de 160 professionnels - nouveaux acteurs et entreprises établies - qui développent au quotidien l'innovation et les cas d'usage du web décentralisé dans tous les pans de l'économie. En levant les obstacles à leur croissance et leur compétitivité, l'Adan œuvre à l'émergence et au rayonnement des champions français et européens au service de notre souveraineté numérique. L'Adan promeut un encadrement adapté, proportionné et catalyseur de l'innovation, mais aussi une meilleure compréhension des nouvelles technologies blockchain, des crypto-actifs et de leurs opportunités.

### **Rédacteur**

- Sébastien Bellanger - Alternant Adan  
[✉ sebastien.bellanger@edhec.com](mailto:sebastien.bellanger@edhec.com)

### **Sous la supervision de :**

- Alizée Van Den Schrieck - Juriste  
[✉ alizee.vandenschrieck@adan.eu](mailto:alizee.vandenschrieck@adan.eu)
- Jules Dubourg - Secrétaire Général  
[✉ jules.dubourg@adan.eu](mailto:jules.dubourg@adan.eu)

L'Adan remercie ses membres qui ont participé à la relecture de cet article.