F

OMPI
ORGANISATION MONDIALE
DE LA PROPRIÉTÉ
INTELLECTUELLE

SCCR/49/6
ORIGINAL : ANGLAIS
DATE : 23 OCTOBRE 2026

Comité permanent du droit d'auteur et des droits connexes

Quarante-neuvième session
Genève, 30 novembre – 4 décembre 2026

ÉTUDE EXPLORATOIRE SUR L'ENTRAÎNEMENT DE L'IA ET LE DROIT D'AUTEUR

réalisée par Daniel J. Gervais, Sung Jai Choi, Melissa Omino, Rodrigo Kopke Salinas, Wang Qian

Étude exploratoire sur l'entraînement de l'IA et le droit d'auteur

AVERTISSEMENT :

La présente étude a été réalisée à la demande du Secrétariat de l'Organisation Mondiale de la Propriété Intellectuelle (OMPI). Les vues exprimées dans le présent document sont celles des auteurs et ne correspondent pas nécessairement à celles des États membres ou du Secrétariat.

REMERCIEMENTS

DANIEL J. GERVAIS est le coauteur principal de cette étude. Il est professeur émérite, titulaire de la chaire Milton R. Underwood de droit à la faculté de droit de Vanderbilt, professeur associé principal à la Vrije Universiteit d'Amsterdam et membre élu de l'American Law Institute. Ses travaux de recherche sont accessibles aux adresses suivantes : www.collectivemanagement.net, www.tripsagreement.com et www.thehumancause.org.

SUNG JAI CHOI est professeur de droit à l'Université Sejong. Il est titulaire d'un LL.M. de la faculté de droit de Columbia. Ses recherches portent sur la propriété intellectuelle et le droit de la concurrence. Il est coauteur de l'ouvrage Copyright Law (2024) et a reçu le Grand Prix d'arbitrage de Corée 2021 en tant que leader de la nouvelle génération.

MELISSA OMINO est directrice du Centre pour le droit de la propriété intellectuelle et des technologies de l'information (CIPIT) à l'université Strathmore de Nairobi. Titulaire d'un doctorat en droit de la propriété intellectuelle et en droit commercial, elle est spécialisée dans la propriété intellectuelle, l'intelligence artificielle, la protection des données et les droits numériques dans les pays du Sud.

RODRIGO KOPKE SALINAS est professeur à la faculté de droit de la FGV de São Paulo (GVlaw) et titulaire d'un LL.M. de l'Université Columbia. Il est membre de la Commission spéciale du droit d'auteur de l'ordre des avocats du Brésil et se spécialise dans le droit des médias et du divertissement.

WANG QIAN est professeur de droit et vice-président du comité académique de l'Université de sciences politiques et de droit de Chine orientale. Il est également vice-président de l'Association chinoise du droit d'auteur et a siégé au comité d'experts chargé de la troisième révision de la loi chinoise sur le droit d'auteur.

REMARQUE : Des modèles de langage ont été utilisés pour vérifier la syntaxe, la grammaire et le contenu de l'étude, ainsi que pour formuler des suggestions rédactionnelles visant à harmoniser ce texte rédigé en collaboration.

TABLE DES MATIÈRES

Introduction 5
Première partie. L'utilisation de contenus protégés dans le développement de modèles d'IA générative 6
Section A. Obtention des données 6
1. Modes de consommation des données selon le type de système 6
2. Le moissonnage de données sur le Web 7
3. Génération augmentée par récupération 8
Section B. Utilisation des données lors de l'entraînement 9
1. Entraînement du modèle 10
2. Optimisation et apprentissage par renforcement 11
Section C. Après l'entraînement 12
1. Utilisations ultérieures du corpus d'entraînement 12
2. Conservation de copies dans le modèle 12
3. La relation entre données d'entrée et résultat 15
Deuxième partie. Autorisation, concession de licence et rémunération 17
Section A. Options dont disposent actuellement les développeurs pour se procurer des éléments susceptibles d'être protégés par le droit d'auteur à des fins de formation 17
1. Les cinq modes d'acquisition 17
2. Accords de collaboration 19
Section B. Principaux éléments de la politique 20
1. Droits concernés par l'entraînement des modèles d'IA générative 20
a) Le droit de reproduction 20
b) Le droit d'adaptation et le droit aux œuvres dérivées 21
c) Communication au public et mise à disposition 21
d) Droit moral 21
e) Information sur le régime des droits et mesures antineutralisation 22
2. Concessions de licence volontaires (collectives et individuelles) 23
a) Concession de licences individuelles et collectives pour les utilisations de l'IA 23
b) Les principaux modèles de licence collective (volontaire, étendue et obligatoire) 27
3. Licences communautaires et de gestion responsable des données linguistiques primaires 31
4. Exceptions TDM réglementaires et clauses générales 33
5. Systèmes de rémunération 37
a) Types de systèmes de rémunération 37
b) Modalités et base de rémunération 38
Troisième partie. Application 40
Section A. Enseignements tirés de la jurisprudence actuelle 40
1. La question du "modèle en tant que copie" 40

2. L’exception relative à l’exploration de textes et de données 43
3. Compétence en matière d’entraînement hors du territoire national 44
4. La question de la substitution 46
Section B. Application des droits face aux développeurs d’IA 50
1. Caractéristiques structurelles du domaine de l’application des droits 50
2. Asymétries entre les parties 51
Section C. Défis dans le domaine de l’application des droits 51
1. Transparence 51
2. Attaques visant les données, filtrage et prolifération des normes de réserve 53
3. La jurisprudence dans son ensemble 56
Conclusion 58

INTRODUCTION

La présente étude a été établie à la demande du Comité permanent du droit d'auteur et des droits connexes (SCCR). À sa quarante-septième session, tenue à Genève du 1er au 5 décembre 2025, le comité a prié le Secrétariat d'établir une étude sur les approches politiques ou réglementaires de la relation entre l'entraînement de l'IA et le droit d'auteur, et les règles connexes, ainsi que les pratiques applicables en matière d'autorisation, d'application des droits et de rémunération en contrepartie de l'utilisation. Cette demande figure dans le résumé présenté par la présidente de cette session (document SCCR/47/SUMMARY[1] du 5 décembre 2025, paragraphe 21).

L'étude est de nature descriptive et comparative. Elle passe en revue les méthodes adoptées dans un certain nombre de ressorts juridiques ainsi que les pratiques en cours d'élaboration, et examine les options proposées sans recommander d'approche spécifique aux États membres ni prendre position sur les questions qu'ils doivent encore résoudre.

L'étude porte sur l'entraînement des modèles d'IA générative, c'est-à-dire l'utilisation de contenus pour la conception et l'amélioration de ces modèles, ainsi que sur les questions d'autorisation, d'application des droits et de rémunération soulevées par cette utilisation lorsque les contenus sont protégés par le droit d'auteur ou les droits connexes. L'intersection plus large entre l'intelligence artificielle et le droit d'auteur va bien au-delà de l'entraînement. Elle englobe notamment la question du statut juridique des résultats produits par un modèle déployé, tant pour déterminer si un résultat particulier porte atteinte aux œuvres sur lesquelles le modèle a été entraîné, que pour établir si et à quelles conditions un résultat peut lui-même être protégé par le droit d'auteur. Les questions relatives aux résultats soulèvent d'autres problématiques que le comité n'a pas demandé de traiter dans la présente étude. Elles ne sont mentionnées que lorsque cela est nécessaire pour brosser un tableau complet de la question de l'entraînement ou pour fournir le contexte nécessaire.

La présente étude s'articule en trois parties. La première partie décrit, d'un point de vue technique et descriptif, les opérations par lesquelles des contenus protégés sont utilisés dans le développement d'un modèle d'IA générative et identifie les étapes auxquelles ces opérations peuvent mettre en jeu les droits exclusifs reconnus par la législation nationale sur le droit d'auteur et les traités internationaux en la matière. La deuxième partie examine la portée et l'étendue des droits concernés ainsi que les mécanismes par lesquels ils sont exercés : les licences volontaires, individuelles ou collectives; ainsi que les exceptions et limitations applicables. La troisième partie traite de l'application des droits et examine les enseignements qui peuvent être tirés de la jurisprudence qui commence à se développer dans un certain nombre de ressorts juridiques. Elle aborde également la nature transfrontière de l'entraînement et les questions connexes d'application des droits, notamment la transparence et les normes relatives à l'expression du droit de retrait, le cas échéant. Les ressorts juridiques cités dans la présente étude le sont à titre d'illustration et ne constituent pas une liste exhaustive. Il s'agit notamment du Brésil, du Canada, de la Chine, de l'Union européenne et de bon nombre de ses États membres, du Kenya, de la République de Corée, du Royaume-Uni et des États-Unis d'Amérique.

---

[1] Voir https://www.wipo.int/meetings/fr/doc_details.jsp?doc_id=653196.

PREMIÈRE PARTIE. L'UTILISATION DE CONTENUS PROTÉGÉS DANS LE DÉVELOPPEMENT DE MODÈLES D'IA GÉNÉRATIVE

Un modèle d'intelligence artificielle (IA) générative est développé par étapes successives et des contenus de tiers sont généralement utilisés à chacune de ces étapes. Certains de ces contenus peuvent être soumis au droit d'auteur et à des droits connexes; d'autres peuvent relever du domaine public, soit en raison de l'expiration de la durée de protection, soit parce qu'ils n'ont jamais été protégés. Les opérations techniques appliquées à ces contenus sont généralement les mêmes, quelle que soit leur situation juridique. Les mêmes processus, notamment l'exploration, la copie, la tokenisation et l'analyse statistique, qui sont expliquées ci-dessous, peuvent s'appliquer aussi bien à un roman récent qu'à un texte du XVIII^{e} siècle relevant du domaine public. La présente partie décrit ces opérations et indique, à titre descriptif, les étapes auxquelles elles peuvent avoir une incidence sur les droits reconnus par la législation nationale sur le droit d'auteur et par les traités internationaux relatifs au droit d'auteur et aux droits connexes.

Trois phases d'entraînement sont présentées tour à tour. La section A traite de l'acquisition des données, c'est-à-dire de la manière dont le développeur du modèle d'IA construit le corpus, souvent composé en grande partie d'objets protégés, sur lequel entraîner le modèle. La section B traite de l'utilisation effective de ces données lors de l'entraînement, y compris le cycle d'entraînement principal et les différents types de réglage fin qui suivent. La section C aborde la situation après l'entraînement, notamment le traitement du corpus d'entraînement, ainsi que la question de savoir si le modèle entraîné lui-même contient des copies des œuvres[2] sur lesquelles il a été entraîné.

SECTION A. OBTENTION DES DONNÉES

Le développement d'un grand modèle de langage commence par la constitution d'un corpus, généralement à très grande échelle et d'origine variée. Les développeurs obtiennent les données de plusieurs manières. Certaines données peuvent faire l'objet d'une licence, auquel cas le contenu peut être mis à disposition avec la licence d'utilisation correspondante. Dans un tel cas de figure, un développeur peut acheter l'accès à un ensemble de données sélectionnées et compilées par un tiers, ou négocier un contrat de licence direct avec un éditeur, un organe de presse ou une organisation de gestion collective. Si aucune licence n'est accordée pour l'utilisation des œuvres protégées, celle-ci peut reposer sur l'hypothèse qu'une exception ou une limitation relative au droit exclusif concerné s'applique. L'une des principales sources de données d'entraînement, en termes de volume et de questions de droit d'auteur qu'elle soulève, consiste en la collecte automatisée de contenu sur le Web ouvert (Internet). L'entraînement peut également recourir à des données synthétiques, c'est-à-dire des données générées par un modèle d'IA existant plutôt que collectées auprès de créateurs humains, en particulier aux stades ultérieurs du développement.

1. Modes de consommation des données selon le type de système

Les besoins en données d'un système ne se limitent généralement pas au cycle d'entraînement initial : les données sont utilisées pour le réentraînement et le réglage fin, pour l'analyse de textes et de données effectuée sur une bibliothèque de recherche tenue à jour, pour la génération augmentée par récupération[3], ainsi que pour la génération d'aperçus et d'extraits. Ces utilisations varient tant par la quantité de contenu qu'elles requièrent que par la facilité avec laquelle leurs sources peuvent être retracées. Le type de système a tendance à influencer les modes de consommation : les systèmes de génération de contenus multimédias et les agents conversationnels basés sur de grands modèles de langage consomment principalement des données issues de la phase d'entraînement; les systèmes personnalisables développés par les

---

[2] Afin de simplifier le texte, le terme "œuvres" doit être compris comme incluant, le cas échéant, les objets protégés par des droits connexes.

[3] La génération augmentée par récupération est examinée plus avant dans la section 3 de la présente partie.

entreprises à partir d'un modèle de langage (également appelé "modèle de base") ajoutent un corpus spécifique au domaine et souvent une bibliothèque de recherche consultée au moment de l'utilisation; enfin, les moteurs de recherche conversationnels nécessitent l'accès à des sources en ligne au moment de la génération. La position d'un système sur ce spectre influe sur l'ensemble des droits exclusifs qu'implique son utilisation des données et sur la structure correspondante de toute licence. Ce point est développé dans la section B.2 de la deuxième partie.

2. Le moissonnage de données sur le Web

Le moissonnage de données sur le Web est une méthode couramment employée par les développeurs de modèles de base pour construire leur corpus d'entraînement. Cette terminologie n'est toutefois pas universellement adoptée. Certaines sources font la distinction entre l'"exploration du Web" (la navigation automatisée à travers des sites Web par un programme qui suit les liens d'une page à l'autre) et le "moissonnage de données sur le Web" (l'extraction sélective de contenu identifié à partir d'une page qui a déjà été récupérée). Dans la pratique, ces deux activités peuvent être combinées : un développeur déploie un robot d'indexation qui suit les liens sur le Web et, à chaque page, identifie et extrait le contenu textuel, visuel ou autre à conserver. Le résultat est souvent une copie de ce contenu, généralement stockée de manière structurée sur l'infrastructure propre du développeur.

Quatre caractéristiques sont pertinentes pour la présente analyse. Premièrement, la technique implique la réalisation de copies. Ces copies peuvent être conservées sur un serveur ou dans un cache, ou encore traitées en mémoire avant une transformation ultérieure. Ces opérations peuvent mettre en jeu le droit de reproduction reconnu à l'article 9.1) de la Convention de Berne, à l'article 1.4) du Traité de l'OMPI sur le droit d'auteur (WCT) et dans les dispositions correspondantes des législations nationales. La déclaration commune concernant l'article 1.4) du WCT confirme que le stockage d'une œuvre protégée sous forme numérique constitue une reproduction et que le droit de reproduction ainsi que les exceptions dont il peut être assorti s'appliquent pleinement dans l'environnement numérique[4]. La question de savoir si le caractère temporaire de certaines copies, ou l'application d'une exception telle que celle relative aux reproductions accessoires et transitoires, modifie l'analyse sera examinée plus loin dans la présente étude. L'article 7 du Traité de l'OMPI sur les interprétations et exécutions et les phonogrammes (WPPT) et le même article du Traité de Beijing sur les interprétations et exécutions audiovisuelles prévoient également un droit concernant les reproductions directes et indirectes "de quelque manière et sous quelque forme que ce soit".

Deuxièmement, la collecte peut être non sélective. Un robot d'indexation ne fait pas toujours la distinction entre les contenus relevant du domaine public, ceux proposés sous une licence ouverte, ceux mis à disposition sans aucune mention de licence, et ceux qui sont protégés et mis à disposition sous réserve de conditions d'utilisation ou protégés par une mesure technique de protection. Par exemple, certains robots d'indexation, mais pas tous, respectent les instructions du fichier robots.txt[5]. De telles distinctions sont parfois établies a posteriori, par le biais d'un filtrage ou d'une déduplication du corpus, mais l'acte initial de copie n'est pas toujours sélectif.

Troisièmement, la visibilité de cette activité pour les titulaires de droits est limitée. Un titulaire de droits peut ignorer que son œuvre a fait l'objet d'un moissonnage de données, par quel robot

---

[4] Traité de l'OMPI sur le droit d'auteur (20 décembre 1996), Déclaration commune concernant l'article 1.4) : "Le droit de reproduction énoncé à l'article 9 de la Convention de Berne et les exceptions dont il peut être assorti s'appliquent pleinement dans l'environnement numérique, en particulier à l'utilisation des œuvres sous forme numérique. Il est entendu que le stockage d'une œuvre protégée sous forme numérique sur un support électronique constitue une reproduction au sens de l'article 9 de la Convention de Berne".

[5] Le protocole d'exclusion des robots, normalisé sous la référence IETF RFC 9309 (2022), est une convention par laquelle l'exploitant d'un site Web place un fichier en texte brut (robots.txt) à la racine d'un site afin d'indiquer quels robots d'indexation automatisés peuvent accéder à quelles parties du site. Le respect de ce protocole est volontaire; aucun mécanisme d'application n'est prévu.

d'indexation, à quel moment ou en vue de quelle utilisation ultérieure. Les obligations de divulgation qui commencent à apparaître dans certaines législations nationales, en particulier au sein de l'Union européenne en vertu du règlement de 2024 sur l'intelligence artificielle (règlement sur l'intelligence artificielle), imposent aux fournisseurs de modèles d'IA à usage général de mettre à la disposition du public un "résumé suffisamment détaillé" du contenu utilisé pour l'entraînement[6]. Il est trop tôt pour évaluer l'effet concret de ces obligations. La question de savoir comment concilier cette obligation de divulgation avec le respect des intérêts commerciaux légitimes des développeurs fait elle-même l'objet d'un débat en cours[7].

Quatrièmement, la réalisation et l'utilisation de copies d'œuvres protégées à des fins d'entraînement peuvent impliquer la suppression ou la modification d'informations sur le régime des droits, ainsi que le contournement de mesures techniques efficaces utilisées par les titulaires de droits dans le cadre de l'exercice de leurs droits en vertu du WCT, du WPPT et du Traité de Beijing.

3. Génération augmentée par récupération

Une technique distincte, mais liée, est la génération augmentée par récupération[8]. Il s'agit d'une méthode par laquelle un système d'IA répond à une requête en consultant d'abord une collection distincte et préalablement constituée de documents, au lieu de se fier uniquement aux données sur lesquelles il a été entraîné. À titre de comparaison, on peut établir un parallèle avec une personne qui répond à une question de mémoire et la même personne qui prend le temps de consulter un ouvrage de référence rangé sur une étagère à côté de son bureau. Lorsqu'une requête est formulée, le système effectue une recherche dans cette collection, extrait les passages les plus pertinents par rapport à la question, les place à côté de la requête et s'en sert pour composer sa réponse. Cette collection est constituée à l'avance en créant et en stockant des copies des documents, et les passages pertinents sont souvent recopiés dans le système au moment où la réponse est générée, ce qui est important pour l'analyse au regard du droit d'auteur exposée ci-après.

Dans un système de génération augmentée par récupération, un corpus externe est tenu à jour, généralement indexé dans une base de données après que chaque document ou fragment de document a été converti en ce qu'on appelle un "plongement vectoriel"[9]. Lorsque le système reçoit une requête d'un utilisateur, celle-ci est convertie en représentation vectorielle et utilisée pour extraire les fragments les plus pertinents du corpus. Le système assemble ensuite la commande qui est transmise au modèle de langage, en combinant la requête de l'utilisateur avec les fragments extraits, que le modèle utilise pour générer un résultat sur la base de son entraînement et des contenus extraits.

---

6 Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (règlement sur l'intelligence artificielle), article 53.1.d) (les fournisseurs de modèles d'IA à usage général, "élaborent et mettent à la disposition du public un résumé suffisamment détaillé du contenu utilisé pour entraîner le modèle d'IA à usage général, conformément à un modèle fourni par le Bureau de l'IA.").

7 Il arrive toutefois que la provenance des contenus moissonnés transparaisse dans le comportement du modèle ainsi créé : les premières versions d'un modèle de génération d'images de premier plan reproduisaient une forme reconnaissable du filigrane d'une banque d'images, ce qui indiquait qu'un grand nombre d'images de cette agence avaient été ingérées sans autorisation. Ce phénomène a été largement médiatisé et a été invoqué dans le litige évoqué dans la section A.1 de la troisième partie (*Getty Images (US) Inc.* c. *Stability AI Ltd.*). Il illustre le fait que, bien que le moissonnage d'une œuvre particulière soit généralement invisible pour le titulaire des droits, le modèle entraîné peut lui-même révéler, à travers ses résultats, que du contenu protégé a été utilisé lors de l'entraînement.

8 Voir la note 2.

9 Les plongements lexicaux représentent chaque jeton sous la forme d'un vecteur dense de grande dimension composé de nombres réels. Ces représentations sont qualifiées de "contextuelles" lorsque la représentation attribuée à un jeton donné dépend des jetons qui l'entourent. Voir Heart of the Matter, note 10, partie II.C.

Dans le cadre qui nous intéresse, la génération augmentée par récupération est parfois présentée comme modifiant les questions de droit d'auteur soulevées par l'entraînement des modèles de langage. Le raisonnement est que le modèle lui-même n'apprend pas à partir du corpus, car celui-ci est conservé séparément et consulté au moment de l'inférence. La construction du corpus indexé nécessite un traitement préalable et, dans une architecture de génération augmentée par récupération classique, le stockage des documents sources ou des fragments qui les composent[10]. La conversion de chaque document en un plongement vectoriel produit une représentation numérique dérivée du document, bien que, contrairement à la copie source stockée, le plongement vectoriel ne soit généralement pas réversible vers le texte original. Au moment de l'exécution (inférence), les fragments pertinents sont copiés dans la fenêtre de contexte du modèle et peuvent être reproduits, en totalité ou en grande partie, dans le résultat. Ces étapes mettent en jeu le droit de reproduction; elles pourraient également mettre en jeu le droit d'adaptation (Convention de Berne, article 12) en fonction de la législation applicable et des caractéristiques de la représentation. L'application éventuelle de ces droits est examinée ci-dessous.

SECTION B. UTILISATION DES DONNÉES LORS DE L'ENTRAÎNEMENT

L'entraînement d'un modèle d'IA générative consiste en une séquence d'opérations généralement décrite comme comportant quatre étapes : le préentraînement, le réglage fin supervisé, la modélisation des récompenses et l'apprentissage par renforcement[11]. La première de ces étapes est généralement la plus exigeante, tant sur le plan du coût de calcul que du volume de données consommées; les trois étapes suivantes, prises dans leur ensemble, sont souvent qualifiées de "réglage fin" ou, parfois, de "post-entraînement". Une activité distincte, mais liée, à savoir l'évaluation et l'optimisation du modèle, accompagne l'entraînement et se poursuit pendant le déploiement du modèle.

Du point de vue du droit d'auteur, le principal élément à retenir est que chacune de ces étapes implique généralement le stockage et le traitement de tout ou partie du corpus sous-jacent. Le processus d'entraînement n'est pas un processus dans lequel les données sont brièvement consultées puis écartées; les données sont stockées, chargées, transformées et utilisées pour mettre à jour le modèle. Lorsque les données comprennent des œuvres protégées, ces opérations impliquent des copies ou des transformations relevant de l'analyse de la reproduction développée dans la deuxième partie.

---

[10] D'un point de vue plus technique, il convient de faire la distinction entre les représentations vectorielles de passages utilisées pour la recherche vectorielle et le texte source à partir duquel elles sont générées. Un plongement vectoriel est une représentation numérique dense produite par un encodeur; ce n'est pas une copie exacte du texte et elle n'est généralement pas réversible par simple décodage du vecteur. Des techniques spécialisées d'"inversion de représentations" peuvent néanmoins permettre de reconstruire une partie de passages sources relativement courts, voire, dans certaines conditions expérimentales, leur intégralité, à partir de leurs représentations. Dans un système classique de génération augmentée par récupération, les documents ou fragments sources sont également conservés sous forme de texte, soit dans le système de stockage des vecteurs lui-même, soit dans un système de stockage de documents distinct lié aux vecteurs. Les représentations vectorielles sont utilisées pour identifier les passages pertinents; c'est alors le texte associé, et non le vecteur, qui est récupéré et fourni au modèle de langage dans le cadre de son contexte d'entrée. Voir John X. Morris *et al.*, "Text Embeddings Reveal (Almost) As Much As Text", Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP) (2023). Pour une description du processus de récupération par un acteur du secteur, voir : https://infohub.delltechnologies.com/en-us/p/demystifying-on-device-intelligent-search-using-rag-architecture/.

[11] La description en quatre étapes utilisée ici est tirée de la littérature technique et est abordée plus en détail dans l'article "The Heart of the Matter: Copyright, AI Training, and LLMs" 71 J Cop Soc'y of the USA 482 (2024) (ci-après "Heart of the Matter"), disponible à l'adresse https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4963711. Un processus similaire en quatre étapes (collecte des données, prétraitement des données, entraînement du modèle, et évaluation et optimisation du modèle) est présenté dans le Guide sur l'applicabilité de la doctrine de l'usage loyal à l'entraînement des modèles d'IA générative (février 2026) publié par le Ministère de la culture, des sports et du tourisme et la Commission coréenne du droit d'auteur (ci-après le "Guide coréen sur l'usage loyal"), page 10.

1. Entraînement du modèle

Le cycle d'entraînement principal, généralement appelé "pré-entraînement", est l'étape au cours de laquelle le modèle acquiert l'essentiel de ses connaissances sur le langage (dans le cas d'un modèle de langage) ou sur les images, l'audio et la vidéo (dans le cas d'un modèle multimodal). Le préentraînement se déroule en plusieurs étapes techniques, chacune étant effectuée sur une copie stockée du corpus. Le contenu textuel est d'abord soumis à une tokenisation, c'est-à-dire la conversion des mots et autres séquences en un flux de jetons numériques issus d'un vocabulaire fixe[12]. Les tokeniseurs modernes opèrent généralement au niveau des sous-mots plutôt qu'au niveau des mots entiers[13]. Cette étape revêt une importance particulière pour la présente analyse car, dans le cas des tokeniseurs opérant au niveau de l'octet utilisés par les principaux modèles contemporains, elle est sans perte et réversible : un texte encodé par un tel tokeniseur puis décodé restitue le texte d'origine; le corpus tokenisé n'est donc pas un résumé ou une abstraction des œuvres, mais un réencodage de celles-ci sous une autre forme.

Les jetons sont ensuite convertis en plongements vectoriels[14]. L'approche qui domine le domaine depuis 2017, le Transformer (le "T" de l'acronyme GPT, qui signifie Generative Pre-trained Transformer), traite ces représentations à l'aide d'un mécanisme appelé "attention"[15]. La caractérisation du plongement vectoriel, ainsi que celle du modèle entraîné lui-même, qui impliquent un degré de transformation plus important, est examinée dans la section C.2 ci-dessous.

Une fois le modèle initialisé, il est entraîné en étant exposé de manière répétée à une séquence de jetons tirés du corpus. Le modèle prédit chaque jeton à tour de rôle en se basant sur les jetons qui le précèdent et est ajusté, en plusieurs petites étapes, de sorte que ses prédictions se rapprochent du texte observé; le corpus n'est donc pas consulté une seule fois, mais utilisé, jeton par jeton, tout au long du cycle d'entraînement. Après avoir traité un volume considérable de jetons tirés du corpus, le modèle a intériorisé une vaste distribution de probabilité couvrant un vaste espace de séquences de jetons possibles. C'est cette distribution, codée dans les paramètres du modèle, qui permet à celui-ci de générer un texte cohérent en réponse à une commande.

La relation entre les données d'entraînement et le modèle entraîné fait l'objet d'un débat approfondi. Selon un point de vue, étant donné que le modèle entraîné ne contient que des paramètres et non les données d'entraînement d'origine, aucune copie des œuvres protégées ne subsiste une fois l'entraînement terminé[16]. Ces paramètres décrivent une distribution de probabilité sur un espace vectoriel extrêmement vaste. Si cette distribution permet au modèle de généraliser et de générer des séquences entièrement nouvelles, elle peut attribuer des probabilités disproportionnellement élevées et concentrées à des séquences spécifiques et exactes rencontrées à plusieurs reprises au cours de l'entraînement. Comme la distribution est

---

12 En ce qui concerne la tokenisation, voir "Heart of the Matter", note 10 ci-dessus, partie II.B. Le "Guide coréen sur l'usage loyal" mentionne la tokenisation comme l'une des procédures techniques de prétraitement des données, au même titre que le nettoyage et la normalisation. Voir "Guide coréen sur l'usage loyal", note 10 ci-dessus, page 10.

13 Le vocabulaire est constitué en partant de caractères individuels, puis en regroupant progressivement les paires adjacentes les plus fréquentes en unités plus grandes, de sorte que les mots courants finissent par être représentés par un seul jeton, tandis que les mots plus rares sont découpés en plusieurs. Le codage par paires d'octets est utilisé par OpenAI pour ses modèles Generative Pretrained Transformer (GPT) ainsi que par un certain nombre d'autres développeurs majeurs. Voir "Heart of the Matter", note 10 ci-dessus, partie II.B.

14 Plus précisément, ils sont projetés dans un espace vectoriel de haute dimension, dans lequel chacun est représenté sous la forme d'un plongement vectoriel.

15 Le mécanisme d'attention permet au modèle, lorsqu'il calcule la représentation d'un jeton donné, de pondérer la pertinence de chaque autre jeton de la séquence, ce qui lui permet de saisir les dépendances au sein de longs passages de texte. Ashish Vaswani *et al.*, "Attention Is All You Need", 31 Conf. Neural Inf. Processing Syst. (NeurIPS) (2017). L'architecture Transformer est à la base des principaux modèles de langage de grande envergure actuels.

16 Voir Heart of the Matter, note 10 ci-dessus, partie II.C.

fortement pondérée en faveur de ces chemins d'entraînement finis, le modèle peut, dans de nombreux cas, être amené à les reproduire mot pour mot ou presque mot pour mot. Ce phénomène est largement documenté dans la littérature technique et est abordé plus en détail dans la section C ci-dessous. Comme expliqué ci-dessous, il existe une différence entre ce phénomène (souvent appelé "mémorisation") et la question de savoir s'il existe une "reproduction" au sein du modèle de toutes les œuvres utilisées pour l'entraînement.

Pour les modèles générant des images, du son ou de la vidéo, les modèles les plus courants sont dits de "diffusion". Ils apprennent en ajoutant progressivement du bruit aux exemples d'entraînement, puis en apprenant à inverser ce processus afin de produire de nouveaux exemples à partir du bruit[17]. Cela diffère des modèles de langage (textuels) mais, dans les deux cas, le processus d'entraînement produit une représentation sophistiquée des données d'entrée.

2. Optimisation et apprentissage par renforcement

Les étapes qui suivent le préentraînement consistent à adapter le modèle préentraîné à des tâches spécifiques, à des cas d'utilisation particuliers et à des normes de comportement précises. La terminologie varie selon les développeurs, mais la séquence comprend le réglage fin supervisé, la modélisation des récompenses et l'apprentissage par renforcement[18].

Le réglage fin supervisé implique un entraînement supplémentaire sur un ensemble de données plus petit et soigneusement sélectionné, composé de paires entrée-sortie du type de celles que le modèle est censé produire. Cet ensemble de données est bien plus petit que le corpus de préentraînement, mais il est sélectionné pour sa qualité et est souvent créé ou révisé par des humains. Chaque exemple de réglage fin, à l'instar de chaque exemple de préentraînement, est chargé et utilisé pour mettre à jour les paramètres du modèle.

La modélisation des récompenses implique la construction d'un modèle de "récompense" distinct qui attribue un score aux réponses candidates du modèle principal selon qu'elles sont plus ou moins souhaitables. Le modèle de récompense est lui-même entraîné à partir de comparaisons, établies par des humains, entre des réponses candidates.

L'apprentissage par renforcement à partir de la rétroaction humaine, une technique largement utilisée pour aligner les modèles d'IA générative sur les préférences humaines, utilise le modèle de récompense pour guider les mises à jour ultérieures du modèle principal.

Chacune de ces étapes nécessite le stockage et le traitement de données d'entraînement. Les volumes sont plus faibles que ceux requis pour le préentraînement, mais les données d'entraînement peuvent elles-mêmes inclure des œuvres protégées, en particulier lorsque les comparaisons fournies par des annotateurs humains sont basées sur de telles œuvres ou en incluent. Le fait de privilégier la qualité plutôt que la quantité des données ne modifie en rien l'analyse au regard du droit d'auteur.

---

[17] Concernant les origines des modèles de diffusion, voir Jascha Sohl-Dickstein *et al.*, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics", Proc. 32nd Int'l Conf. on Machine Learning (ICML) (2015); Jonathan Ho, Ajay Jain et Pieter Abbeel, "Denoising Diffusion Probabilistic Models", 34 Conf. Neural Inf. Processing Syst. (NeurIPS) (2020); et, concernant l'architecture de diffusion latente utilisée dans Stable Diffusion et en cause dans le litige Getty évoqué dans la troisième partie, Robin Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). Pour une introduction accessible, voir également Yang Song et Stefano Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution", 33 Conf. Neural Inf. Processing Syst. (NeurIPS) (2019), disponible à l'adresse http://papers.neurips.cc/paper/9361-generative-modeling-by-estimating-gradients-of-the-data-distribution.pdf. Le propos général de la présente étude ne repose pas sur les détails mathématiques d'une classe particulière de modèles.

[18] Voir "Heart of the Matter", note 10 ci-dessus, partie II.D. Les termes "pré-entraînement", "réglage fin supervisé", "modélisation des récompenses" et "apprentissage par renforcement" sont couramment utilisés par les développeurs d'IA.

L'évaluation et l'optimisation du modèle, qui se déroulent parallèlement aux étapes d'entraînement et continuent après le déploiement, impliquent l'utilisation d'ensembles de données d'évaluation pour mesurer les performances du modèle et identifier les domaines dans lesquels un entraînement supplémentaire est nécessaire. Les ensembles de données d'évaluation peuvent eux-mêmes inclure des œuvres protégées, en particulier lorsque les jeux de données de référence ont été constitués à partir de corpus de test accessibles au public.

SECTION C. APRÈS L'ENTRAÎNEMENT

Trois questions distinctes se posent une fois l'entraînement terminé. La première est de savoir si le corpus est conservé, par qui et à quelles fins ultérieures. La deuxième est de savoir si, et en quel sens, le modèle contient des copies des œuvres sur lesquelles il a été entraîné. La troisième est de savoir si les résultats reproduisent, adaptent ou mettent en jeu d'une autre manière les œuvres utilisées lors de l'entraînement.

1. Utilisations ultérieures du corpus d'entraînement

Le corpus d'entraînement reste pertinent une fois le cycle d'entraînement terminé. Il peut être conservé par le développeur à plusieurs fins : pour évaluer de futures versions du modèle, pour faciliter un réglage fin supplémentaire sur des sous-ensembles particuliers, pour permettre la poursuite des recherches, ou pour être mis à la disposition de tiers sous licence ou selon d'autres modalités. Il peut être combiné avec de nouveaux éléments afin de former un corpus élargi destiné à la prochaine génération de modèles. Il peut être vendu ou partagé (par exemple, avec des organisations affiliées). Dans certains cas, des dérivés du corpus, notamment des sous-ensembles filtrés ou dédupliqués, sont mis à la disposition du public sous des licences permissives.

Chacune de ces utilisations ultérieures peut impliquer des actes de reproduction, de distribution et de mise à disposition au sens des règles nationales et internationales applicables. En d'autres termes, un corpus qui a rempli son objectif d'entraînement peut continuer d'être utilisé pendant des années, et son utilisation ultérieure obéit aux mêmes règles que sa constitution initiale.

2. Conservation de copies dans le modèle

La question de savoir si le modèle entraîné contient lui-même des "copies" des œuvres sur lesquelles il a été entraîné est l'une des questions les plus complexes sur le plan technique abordées dans cette partie, et c'est précisément ce qui distingue l'IA générative des technologies antérieures. Les positions défendues par les développeurs d'IA, d'une part, et par les titulaires de droits, d'autre part, étaient jusqu'à récemment diamétralement opposées[19]. Les développeurs d'IA ont généralement fait valoir que le modèle entraîné ne contient que des paramètres, et non les œuvres elles-mêmes, et qu'aucune copie n'est conservée au sens de la législation sur le droit d'auteur. Les titulaires de droits ont fait valoir que ce modèle constituait, en réalité, une copie de l'intégralité du corpus. Les données empiriques recueillies au cours des deux dernières années semblent indiquer que ces deux positions ont tendance à simplifier à l'excès la relation entre la mémorisation et le modèle. De plus, quelle que soit l'issue de ce désaccord, cela n'affecte en rien les actes antérieurs : l'assemblage du corpus et son utilisation

[19] Cette divergence va au-delà des positions adoptées dans le cadre des litiges pour s'étendre aux priorités en matière de conception. Une étude Delphi en plusieurs étapes, menée auprès d'éthiciens spécialisés en IA, d'experts juridiques, de praticiens de l'IA musicale et d'artistes, afin de valider un cadre de conception responsable pour les systèmes musicaux génératifs, a mis en évidence un alignement "protectionniste" entre les artistes et les experts juridiques sur les questions de souveraineté des données, contrastant avec l'accent mis par les praticiens sur la faisabilité. Elle indiquait en conclusion que de tels cadres fonctionnent mieux en tant que couche de traduction entre les priorités divergentes des parties prenantes, notamment le compromis entre la transparence des modèles et les intérêts en matière de propriété intellectuelle, plutôt qu'en tant qu'instruments issus d'un consensus forcé. J. de Berardinis *et al.*, "RAIM : Operationalising Trustworthy AI in Music Generation" (RAIM : Mise en œuvre d'une IA fiable dans la génération musicale), IEEE Transactions on Technology and Society, numéro spécial consacré à l'innovation éthique dans les technologies musicales (à paraître en 2026), site Web du projet https://casbahlab.github.io/raim-framework/ consulté le 25 août 2026.

à des fins de formation impliquent des reproductions dont la qualification juridique est indépendante du fait que le modèle achevé contienne lui-même des copies.

Deux types d'éléments techniques peuvent être pris en considération dans l'analyse. La mémorisation est une propriété du modèle entraîné qui permet de retrouver, avec une fidélité suffisante, des séquences d'entraînement spécifiques dans les paramètres du modèle. Étant donné qu'il est difficile d'examiner directement la structure interne d'un modèle, la mémorisation est étudiée par l'extraction, c'est-à-dire la capacité démontrée du modèle, sous l'effet d'une incitation appropriée, à reproduire des séquences issues de ses données d'apprentissage, mot pour mot ou presque. L'extraction sert alors de preuve de la mémorisation. L'échec d'une procédure d'extraction donnée ne prouve pas qu'une œuvre n'ait pas été mémorisée. Une étude systématique menée par Cooper et ses collaborateurs a consisté à appliquer une technique d'extraction probabiliste à 200 ouvrages sur 14 modèles linguistiques à poids libre; elle a révélé que le degré de mémorisation variait considérablement tant d'un modèle à l'autre que d'un ouvrage à l'autre[20]. La plupart des plus grands modèles n'ont pas mémorisé la majeure partie des livres, ni dans leur intégralité ni en partie. Lorsque la mémorisation avait lieu, elle pouvait toutefois être considérable : l'étude indiquait qu'un modèle largement utilisé avait mémorisé presque intégralement le premier roman de la série Harry Potter et le roman 1984 de George Orwell; dans le cas du premier, l'étude a démontré que la quasi-totalité du livre pouvait être reproduite presque mot pour mot.

Ces démonstrations empiriques de la mémorisation ne font toutefois pas l'unanimité. Les développeurs d'IA ont réagi en faisant valoir que les techniques de "prompting" nécessaires pour obtenir des reproductions mot pour mot sont "antagonistes", ne reflètent pas la manière dont les systèmes sont utilisés dans la pratique et peuvent elles-mêmes constituer une manipulation ou une utilisation abusive du produit qui sort du cadre d'une utilisation normale (et, dans certains cas, des conditions générales d'utilisation). Dans le cadre du litige opposant le New York Times à OpenAI, le défendeur a demandé le rejet des chefs d'accusation de contrefaçon directe du droit d'auteur, en invoquant notamment le fait que le plaignant avait utilisé ce que le défendeur qualifiait de "consignes élaborées et manipulées" pour produire la régurgitation dont il se plaignait[21]. En avril 2025, le Tribunal fédéral du district Sud de New York a refusé de rejeter les demandes sur ce motif, et le litige a depuis évolué, au cours de la phase d'instruction, passant de la question de savoir si la "régurgitation" peut être provoquée artificiellement à celle de sa fréquence d'apparition spontanée dans la pratique, ce qui a abouti à une décision imposant la production d'un échantillon représentatif de 20 millions de journaux

---

[20] Feder Cooper, Aaron Gokaslan, Ahmed Ahmed, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho et Percy Liang, "Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models" (Extraction de passages mémorisés de livres (protégés par le droit d'auteur) à partir de modèles linguistiques à poids ouvert", arXiv:2505.12546 (2025), version du 30 novembre 2025 (également publié sous le titre "Stanford Public Law Working Paper / WVU College of Law Research Paper n° 2025-005"). Il convient de citer dans son intégralité la formulation de la question par les auteurs (page 1) : "Dans les litiges en matière de droit d'auteur portant sur l'IA générative, les plaignants et les défendeurs avancent souvent des arguments radicaux et contradictoires quant à la mesure dans laquelle les grands modèles linguistiques (LLM) ont mémorisé les expressions protégées des plaignants. En nous appuyant à la fois sur l'apprentissage automatique et sur le droit d'auteur, nous démontrons que ces positions polarisées simplifient à l'extrême la relation entre la mémorisation et le droit d'auteur". Concernant l'idée selon laquelle le contenu mémorisé doit être encodé dans les paramètres du modèle, car "il ne pourrait se trouver nulle part ailleurs", voir également A. Feder Cooper et James Grimmelmann (cités dans cet ouvrage à la page 25).

[21] *The New York Times Company* c. *Microsoft Corp.*, n° 1:23-cv-11195, Mémoire à l'appui de la requête en irrecevabilité des défendeurs (S.D.N.Y., 26 février 2024); Arrêt et ordonnance (S.D.N.Y., 4 avril 2025) (J. Stein) (rejetant, dans la partie pertinente, les requêtes en irrecevabilité des chefs d'accusation de contrefaçon directe et de contrefaçon par contribution). Pour connaître la position officielle d'OpenAI sur cette affaire, voir OpenAI, "Reporting the Facts About the New York Times' Lawsuit", disponible à l'adresse openai.com/new-york-times.

d'interactions d'utilisateurs réels[22]. Ce litige, ainsi que la portée de cette ordonnance, sont examinés plus en détail dans la partie III, section A.4, de la présente étude.

En résumé, les résultats empiriques remettent en cause ces deux positions opposées : le modèle entraîné n'est pas, d'une manière générale, une "copie" (copy) de son corpus d'entraînement, mais il n'en est pas non plus "indépendant" (free). La portée exacte varie selon les modèles.

La deuxième source d'informations est la littérature plus générale consacrée aux attaques par extraction de données, qui décrit les circonstances dans lesquelles des exemples d'apprentissage individuels peuvent être récupérés à partir d'un modèle entraîné grâce à des consignes ciblées[23]. Parmi les éléments récupérés figurent des passages de texte assez longs, des images identifiables et des informations permettant d'identifier des personnes. Ces travaux ont été cités comme preuve que l'extractibilité peut être une propriété structurelle des grands modèles génératifs, plutôt que le simple résultat d'une manipulation de la part de l'utilisateur. Certaines entreprises spécialisées dans l'IA affirment que ce type de consigne constitue une manipulation déloyale ou une violation des conditions d'utilisation[24]. Les évaluations de ce type sont toutefois devenues plus complexes au cours de l'année écoulée : les systèmes déployés intègrent de plus en plus souvent des filtres et d'autres mécanismes de protection entre le modèle et l'utilisateur afin de limiter la possibilité d'extraire le contenu mémorisé; des études récentes ont dû recourir à des techniques antagonistes pour certains systèmes, tandis que d'autres ont permis d'obtenir le contenu protégé sans y avoir recours[25].

Une analogie pourrait s'avérer utile pour comprendre la situation. C'est une analogie avec les programmes informatiques[26]. Lorsque le code source d'un programme (souvent écrit dans un langage s'apparentant à un langage naturel) est compilé sous forme exécutable, le résultat diffère du code source et n'est pas (facilement) lisible par un être humain. La compilation n'efface toutefois pas l'original : le fichier exécutable conserve les informations du programme source et, par ingénierie inverse, il est possible de reconstituer l'original dans une large mesure. À cet égard, le code exécutable reste similaire au programme source. Les étapes qui permettent d'obtenir un modèle entraîné à partir d'un corpus d'entraînement sont, si l'on suit cette analogie, sans doute de nature comparable. Dans les deux cas, une source intelligible (un programme sous forme de code source; un corpus d'œuvres) est transformée en un objet numérique fonctionnel qui ne lui ressemble plus, mais qui continue de véhiculer des informations qui en découlent. Cette analogie a toutefois ses limites : la compilation est un processus déterministe

---

22 Dans l'affaire OpenAI, Inc. relative à une violation du droit d'auteur, n° 1:25-md-03143-SHS-OTW (S.D.N.Y.), arrêt et ordonnance (J. Stein, 5 janvier 2026), confirmant l'arrêt et l'ordonnance (M.J. Wang, 2 décembre 2025) (rejetant la requête en réexamen d'OpenAI) ainsi que l'ordonnance de communication de pièces sous-jacente de novembre 2025. Le tribunal a ordonné la production d'un échantillon aléatoire et anonymisé de 20 millions de journaux d'interactions d'utilisateurs de ChatGPT, tirés de l'ensemble des données conservées, estimant que des preuves générales relatives aux résultats destinés au grand public étaient pertinentes pour l'analyse de l'usage loyal, en particulier pour déterminer si les résultats générés par le modèle constituaient un substitut commercial aux œuvres des plaignants.

23 Parmi les travaux antérieurs allant dans le même sens, on peut citer celui d'Antonia Karamolegkou *et al.*, "Copyright Violations and Large Language Models", arXiv:2310.13771 (2023) (qui démontre, sur un large éventail de modèles linguistiques, que des exemples individuels de données d'entraînement peuvent être récupérés par le biais d'attaques par extraction de données).

24 Voir Nicholas Carlini *et al.*, "Extracting Training Data from Large Language Models", 30e symposium USENIX sur la sécurité (2021), p. 2633; Milad Nasr *et al.*, "Scalable Extraction of Training Data from Aligned, Production Language Models", ICLR 2025 (prolongeant Nasr *et al.*, arXiv:2311.17035 (2023)).

25 Ahmed Ahmed *et al.*, "Extracting Books from Production Language Models", arXiv:2601.02671 (2026) (concluant que l'utilisation de consignes antagonistes était nécessaire pour extraire des passages mémorisés de livres de certains systèmes déployés, mais pas d'autres).

26 Voir *Heart of the Matter*, note 10 ci-dessus, partie II.C (établissant une analogie entre l'entraînement d'un modèle linguistique et la compilation d'un programme informatique sous une forme exécutable, et soulignant que, dans les deux cas, "l'œuvre originale protégée par le droit d'auteur" est "traduite sous une autre forme qui n'est pas directement compréhensible par les humains", mais dans laquelle le contenu de l'original "n'est [pas] absent de cette version compréhensible par la machine").

qui transpose l'intégralité du programme source dans le fichier exécutable, tandis que l'encodage des données d'apprentissage dans un modèle est sélectif et en partie probabiliste, de sorte que l'analogie ne vaut, tout au plus, que pour le sous-ensemble des œuvres dont il est démontré qu'elles ont été mémorisées; des études récentes soutiennent que c'est précisément sur cette différence entre une récupérabilité déterministe et une récupérabilité purement probabiliste que portera vraisemblablement la qualification juridique du modèle[27].

Les éléments techniques examinés ici indiquent qu'un modèle entraîné conserve les informations issues des œuvres figurant dans ses données d'entraînement, même si la transformation est importante et que l'accès à une œuvre donnée peut être partiel et probabiliste. La manière dont cette observation technique se traduit par la qualification juridique du modèle et par la question de savoir si ce dernier constitue, au regard du droit d'auteur, une copie ou une adaptation/œuvre dérivée des œuvres sur lesquelles il a été entraîné fait l'objet d'une analyse continue devant les tribunaux et dans la littérature spécialisée, comme l'explique cette étude plus loin.

On peut formuler deux remarques à ce stade. D'une part, la mémorisation est invoquée comme argument tant au stade de l'entraînement – où elle sert à démontrer que l'entraînement du modèle a impliqué l'encodage de copies – qu'au stade de la production, où elle sert à démontrer que le modèle déployé est capable de générer des œuvres protégées sous une forme qui se substitue aux originaux. D'autre part, les démonstrations sur lesquelles s'appuient les plaignants portent sur des œuvres spécifiques figurant dans des modèles spécifiques. Ils tendent à confirmer que la mémorisation peut jouer un rôle important pour des œuvres spécifiques, individuelles et fréquemment rencontrées, mais ne démontrent pas qu'il s'agisse d'une propriété commune à tous les modèles ni que les paramètres d'un modèle se réduisent, de manière générale, à des copies stockées. La qualification juridique de ces questions est abordée dans la partie III[28].

Enfin, il convient également de noter qu'un modèle peut évoluer une fois l'apprentissage terminé, par le biais d'un affinage, d'un apprentissage d'alignement ou de tentatives de "désapprentissage" de la machine ciblant des contenus spécifiques. Le fait qu'un modèle cesse de reproduire une œuvre donnée à la suite d'une telle intervention ne prouve pas pour autant qu'il ne l'ait jamais encodée. La littérature technique indique qu'il reste difficile de supprimer de manière fiable des œuvres spécifiques d'un modèle entraîné, et que de telles interventions peuvent nuire au modèle de manière imprévisible[29].

3. La relation entre données d'entrée et résultat

Une fois le modèle déployé, il fonctionne en recevant des données d'entrée de l'utilisateur, ou une consigne, et en générant un résultat. Le résultat peut être du texte, du code, une image, un

---

[27] M.A. Lemley et A.F. Cooper, "Probabilistic 'Copies' in Generative AI Models", Berkeley Technology Law Journal (2026), disponible à l'adresse https://ssrn.com/abstract=7067878 (soutenant, en s'appuyant sur la littérature relative à l'extraction empirique évoquée dans la présente partie, qu'un modèle ne doit être considéré comme contenant une "copie" d'une œuvre particulière que lorsque celle-ci peut être extraite de manière simple et, dans les cas les plus marqués, déterministe; que lorsque l'extraction ne réussit que de manière probabiliste, après de nombreuses tentatives, ce qui est stocké n'est pas une copie de l'œuvre, mais un ensemble d'"ingrédients"; et que la question relève donc de la preuve, à laquelle il convient de répondre œuvre par œuvre et modèle par modèle).

[28] Concernant la procédure regroupée, voir l'affaire "In re OpenAI, Inc. Copyright Infringement Litigation", n° 1:25-md-03143-SHS-OTW (S.D.N.Y.), regroupant les actions intentées par le New York Times et par des groupes d'auteurs contre OpenAI et Microsoft. La description du modèle entraîné comme un mécanisme de stockage et de récupération, et celle des paramètres comme une forme compressée des œuvres intégrées, reflète les arguments des demandeurs; la description contraire des paramètres, à savoir qu'il s'agit de généralisations statistiques codant des schémas linguistiques, reflète la position des défendeurs. En ce qui concerne la littérature empirique mettant en garde contre ces deux interprétations, voir l'étude de Cooper *et al.* citée plus haut dans cette section.

[29] A. Feder Cooper *et al.*, "Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice" (Le désapprentissage des machines ne fonctionne pas comme vous le pensez : enseignements pour la politique, la recherche et la pratique en matière d'IA générative), 39e conférence. Neural Inf. Processing Syst. (NeurIPS) (2025).

fichier audio, une vidéo ou tout autre type de contenu. La relation entre données d'entrée et résultat s'apparente à certains égards à celle qui existe entre une requête et une réponse dans un moteur de recherche, mais cette analogie n'est que partielle. Le modèle ne consulte pas un corpus indexé pour en extraire des passages; il génère du contenu nouveau en échantillonnant la distribution de probabilité qu'il a intériorisée au cours de l'entraînement. Dans certains cas, ce qu'il génère est étroitement lié à des œuvres spécifiques présentes dans ses données d'apprentissage; dans d'autres cas, les données de sortie sont une composition originale qui ne ressemble étroitement à aucun exemple d'apprentissage en particulier.

Trois éléments méritent particulièrement d'être soulignés. Premièrement, le résultat peut être une reproduction mot pour mot ou quasi mot pour mot d'une œuvre figurant dans les données d'entraînement, auquel cas il s'agit d'une reproduction de l'œuvre originale[30]. Cela a été démontré dans le cadre d'études portant sur des modèles linguistiques reproduisant des extraits d'ouvrages protégés par le droit d'auteur et sur des modèles de génération d'images reproduisant des photographies et des œuvres d'art protégées[31]. La question de savoir si une telle reproduction constitue une atteinte dépendra d'un certain nombre d'autres éléments, notamment l'applicabilité des exceptions et des limitations, le rôle des filtres de sortie éventuellement mis en place par le développeur, ainsi que la répartition des responsabilités entre le développeur du modèle et l'utilisateur qui a saisi la consigne. Ces questions sont brièvement abordées plus loin dans cette étude[32].

Deuxièmement, le résultat peut être très similaire à une œuvre figurant dans les données d'entraînement sans pour autant en être une reproduction mot pour mot. Cela pourrait constituer une atteinte au titre, par exemple, de la doctrine de la similitude substantielle, un critère bien connu dans plusieurs législations nationales sur le droit d'auteur.

Troisièmement, le résultat peut être entièrement nouveau, en ce sens qu'aucune œuvre particulière figurant dans les données d'entraînement ne peut être identifiée comme sa source. Un modèle génératif est capable de produire du contenu qui ne correspond à aucun exemple d'apprentissage en particulier. En effet, dans la plupart des cas, c'est exactement ce qu'il fait. Dans de tels cas, le lien entre le résultat et une œuvre protégée spécifique est atténué. La question de savoir si le résultat reste "fondé sur" les données d'entraînement, au sens où l'entendrait le droit d'auteur, a été soumise aux tribunaux de plusieurs juridictions, mais aucun consensus ne s'est encore dégagé à ce sujet.

La première partie recense donc les utilisations d'œuvres protégées, ainsi que les questions relatives à leur conservation ou à leur reproduction, à différentes étapes du développement et du fonctionnement d'un modèle d'IA générative : dans le corpus constitué au départ, dans les opérations de stockage et de traitement appliquées au corpus pendant l'entraînement, dans le modèle entraîné lui-même, et dans les résultats produits après son déploiement. À chaque étape, les principaux droits exclusifs reconnus par la législation nationale sur le droit d'auteur et par les traités internationaux peuvent être en jeu. La question de savoir si chaque reproduction peut constituer une atteinte présumée (sous réserve d'une exception ou d'une limitation), de même que la question de savoir quels mécanismes existent ou pourraient être élaborés pour autoriser les utilisations concernées, sont abordées dans les parties suivantes.

---

30 Lorsque le résultat généré par un modèle LLM présente des similitudes avec une œuvre protégée préexistante que le modèle n'a pas utilisée pour son apprentissage et à laquelle il n'a pas eu accès d'une autre manière, il est beaucoup moins évident de conclure qu'une "copie" a été réalisée.

31 La reproduction mot pour mot par des modèles linguistiques a été attestée dans des études universitaires et dans des dossiers judiciaires; la reproduction de photographies et d'œuvres d'art reconnaissables par des modèles de génération d'images a été attestée dans le cadre des modèles de diffusion. Le nombre d'exemples documentés n'a pas d'incidence directe sur la qualification juridique d'une production donnée, celle-ci dépendant de la législation nationale applicable et des faits propres à chaque affaire.

32 En ce qui concerne la relation entre la consigne, le modèle et le résultat, ainsi que la question encore en suspens de la répartition des responsabilités entre le développeur et l'utilisateur, voir la partie III ci-dessous.

DEUXIÈME PARTIE. AUTORISATION, CONCESSION DE LICENCE ET RÉMUNÉRATION

La première partie de cette étude a permis de recenser, sur le plan technique, les cas dans lesquels les principaux droits exclusifs reconnus par la législation nationale sur le droit d'auteur et par les traités internationaux relatifs au droit d'auteur peuvent intervenir dans l'élaboration et l'exploitation d'un modèle d'IA générative. La présente partie passe de ces droits aux mécanismes permettant de les exercer, d'y renoncer ou de les remplacer dans le contexte de l'IA. La question de savoir si les utilisations en cause sont licites sans autorisation et, dans le cas contraire, celle de savoir comment les développeurs peuvent mettre leurs activités en conformité avec la loi et comment les titulaires de droits doivent être indemnisés dans le cadre de ce processus, renvoient à un ensemble unique de principes juridiques : le droit applicable aux licences de droit d'auteur, individuelles et collectives, volontaires et légales, ainsi que les exceptions et limitations qui conditionnent la nécessité même d'une licence.

Cette partie est structurée comme suit : La section A examine les principaux canaux par lesquels les développeurs en IA se procurent actuellement les œuvres protégées par le droit d'auteur sur lesquelles ils entraînent leurs modèles, ainsi que le statut juridique de chacun de ces canaux. La section B aborde les principaux enjeux politiques qui déterminent actuellement les pratiques en matière de licences : les droits concernés par l'entraînement des modèles, les licences volontaires, les licences collectives sous leurs trois formes principales, la situation en ce qui concerne l'exception européenne relative à l'exploration de textes et de données ainsi que le débat scientifique sur sa pertinence pour l'entraînement des modèles d'IA, et enfin les mécanismes de rémunération. Les pays cités fournissent des exemples d'approches adoptées dans différentes régions du monde; ils ne constituent pas l'objet de l'analyse et la présente étude n'est pas exhaustive.

SECTION A. OPTIONS DONT DISPOSENT ACTUELLEMENT LES DÉVELOPPEURS POUR SE PROCURER DES ÉLÉMENTS SUSCEPTIBLES D'ÊTRE PROTÉGÉS PAR LE DROIT D'AUTEUR À DES FINS DE FORMATION

1. Les cinq modes d'acquisition

Dans la pratique, les développeurs en IA se procurent leurs données d'entraînement selon l'une des cinq méthodes suivantes, ou une combinaison de celles-ci. L'une d'entre elles consiste à utiliser des ensembles de données ouverts ou sélectionnés, créés par des tiers et proposés, parfois gratuitement, parfois moyennant paiement. Un exemple marquant est celui de Common Crawl, un référentiel de pages Web mis à jour en continu, exploré et mis à disposition par une fondation à but non lucratif, et qui sert de base, directement ou indirectement, aux corpus d'entraînement textuels de la plupart des grands modèles linguistiques. LAION-5B et ses prédécesseurs constituent l'équivalent des modèles de génération d'images[33]. Ces ensembles de données sont eux-mêmes le résultat d'opérations de collecte automatisées à grande échelle et peuvent soulever des questions de droit d'auteur liées à la collecte sous-jacente. La situation juridique dépendra des modalités d'obtention des données sources et de la législation applicable aux utilisations ultérieures de l'ensemble de données. Une équivalence dans le secteur public à ces ensembles de données sélectionnés commence à voir le jour : en République de Corée, le système de licences de données publiques (Gonggong Nuri) a été mis à jour afin d'inclure une catégorie de licence en vertu de laquelle les ensembles de données

---

[33] Common Crawl est présenté à l'adresse https://commoncrawl.org. L'ensemble de données LAION-5B (utilisé pour la conception de Stable Diffusion et d'autres modèles de génération d'images) et son prédécesseur, LAION-400M, sont des ensembles de données image-texte accessibles au public, compilés par le Large-scale Artificial Intelligence Open Network. L'ensemble de données lui-même ne contient pas les images sous-jacentes, mais des paires composées d'URL et de descriptions textuelles; les images sont récupérées au moment de l'entraînement à partir de leurs emplacements d'origine. Concernant la qualification juridique des opérations de compilation d'ensembles de données en tant qu'"exploration de textes et de données", voir l'affaire *Kneschke* c. *LAION e.V.*, Tribunal régional de Hambourg, 27 septembre 2024 (310 O 227/23), et, en appel, l'affaire *Kneschke* c. *LAION e.V.*, Cour d'appel de Hambourg, 10 décembre 2025 (5 U 104/24).

publiques comportent une mention expresse indiquant que leur utilisation pour l'entraînement de modèles d'IA est autorisée[34].

Une autre source est ce qu'on appelle les "bibliothèques fantômes", de vastes référentiels non autorisés de livres, d'articles, de musique et d'autres œuvres, que plusieurs modèles largement utilisés auraient, selon certaines informations ou allégations, utilisés, soit directement, soit par l'intermédiaire d'ensembles de données qui les incluent. Les conclusions relatives au piratage aux États-Unis, évoquées dans la section A de la partie III, ci-dessous, illustrent les risques juridiques que comporte le recours à un tel référentiel.

À titre de comparaison, certains tribunaux ont établi une distinction entre l'acquisition illicite d'œuvres et la question distincte de savoir si l'utilisation de ces œuvres à des fins de formation relève d'une exception ou d'une limitation. Dans l'une des décisions rendues aux États-Unis d'Amérique évoquées ci-dessous, le téléchargement et la conservation de copies pirates ont été jugés ne pas relever de l'usage loyal, tandis que l'entraînement d'un modèle à partir de copies acquises légalement des mêmes œuvres a été considéré comme une utilisation transformatrice et loyale. La légalité de l'acquisition et celle de la formation sont, en ce sens, analytiquement indépendantes l'une de l'autre. Cette distinction est importante car, au sein d'un même modèle, les données d'entraînement peuvent provenir à la fois de sources licites et illicites, ce qui complique les questions de licence et d'application de la loi. Un développeur peut donc avoir besoin à la fois d'une acquisition légale du corpus et d'une base légale pour l'entraînement lui-même, sous réserve de la ou des législations nationales applicables.

La troisième approche concerne les licences individuelles et collectives. Un développeur peut acheter un accès à un corpus sélectionné ou négocier cet accès auprès d'un éditeur, d'un organe de presse, d'un organisme de radiodiffusion, d'une photothèque, d'un gestionnaire de base de données universitaire ou de tout autre titulaire de droits institutionnel. La concession de licences est devenue une pratique courante au cours des trois dernières années, les principaux développeurs en IA ayant conclu des accords avec de grands organes de presse et des banques d'images en Europe, en Amérique et en Asie[35]. La particularité structurelle de la concession de licences réside dans le fait que le développeur obtient une chaîne de titres faisant autorité sur l'œuvre sous licence, du moins en ce qui concerne les droits propres au donneur de licence. Les limites de cette chaîne sont abordées dans la section B.2 ci-dessous. Dans certains cas, le contrat prévoit la fourniture du contenu.

La quatrième méthode, qui constitue l'une des principales en termes de volume, consiste à collecter directement des données sur le Web ouvert grâce à des techniques d'exploration du Web et de moissonnage de données sur le Web mises en œuvre par le développeur lui-même, et non par un tiers. Les procédures techniques sont décrites dans la section A.2 de la première partie de la présente étude. L'ampleur de cette recherche automatisée est considérable : selon certaines estimations, les agents automatisés représenteraient désormais la majeure partie du trafic sur l'Internet au niveau mondial[36]. L'exploration de sites Web est souvent effectuée sans le consentement préalable des titulaires des droits sur les contenus récupérés, en s'appuyant sur l'une des trois bases juridiques suivantes : soit les contenus relèvent du domaine public, soit ils ont fait l'objet d'une licence valablement accordée au public

---

[34] Il s'agit du système Gonggong Nuri (licence libre du Gouvernement de la République de Corée), dans le cadre duquel les institutions publiques mettent à disposition des données publiques en vue de leur réutilisation selon des catégories de licences normalisées; le cadre actualisé ajoute une catégorie autorisant expressément l'utilisation de ces données pour l'entraînement de modèles d'IA.

[35] En ce qui concerne les accords de licence bilatéraux entre les développeurs en IA et les éditeurs, voir (entre autres) les accords conclus par OpenAI avec l'Associated Press, News Corp, Le Monde, Prisa Media, Axel Springer, Vox Media, The Atlantic et Hearst; par Google avec Reddit et l'Associated Press; et par Anthropic avec divers éditeurs. Les informations relatives à ces opérations sont divulguées avec plus ou moins de détails, et leurs conditions financières relèvent du secret commercial.

[36] Voir Imperva, "Bad Bot Report 2025" (douzième édition, 2025), qui indique que le trafic automatisé représentait 51% de l'ensemble du trafic Internet en 2024, dépassant pour la première fois depuis 10 ans le trafic généré par les utilisateurs humains, les "bad bots" représentant quant à eux 37% de ce trafic.

à des conditions permissives; soit l'opération relève d'une exception ou d'une limitation prévue par la législation applicable en matière de droit d'auteur (dans l'Union européenne, l'exception relative à l'exploration de textes et de données prévue à l'article 4 de la directive CDSM; aux États-Unis d'Amérique et, en vertu de l'article 35-5 de la loi sur le droit d'auteur, en République de Corée, une doctrine générale d'usage loyal; à Singapour et au Japon, des dispositions légales spécifiques); ou que l'opération n'engage en rien le droit de reproduction, car les copies réalisées sont temporaires ou transitoires, ou encore que, dans certains ressorts juridiques, ces copies relèvent d'une exception expresse concernant les actes de reproduction temporaires (comme en vertu de l'article 5.1) de la directive "InfoSoc" dans l'Union européenne). Ces théories ne s'excluent pas mutuellement et un développeur peut recourir à différentes théories pour différentes parties du corpus.

Le cinquième type est celui des données synthétiques : il s'agit de données générées par un modèle d'IA existant et utilisées pour entraîner ou affiner un autre modèle. Les données synthétiques constituent un outil de plus en plus utilisé, le plus souvent dans les dernières étapes de l'élaboration d'un modèle, notamment pour le réglage fin supervisé et la modélisation des récompenses[37]. Du point de vue du droit d'auteur, les données synthétiques soulèvent une question différente. Si le modèle qui l'a généré a été entraîné sur des œuvres protégées, les résultats synthétiques reproduisent-ils ou adaptent-ils également des expressions protégées? La question n'a pas encore été soumise à un tribunal de manière à obtenir une réponse faisant autorité.

2. Accords de collaboration

Enfin, outre ces cinq grands axes, on observe dans certains ressorts juridiques des formes de collaboration organisée entre les développeurs en IA et les titulaires de droits institutionnels ou publics, particulièrement pertinentes pour certaines applications sectorielles. La Chine en est un exemple. L'élaboration de grands modèles à domaine vertical a donné lieu à des collaborations entre, d'une part, un développeur de modèles d'IA à usage général et, d'autre part, des chaînes de radio et de télévision de Hong Kong (pour les données vocales en cantonais); entre des développeurs et des éditeurs en ligne disposant d'un vaste répertoire d'auteurs sous contrat (pour les romans numériques); entre des développeurs et de grandes banques d'images détenant des milliards d'images par l'intermédiaire de transferts ou de licences exclusives (pour les données d'entraînement visuelles); et entre des développeurs et des gestionnaires de bases de données universitaires travaillant conjointement à la création de modèles spécialisés. On estime que ce phénomène est très répandu, même si les accords eux-mêmes ne sont souvent pas désignés comme des licences de droit d'auteur et que leur champ d'application précis n'est pas toujours précisé. Comme nous le verrons dans la section B.2 ci-dessous, la limite structurelle de cette approche réside dans le fait que l'éditeur, le diffuseur ou le dépositaire ne détient généralement que certains droits qui lui ont été concédés par l'auteur d'origine, et qu'il n'est pas certain que ces droits couvrent l'entraînement des IA.

[37] Les données synthétiques sont des données générées par un modèle d'IA existant et utilisées pour l'entraînement ou le réglage fin d'un autre modèle. Cela soulève des questions spécifiques en matière de droit d'auteur lorsque le modèle générateur a lui-même été entraîné à partir d'œuvres protégées. Concernant son utilisation croissante aux dernières étapes du développement de modèles, voir R. Liu *et al*., "Best Practices and Lessons Learned on Synthetic Data", arXiv:2404.07503 (2024), présenté lors de la première conférence sur la modélisation du langage.

SECTION B. PRINCIPAUX ÉLÉMENTS DE LA POLITIQUE

1. Droits concernés par l'entraînement des modèles d'IA générative

a) Le droit de reproduction

Les opérations techniques décrites dans la première partie relèvent, à première vue, d'un certain nombre de droits exclusifs reconnus par les législations nationales sur le droit d'auteur et par les traités internationaux relatifs au droit d'auteur. Le droit de reproduction est au cœur de l'analyse[38]. La constitution d'un corpus d'entraînement implique la création de copies; le stockage et le traitement de ce corpus au cours des différentes étapes de l'entraînement impliquent d'autres copies, certaines permanentes et d'autres temporaires; certains aspects du modèle entraîné, selon l'analyse technique exposée dans la section C.2 de la première partie, peuvent eux-mêmes être considérés comme des copies des œuvres sur lesquelles le modèle a été entraîné. Chacune de ces opérations relève, à première vue, du droit de reproduction tel qu'il est reconnu à l'article 9.1) de la Convention de Berne, à l'article 1.4) du Traité de l'OMPI sur le droit d'auteur, à l'article 7 du WPPT et du Traité de Beijing, ainsi que dans les dispositions correspondantes des législations nationales.

Le traitement des copies temporaires et temporaires mérite une analyse distincte, car deux techniques juridiques distinctes aboutissent à des résultats pratiques similaires : dans certains ressorts juridiques, ces copies ne sont pas considérées comme relevant du champ d'application du droit de reproduction, tandis que dans d'autres, elles font l'objet d'une exception expresse. La Chine constitue l'exemple principal de cette première technique : la reproduction temporaire a toujours été considérée comme ne relevant pas du champ d'application de ce droit. Dans la pratique judiciaire chinoise, les reproductions temporaires ainsi exclues sont les copies éphémères créées dans la mémoire vive ou dans une mémoire cache, qui sont strictement nécessaires au traitement des données; il n'existe pas de position établie quant à la durée exacte pendant laquelle une copie peut subsister tout en étant considérée comme temporaire. L'article 10.1)5) de la loi sur le droit d'auteur définit la reproduction comme le droit de produire une ou plusieurs copies d'une œuvre par des moyens tels que l'impression, la photocopie, la gravure, l'enregistrement, l'enregistrement vidéo, la duplication, la photographie ou la numérisation[39], et la délégation chinoise, parmi d'autres, a voté contre la déclaration commune adoptée lors de la conférence diplomatique, qui aurait clairement inclus le stockage numérique temporaire dans le champ d'application de ce droit[40]. Le règlement relatif à la protection du droit de communication au public sur les réseaux informatiques (2006) ne comportait pas de disposition reconnaissant la reproduction temporaire comme un acte de reproduction, et les tribunaux chinois ont appliqué cette interprétation dans des affaires portant sur le transcodage et d'autres opérations numériques transitoires similaires. Dans cette optique, l'acte d'entraînement en lui-même, dans la mesure où il n'implique qu'un accès temporaire à des œuvres stockées en mémoire ou dans le cache, qui disparaît une fois la séance d'entraînement terminée, pourrait ne pas relever du droit de reproduction en vertu de la législation chinoise, bien que la constitution de la base de données sous-jacente contenant ces œuvres puisse certainement en relever. Dans d'autres ressorts juridiques, on parvient au même résultat pratique par l'intermédiaire d'une exception expresse : dans l'Union européenne, l'article 5.1) de la directive "InfoSoc" prévoit une exception strictement délimitée pour les copies temporaires et accessoires, plutôt que de les exclure d'emblée du champ d'application du droit. Les copies en question sont, sur le fond, identiques dans les deux systèmes; la différence

---

38 Voir la première partie, sections A.1 et A.2.

39 Loi sur le droit d'auteur de la République populaire de Chine, article 10.1)5).

40 Voir le document de l'OMPI CRNR/DC/101, compte rendu analytique, séance plénière, paragraphes 497 à 502. La déclaration commune relative à l'article 1.4) du WCT a néanmoins été adoptée par la conférence diplomatique; la Chine a voté contre.

réside dans la technique juridique employée[41]. Dans la plupart des autres ressorts juridiques examinés ici, l'acte d'entraînement est généralement considéré comme relevant du droit de reproduction, et la question qui se pose alors porte sur l'existence d'une exception ou d'une limitation.

b) Le droit d'adaptation et le droit aux œuvres dérivées

Outre le droit de reproduction, trois autres droits sont ou peuvent être en jeu. Le droit d'adaptation, reconnu par l'article 12 de la Convention de Berne et par les dispositions nationales correspondantes, est invoqué lors de la phase de création de l'œuvre dérivée, car la création de celle-ci est considérée comme entraînant une transformation ou une modification des œuvres sous-jacentes. Par exemple, la législation brésilienne mentionne expressément le droit d'adaptation parmi les formes d'utilisation nécessitant une autorisation préalable et expresse[42], et le Secrétariat au droit d'auteur et à la propriété intellectuelle du Ministère de la culture a estimé que la production générée relevait du droit d'adaptation ainsi que du droit de reproduction. La situation en droit canadien, français et allemand est globalement compatible.

c) Communication au public et mise à disposition

Le droit de mise à la disposition du public, prévu à l'article 8 du Traité de l'OMPI sur le droit d'auteur et aux articles 10 et 14 du Traité de l'OMPI sur les interprétations et exécutions et les phonogrammes, s'applique principalement au stade de la production, lorsque le modèle utilisé renvoie du contenu à l'utilisateur, si ce contenu contient une copie d'une expression protégée. Au Brésil, qui n'est pas partie contractante du WCT ni du WPPT, le Superior Tribunal de Justiça a jugé en 2017, dans l'affaire REsp 1.559.264-RJ, que la diffusion de musique en continu sur l'Internet relève du droit de communication au public, qui englobe ce que d'autres ressorts juridiques pourraient qualifier de droit de mise à disposition[43]. Cela s'inscrit dans la lignée de la "solution globale" adoptée lors de la conférence diplomatique de 1996, en vertu de laquelle le droit de mise à disposition peut être mis en œuvre par l'intermédiaire du droit de communication au public, du droit de distribution ou d'une combinaison des deux. L'application d'un raisonnement analogue aux résultats d'un modèle d'IA déployé n'a pas encore été examinée par les tribunaux brésiliens[44]. Les artistes interprètes et les producteurs de phonogrammes détiennent des droits parallèles de mise à disposition en vertu des articles 10 et 14 du WPPT, mentionnés ci-dessus, de sorte que l'analyse dépasse le cadre du droit d'auteur.

d) Droit moral

Le droit moral peut également être mis en jeu. C'est le droit à l'intégrité qui est le plus souvent invoqué. Au Brésil, la fragmentation des œuvres au cours du processus d'entraînement est considérée comme une atteinte potentielle au droit à l'intégrité, et l'absence de mention des

---

41 Article 5.1) de la directive 2001/29/CE du Parlement européen et du Conseil du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information [2001] JO L 167/10 ("directive InfoSoc"). Cette exception s'applique aux actes de reproduction temporaires qui sont éphémères ou accessoires, qui constituent une partie intégrante et essentielle d'un processus technologique, dont le seul but est de permettre une transmission sur un réseau entre des tiers par un intermédiaire ou dans le cadre d'une utilisation licite, et qui n'ont aucune valeur économique en soi.

42 Loi brésilienne sur le droit d'auteur (loi n° 9 610/1998), article 29, points III et IV (adaptation, arrangement musical, traduction et autres transformations). La position du Ministère de la culture a notamment été exposée dans une note technique adressée à l'Autorité nationale de protection des données en octobre 2024 concernant l'entraînement des modèles d'un grand développeur, en invoquant l'article 29.III) (transformation en vue de la création d'œuvres dérivées); voir le rapport disponible à l'adresse https://nucleo.jor.br/curtas/2024/10/10/ministerio-cultura-prejuizos-direitos-ia-meta/, consulté le 7 juillet 2026.

43 Tribunal supérieur de justice, REsp 1.559.264-RJ (2017) (jugement selon lequel la diffusion de musique en continu sur l'Internet constitue une forme d'exécution publique et relève des droits de communication au public, et selon lequel l'ECAD est habilitée à percevoir des redevances à ce titre).

44 Loi brésilienne sur le droit d'auteur (loi n° 9 610/1998), article 5.V) (définition de la "communication au public") et article 29.VII) (diffusion par tout système permettant à l'utilisateur de sélectionner l'œuvre).

auteurs des documents utilisés comme une atteinte potentielle au droit d'attribution[45]. La loi canadienne sur le droit d'auteur reconnaît à la fois le droit à l'intégrité et le droit à la mention de l'auteur comme des prérogatives attachées au droit moral de l'auteur[46]. Dans plusieurs ressorts juridiques régis par le droit d'auteur, la question de savoir si l'utilisation d'une œuvre dans un corpus d'entraînement, sans mention de la source et d'une manière susceptible de générer des résultats perçus comme des versions dénigrantes ou déformées de l'original, constitue une atteinte au droit moral a été soulevée dans la littérature universitaire, mais n'a pas encore fait l'objet d'une réponse faisant autorité.

Enfin, au sein de l'Union européenne et dans certains ressorts juridiques ayant adopté une protection similaire, le droit *sui generis* relatif aux bases de données protégées par la directive 96/9/CE peut être invoqué lorsque le corpus d'entraînement est constitué par l'extraction et la réutilisation de parties substantielles d'une base de données protégée[47]. Ce droit est indépendant du droit d'auteur et obéit à ses propres critères d'"investissement substantiel" et de "partie substantielle" (pour le "droit d'extraction"). Son champ d'application à la constitution de corpus d'entraînement pour l'IA reste encore à définir, notamment en ce qui concerne les bases de données d'informations juridiques, les agrégateurs de revues scientifiques et les vastes bibliothèques d'images.

Les droits susceptibles d'être concernés par l'entraînement d'un modèle d'IA générative pourraient donc aller au-delà du droit de reproduction. La question de savoir si une licence bilatérale ou collective autorise les utilisations en cause dépend donc des droits qu'elle couvre; une licence limitée à la reproduction peut ne pas couvrir les droits liés à l'adaptation, à la mise à disposition, au droit moral ou, le cas échéant, au droit *sui generis* sur les bases de données. Un autre point, abordé dans la partie III dans le cadre de l'application transfrontalière, est que l'ensemble des droits et la portée précise de chacun d'entre eux varient d'un ressort juridique à l'autre, de sorte qu'un accord de licence destiné à régler les droits dans un ressort juridique peut exposer le développeur à des risques dans un autre[48].

e) Information sur le régime des droits et mesures antineutralisation

En vertu des articles 11 et 12 du WCT, des articles 18 et 19 du WPPT et des articles 15 et 16 du Traité de Beijing, les titulaires de droits bénéficient d'une protection juridique adéquate et de recours juridiques efficaces contre la neutralisation des mesures techniques efficaces qu'ils utilisent dans le cadre de l'exercice de leurs droits en vertu de ces instruments et de la Convention de Berne, ainsi qu'à des recours juridiques adéquats et efficaces contre toute personne qui accomplit sciemment les actes suivants, en sachant, ou, en ce qui concerne les recours civils, en ayant des motifs raisonnables de savoir, que cela induira, permettra, facilitera ou dissimulera une atteinte à tout droit couvert par ces traités ou par la Convention de Berne :

i) supprimer ou modifier, sans y être habilitée, toute information relative au régime des droits se présentant sous forme électronique;

ii) la distribution, l'importation en vue de la distribution, la diffusion ou la communication au public, sans autorisation, d'œuvres ou de copies d'œuvres, de phonogrammes ou d'exécutions

---

45 Loi brésilienne sur le droit d'auteur (loi n° 9 610/1998), article 24, points I à IV. Les artistes interprètes jouissent également, en vertu de l'article 92, du droit moral d'intégrité et de paternité sur leurs prestations.

46 Loi sur le droit d'auteur, L.R.C. Loi de 1985, chap. C-42, articles 14.1 et 28.1.

47 Directive 96/9/CE du Parlement européen et du Conseil du 11 mars 1996 concernant la protection juridique des bases de données, JO L 77 page 20, article 7 (le droit *sui generis*). Voir également CJUE, C-203/02, *British Horseracing Board* c*. William Hill*, ECLI:EU:C:2004:695, concernant les conditions relatives au critère de l'investissement substantiel.

48 Voir *Toronto Star Newspapers Ltd* c*. OpenAI Inc.*, 2025 ONSC 6217 (juge Kimmel, 7 novembre 2025) (rejetant la requête d'OpenAI relative à la compétence en raison de l'existence d'un lien réel et substantiel avec l'Ontario).

et (en vertu des instruments respectifs) en sachant que des informations relatives à la gestion électronique des droits ont été supprimées ou modifiées sans autorisation.

À notre connaissance, aucune action en justice n'avait abouti, à la date de rédaction du présent article, contre des entreprises concernant l'entraînement des modèles d'IA.

2. Concessions de licence volontaires (collectives et individuelles)

a) Concession de licences individuelles et collectives pour les utilisations de l'IA

La concession de licences volontaires consiste en la négociation d'une autorisation entre un développeur d'IA et un titulaire de droits ou un représentant de plusieurs titulaires de droits, notamment une organisation de gestion collective ou un autre intermédiaire, sans qu'il y ait de mandat légal. C'est la règle par défaut dans tous les ressorts juridiques étudiés.

Deux caractéristiques structurelles déterminent la forme que prend, dans la pratique, la concession de licences volontaires. La première est l'asymétrie entre, d'une part, le petit nombre de grands titulaires institutionnels de droits capables de négocier avec les développeurs à des conditions commercialement intéressantes et, d'autre part, le nombre très important d'auteurs individuels et de petits titulaires de droits pour lesquels une telle négociation est souvent impossible en raison des coûts de transaction, d'un manque d'expertise ou d'autres facteurs. Les principaux accords bilatéraux conclus au cours des trois dernières années ont été signés entre des développeurs d'IA et de grands éditeurs, des organes de presse, des chaînes de télévision et des banques d'images; il n'existe aucune infrastructure comparable pour les auteurs, photographes, illustrateurs ou musiciens indépendants, qui dépendent souvent des accords conclus par leurs éditeurs et autres titulaires de licence pour obtenir une part de la valeur générée.

L'ampleur de certains de ces accords est désormais considérable. Un contrat de cinq ans conclu entre un grand éditeur de presse et un développeur d'IA aurait dépassé les 250 millions de dollars É.-U.[49], et la concession de licences de contenu à des fins informatiques est devenue une source de revenus importante pour les grandes banques d'images : les résultats annuels 2025 de Shutterstock, Inc. font état de 203,3 millions de dollars É.-U. de chiffre d'affaires dans les secteurs "Données, distribution et services", ce qui représente environ 20,5% de son chiffre d'affaires total de 989,9 millions de dollars É.-U.; Getty Images a quant à elle annoncé séparément un chiffre d'affaires annuel de 981,3 millions de dollars É.-U. pour 2025. Les deux sociétés ont annoncé une fusion entre égaux le 7 janvier 2025; Getty Images a annoncé le 30 juin 2026 qu'elle mettrait fin à l'accord de fusion après que l'Autorité britannique de la concurrence et des marchés (CMA) a subordonné son autorisation à la cession des activités éditoriales de Shutterstock. L'accord aurait pris fin le 7 juillet 2026[50]. Le caractère sélectif de ce marché est illustré par la presse française : en mars 2024, Le Monde a conclu un accord avec OpenAI, qu'il a qualifié de premier accord entre un média français et un grand

---

[49] Selon le Wall Street Journal, l'accord de cinq ans conclu entre News Corp et OpenAI s'élèverait à plus de 250 millions de dollars É.-U. Les conditions financières de ces accords sont généralement confidentielles, et le montant est relayé par la presse plutôt que confirmé publiquement par les parties. Voir "OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over $250 Million" (The Wall Street Journal, 22 mai 2024) <https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba> consulté le 3 juillet 2026; "Sam Altman's OpenAI signs content agreement with News Corp" (Reuters, 22 mai 2024) <https://finance.yahoo.com/news/sam-altmans-openai-signs-content-201808782.html> consulté le 3 juillet 2026.

[50] Shutterstock, "Shutterstock Reports Full Year 2025 and Fourth Quarter Financial Results" (17 février 2026) <https://investor.shutterstock.com/node/14551/pdf> consulté le 3 juillet 2026; Getty Images Holdings, Inc, "Getty Images Reports Fourth Quarter and Full Year 2025 Results" (16 mars 2026) <https://investors.gettyimages.com/node/15726/pdf> consulté le 3 juillet 2026; Getty Images Holdings, Inc et Shutterstock, Inc, "Getty Images and Shutterstock to Merge, Creating a Premier Visual Content Company" (7 janvier 2025) <https://www.sec.gov/Archives/edgar/data/1898496/000110465925001732/tm252186d1_ex99-1.htm> consulté le 3 juillet 2026. Concernant l'annulation de la fusion, voir l'Autorité de la concurrence et des marchés, "Getty abandons Shutterstock merger plans" (juillet 2026) <https://www.gov.uk/government/news/getty-abandons-shutterstock-merger-plans>, consulté le 31 juillet 2026.

développeur d'IA générative[51]. Plus tard dans l'année, l'APIG a cherché à engager des discussions collectives avec OpenAI, même si les informations disponibles indiquent que ces discussions n'ont pas abouti à un contrat de licence élargi à cette époque[52]. Une réponse à cette asymétrie commence à voir le jour sous la forme d'intermédiaires qui regroupent les droits des petits éditeurs : un fournisseur d'informations économiques indique que plusieurs milliers d'éditeurs de son réseau, issus d'un vivier bien plus vaste, ont choisi d'accorder une licence d'utilisation de leur contenu à des fins d'IA[53], et le développeur d'un moteur de réponses sous licence fait état d'un regroupement comparable de publications participantes[54].

Le deuxième point concerne la question de la chaîne de titres. Dans de nombreux ressorts juridiques, lorsque l'éditeur, le diffuseur ou le dépositaire s'est vu accorder certains droits par l'auteur concerné au titre d'un contrat de licence ou de cession conclu avant l'avènement de l'IA générative, la question de savoir si ces droits s'étendent à l'utilisation de l'œuvre dans le cadre de l'entraînement d'un modèle d'IA est à la fois juridiquement incertaine et non prévue par le contrat. Cette incertitude n'est pas propre à l'IA : des questions similaires se posent dans le cadre des contrats entre auteurs et éditeurs de musique, entre artistes interprètes ou exécutants et producteurs de phonogrammes, ainsi qu'entre producteurs audiovisuels et artistes engagés pour une production, chaque fois qu'un mode d'exploitation n'avait pas été envisagé au moment de la conclusion du contrat; ce qui distingue le contexte de l'IA, c'est l'ampleur du phénomène et la difficulté pratique que représente la renégociation. En Chine, par exemple, la plupart des auteurs accordent à leurs éditeurs le droit exclusif de publier leurs œuvres sur papier et, de plus en plus, sous forme numérique. Il s'agit des droits de reproduction de l'œuvre sur papier, de distribution d'exemplaires papier, ainsi que de reproduction et de distribution d'exemplaires numériques. Aucun de ces types d'utilisation n'est naturellement interprété comme incluant l'utilisation de l'œuvre dans l'entraînement d'un modèle d'IA. Les fournisseurs de corpus et les développeurs d'IA qui obtiennent des licences de droit d'auteur auprès des éditeurs estiment donc que la légalité des activités autorisées par ces licences reste incertaine vis-à-vis des auteurs concernés. Même lorsque le développeur a obtenu une licence auprès du titulaire des droits au sein de l'institution, il y a un risque résiduel de poursuite de la part de l'auteur concerné. Selon le ressort juridique, le risque ne se limite pas au donneur de licence : celui-ci peut être tenu responsable d'un manquement au contrat ou à la loi, et le développeur, en tant que partie qui bénéficie en dernier ressort de l'utilisation, peut lui-même rester exposé à des actions en contrefaçon intentées par les auteurs concernés, malgré la licence. Dans le secteur audiovisuel, les contrats de licence types permettraient, semble-t-il, de parer à ce risque sur le plan contractuel : le donneur de licence déclare et garantit qu'il détient tous les droits nécessaires à l'utilisation à des fins de formation. En ce qui concerne les articles de presse rédigés par des journalistes et des rédacteurs avant l'entrée en vigueur, le 1er juin 2021, de l'amendement applicable à la loi chinoise sur le droit d'auteur, le droit d'auteur sur ces articles appartenait généralement au journaliste ou au rédacteur, et non à l'agence de presse. Cela signifiait que même une licence concédée par l'agence de presse

---

[51] Cet accord, conclu en mars 2024, a été présenté par Le Monde comme le premier accord entre un média français et un grand développeur de modèles d'IA générative. Voir "Le Monde signs artificial intelligence partnership agreement with OpenAI" (Le Monde, 13 mars 2024) <https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html> consulté le 3 juillet 2026.

[52] L'organisme en question est l'Alliance de la presse d'information générale (APIG) qui, conjointement avec le Syndicat des éditeurs de la presse magazine (SEPM), a officiellement demandé l'ouverture de négociations avec OpenAI en juin 2024; l'éditeur aurait répondu qu'il ne voyait pas la nécessité de conclure de nouveaux contrats de licence avec les éditeurs français à ce stade. Voir "Intelligence artificielle : un accord de partenariat entre Le Monde et OpenAI" (Le Monde, 13 mars 2024) <https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html> consulté le 3 juillet 2026.

[53] Selon Factiva, filiale de Dow Jones, d'ici la fin de 2025, environ 5 000 éditeurs, sur un réseau d'environ 30 000, auront choisi de concéder des licences d'utilisation de leurs contenus dans le cadre de systèmes d'IA générative. Voir Dow Jones, "Factiva Sees Growing Publisher Participation in AI Licensing" (Dow Jones, 2025) <https://www.dowjones.com>, consulté le 3 juillet 2026.

[54] ProRata.ai décrit son réseau, selon ses propres termes, comme regroupant "plus d'un millier de publications et de créateurs de contenu".

n'aurait pas autorisé l'utilisation de cet article pour l'entraînement d'un modèle d'IA. Des contraintes de ce type peuvent elles-mêmes entraver l'émergence d'un marché des licences volontaires, car une licence délivrée par un intermédiaire dont les droits sont incertains ajoute un maillon supplémentaire de contrefaçon potentielle à la chaîne de titres. Il y a également une dimension distributive : lorsque la concession de licences est assurée par des intermédiaires tels que les éditeurs, les diffuseurs et les dépositaires, l'asymétrie du pouvoir de négociation entre les auteurs ou artistes interprètes ou exécutants individuels et ces institutions fait que les avantages liés à la concession de licences peuvent ne pas profiter aux personnes dont les œuvres sont utilisées, une considération pertinente pour le choix entre les modèles de concession de licences abordés dans la section B.

La pratique chinoise montre toutefois que, même dans le cadre de ces contraintes structurelles, des contrats de licences volontaires à grande échelle voient le jour dans des contextes sectoriels et verticaux. Les chaînes de radio et de télévision de Hong Kong s'associent à des développeurs d'IA pour fournir des données vocales en cantonais de haute qualité. Les romans numériques sont proposés grâce à une collaboration entre des développeurs et des éditeurs en ligne qui disposent d'un vaste réseau d'auteurs sous contrat. Les images sont fournies grâce à une collaboration entre les développeurs et les principales banques d'images, qui les détiennent sur la base de cessions ou de licences exclusives octroyées par les photographes. Les articles scientifiques sont fournis grâce à une collaboration entre les développeurs et les exploitants de bases de données universitaires, qui deviennent parfois co-développeurs du modèle propre au domaine ainsi créé. Certains fournisseurs de corpus produisent également leurs propres supports de formation, en faisant appel à des photographes, des mannequins et des enseignants afin de céder ou de concéder sous licence les œuvres ainsi créées pour qu'elles soient utilisées dans des ensembles de données d'entraînement.

D'autres ressorts juridiques présentent des tendances similaires, bien que les dispositions institutionnelles y soient différentes. Au Brésil, le régime juridique applicable à la concession de licences volontaires suit les règles générales de la loi sur le droit d'auteur. L'obligation d'une autorisation préalable et expresse prévue à l'article 29, ainsi que l'interprétation restrictive des accords relatifs au droit d'auteur prévus aux articles 4, 31 et 49, impliquent que, même si les droits patrimoniaux avaient été entièrement cédés avant l'apparition de l'IA générative, l'utilisation de l'œuvre pour l'entraînement de l'IA serait contestable, car le contrat n'aurait pas pu prévoir spécifiquement une technologie qui n'existait pas encore. On retrouve des doctrines similaires d'interprétation stricte dans les systèmes juridiques français, italien et d'autres ressorts juridiques appliquant la doctrine du droit d'auteur[55].

Lorsque la loi prévoit un droit d'auteur explicite concernant les mises en scène, les représentations ou d'autres objets définis, le domaine de la concession de licences volontaires peut s'avérer plus simple, car les droits sous-jacents sont plus clairs. La situation peut se compliquer pour les catégories d'œuvres créées par de nombreux contributeurs, telles que les œuvres audiovisuelles et les phonogrammes, où l'incertitude juridique ou contractuelle entre les producteurs et les contributeurs individuels peut soit entraver la concession de licences volontaires, soit priver ces contributeurs des avantages qui en découlent. La législation sénégalaise, qui reconnaît expressément les mises en scène comme des œuvres protégées[56], en est un exemple, bien que l'application concrète de cette reconnaissance à l'entraînement de l'IA n'ait pas encore été mise à l'épreuve. Au Kenya, la loi sur le droit d'auteur (chapitre 130) reconnaît l'ensemble standard des droits exclusifs et prévoit la gestion collective du droit d'auteur par le biais d'organisations de gestion collective, mais l'infrastructure institutionnelle n'a

---

[55] Concernant la règle relative à la propriété, en vigueur avant 2021 dans la loi chinoise sur le droit d'auteur, pour les articles de presse rédigés par des journalistes et des rédacteurs, voir la loi sur le droit d'auteur de la République populaire de Chine (telle que modifiée en 2020, en vigueur à compter du 1er juin 2021), article 18 (œuvres créées dans le cadre d'un contrat de louage ou de services).

[56] Article 2 de la loi sénégalaise sur le droit d'auteur (loi n° 2008-09 du 25 janvier 2008). Des dispositions similaires figurent dans l'annexe VII de l'Accord de Bangui de l'OAPI (Acte révisé de 2015).

pas été conçue pour une utilisation informatique, et la concession de licences volontaires pour l'entraînement de l'IA n'est pas encore une pratique avérée. Au Canada, la concession de licences volontaires constitue le mécanisme par défaut et la négociation des contrats est l'activité centrale. Le principal litige en cours, *Toronto Star Newspapers Ltd* c*. OpenAI Inc.*, s'inscrit dans le cadre d'un différend portant sur une utilisation non autorisée plutôt que sur la concession de licences proprement dite; il est abordé dans la troisième partie de la présente étude.

L'une des caractéristiques récurrentes de la concession de licences volontaires, qu'elles soient individuelles ou collectives, réside dans l'adaptation des conditions de licence au profil d'utilisation des données du système pour lequel celles-ci sont recherchées, profil qui varie considérablement selon les catégories de systèmes d'IA générative actuellement utilisés. Lorsque les données ne sont utilisées qu'une seule fois, pour l'entraînement initial ou le réglage fin d'un modèle dont les pondérations sont ensuite définies, la structure tarifaire courante consiste en une redevance unique ou un partage des recettes pour un corpus défini. Dans les situations où le système a besoin d'un accès permanent au contenu sous licence lors de la génération de résultats, comme c'est le cas pour les systèmes de génération augmentée par récupération (RAG) et ceux qui effectuent une analyse continue de textes et de données sur une bibliothèque, cet accès ne se limite pas à une simple instance, mais constitue une relation continue. Les modèles de rémunération qui se sont développés dans la pratique en tiennent compte, qu'il s'agisse de taxes récurrentes ou de frais facturés par exploration ou par inférence. Outre le prix, les principaux points que les parties doivent régler concernent l'étendue de l'utilisation autorisée, la durée de la licence, son champ d'application territorial, les conséquences en cas de résiliation et la possibilité de céder le modèle à des tiers. Dans le contexte de l'IA, la question se pose de savoir si la résiliation implique que le développeur doit cesser d'utiliser un modèle déjà entraîné sur le contenu sous licence, le réentraîner ou détruire les copies sauvegardées, aucune de ces options ne s'inscrivant clairement dans le cadre des mesures correctives prévues pour les relations de licence antérieures. La résiliation illustre bien cette difficulté : dans le cadre d'une licence traditionnelle, la résiliation rétablit le contrôle du titulaire des droits sur les utilisations futures; or, un titulaire de droits qui résilie une licence relative à l'entraînement d'une IA ne peut, dans la pratique, contraindre le développeur à cesser d'utiliser ce que l'entraînement a déjà ancré dans le modèle, en particulier lorsque le catalogue concerné est restreint par rapport au corpus; la solution, si elle existe, devrait passer par un réentraînement ou des tentatives de "désapprentissage" de la machine, dont les limites sont soulignées dans la partie I. Même lorsque de telles interventions s'avèrent réalisables, elles entraînent des coûts opérationnels et de traitement, que les parties devront peut-être répartir dans le contrat lui-même, ce qui ajoute une difficulté supplémentaire à la négociation de ces licences. Lorsque ces pondérations de modèle sont communiquées ou vendues à des utilisateurs en aval, la question de la transmissibilité se pose : l'obligation contractuelle liée à une licence s'applique-t-elle également aux pondérations de modèle? Une question connexe, mais distincte, consiste à savoir si la rémunération en aval pourrait être calculée à partir du modèle lui-même, par exemple en tenant compte de la contribution du contenu concédé sous licence aux pondérations. Ces questions n'ont pas encore été tranchées dans les contrats qui ont été rendus publics et sont pertinentes pour le choix éventuel entre les modèles de rémunération évoqués à la section B.6 ci-dessous.

Dans le haut de gamme, la concession de licences volontaires dans les secteurs de la musique et de l'audiovisuel ont principalement pris la forme de partenariats individualisés qui associent la licence à des conditions commerciales allant bien au-delà de la simple utilisation des œuvres protégées. Au cours de l'année 2025, les grandes maisons de disques sont passées de la voie judiciaire à la concession de licences avec les principaux développeurs d'outils musicaux fondés sur l'IA : deux de ces développeurs ont trouvé un accord concernant les plaintes pour contrefaçon dont ils faisaient l'objet et ont accepté de mettre en place de nouvelles plateformes sous licence; l'un de ces accords s'est accompagné de l'acquisition, par le développeur, d'un service de découverte de concerts auprès de la maison de disques, tandis qu'un service de streaming a été convenu séparément avec les trois grandes maisons de disques et deux autres

titulaires de droits afin de développer des produits musicaux fondés sur l'IA selon un principe d'adhésion volontaire[57]. Ces accords ont été précédés par des déclarations de principe collectives émanant du secteur : des coalitions regroupant des organismes professionnels, des associations d'artistes et des entreprises technologiques ont défini des principes selon lesquels l'utilisation d'un répertoire protégé dans le cadre du développement de l'IA est subordonnée à une autorisation et à la concession d'une licence, et ont appelé à la transparence dans les pratiques d'entraînement[58]. Dans le secteur audiovisuel, un contrat de licence annoncé en décembre 2025, aux termes duquel un studio devait mettre à disposition plus de 200 de ses personnages sur une plateforme de création de vidéos et prendre une participation de 1 milliard de dollars É.-U. dans la société de développement, aurait été abandonné en mars 2026 lorsque cette dernière a retiré la plateforme[59]. Ces dispositions portent en partie sur la situation des produits, qui ne relève pas du champ d'application de la présente étude; elles ne sont mentionnées ici que pour illustrer les structures de concession de licences sur mesure et à forte valeur ajoutée qui voient actuellement le jour, ainsi que leur fragilité.

b) Les principaux modèles de licence collective (volontaire, étendue et obligatoire)

La forme la plus courante de concession de licences collectives passe par la gestion collective volontaire. Une organisation de gestion collective se voit confier, par contrat, la gestion de certains droits par ses membres; elle négocie et conclut des contrats de licence dans le cadre de son mandat; les recettes sont réparties entre ses membres. La plupart des ressorts juridiques dotés d'un système de droit d'auteur bien développé disposent d'organisations de gestion collective à caractère volontaire, mais les droits et les œuvres qu'ils gèrent varient considérablement. Les sociétés de gestion collective des droits de représentation ou d'exécution publique d'œuvres musicales en sont l'exemple le plus courant. En Chine, cinq organisations de gestion collective sont opérationnelles ou en passe de le devenir : la Music Copyright Society of China, la China Audio-Video Copyright Association, la China Written Works Copyright Society, l'Images Copyright Society of China et la China Film Copyright Association[60]. Les organisations de gestion collective ont publiquement fait part de leur volonté de concéder des licences pour l'entraînement de l'IA, mais à la date de la présente étude, aucun contrat de licence collective majeur concernant l'entraînement de l'IA n'avait été conclu en Chine[61]. Dans le contexte chinois, la raison structurelle tient au fait que le mandat de

---

[57] Les sociétés Universal Music Group et Warner Music Group ont chacune réglé leur litige contre Udio fin 2025 et conclu des contrats de licence pour de nouvelles plateformes; Warner Music Group a par ailleurs conclu un accord avec Suno, qui prévoyait notamment l'acquisition par Suno du service de découverte de concerts Songkick. Un autre développeur, KLAY Vision, a conclu des accords avec les principales maisons de disques et leurs filiales d'édition. Ce partenariat dans le domaine du streaming a été annoncé par Spotify, en collaboration avec Sony Music Group, Universal Music Group, Warner Music Group, Merlin et Believe, en octobre 2025.

[58] La campagne "Human Artistry", lancée en 2023 par une coalition de plus de 150 organisations, parmi lesquelles figurent des organismes majeurs de l'industrie musicale et des associations représentant les artistes, énonce sept principes fondamentaux <https://www.humanartistrycampaign.com> consulté le 25 août 2026; les Principles for Music Creation with AI (principes pour la création musicale avec l'IA), présentés par Roland Corporation et Universal Music Group et approuvés par plus de 50 organisations <https://aiformusic.info> consulté le 25 août 2026. Ces deux initiatives sont abordées et replacées dans un cadre plus large de conception responsable dans l'article de de Berardinis *et al.*, "RAIM: Operationalising Trustworthy AI in Music Generation", cité dans la partie I, note de bas de page n° 17, de la présente étude.

[59] L'accord entre The Walt Disney Company et OpenAI, annoncé le 11 décembre 2025, portait sur plus de 200 personnages et prévoyait un investissement en capital d'un milliard de dollars É.-U.; il aurait été abandonné à la suite du retrait par OpenAI de son produit de génération vidéo "Sora" le 24 mars 2026, aucun investissement n'ayant été finalisé.

[60] Les trois premières sont généralement désignées par les acronymes MCSC, AVCA et CWWCS; l'Images Copyright Society of China et la China Film Copyright Association sont les deux qui ne sont pas encore pleinement opérationnelles.

[61] Les principales organisations de gestion collectives chinoises ont publiquement fait part de leur volonté de concéder des licences pour l'entraînement de l'IA, mais aucun contrat de licence significatif n'a été conclu à ce jour.

l'organisation de gestion collective se limite aux droits de ses membres, et que le développeur est exposé à un risque résiduel concernant les œuvres dont les auteurs ne sont pas membres.

On observe des caractéristiques similaires en République de Corée, avec la Korea Music Copyright Association et des organismes du même type. Le caractère limité du répertoire collectif ne semble constituer qu'une partie de l'explication de l'échec des négociations en vue de la conclusion de licences. Dans le cadre des régimes réglementaires actuels, les développeurs peuvent être peu incités à demander spontanément des autorisations. Deux autres facteurs semblent expliquer pourquoi des accords sont néanmoins conclus aux États-Unis d'Amérique. D'une part, la structure du marché permet à un petit nombre de grands titulaires de droits et de fournisseurs de contenus de négocier directement avec les développeurs. D'autre part, le risque de poursuites, notamment de dommages-intérêts, incite à parvenir à un accord. Les contrats de licence relatifs au règlement des litiges décrits à la section B.2 en sont des exemples. Pour expliquer pourquoi la concession de licences n'a pas encore pris son essor en Chine et en République de Corée, il faut donc mettre en balance ces facteurs commerciaux et juridiques avec les limites du mandat collectif. À la date de la présente étude, le Ministère de la culture, des sports et du tourisme ainsi que la Commission coréenne du droit d'auteur de la République de Corée étaient engagés dans une démarche visant à améliorer le cadre réglementaire.

Sur le plan institutionnel, les États-Unis d'Amérique disposent d'un réseau dense d'organismes de gestion collective sectoriels. Les sociétés de gestion des droits d'exécution (ASCAP, BMI, SESAC, Global Music Rights et AllTrack) gèrent les droits d'exécution publique des œuvres musicales; le Mechanical Licensing Collective, créé en vertu de la loi de 2018 sur la modernisation de la musique (Music Modernization Act), gère les licences mécaniques générales pour les utilisations numériques des œuvres musicales[62]; SoundExchange gère les redevances légales relatives à l'exécution numérique des enregistrements sonores en vertu de l'article 114 de la loi sur le droit d'auteur; le Copyright Clearance Center, Inc. (CCC) gère les droits de reproduction des textes publiés; et l'Artists Rights Society (ARS) gère certains droits de reproduction dans le domaine des arts visuels. Chaque organisation exerce ses activités dans son secteur et selon les modalités négociées avec ses membres. Face à l'émergence de l'entraînement de l'IA comme utilisation des œuvres protégées, le CCC a étendu sa licence d'entreprise annuelle afin de couvrir certaines utilisations liées à l'IA et a annoncé son intention de proposer plusieurs licences supplémentaires spécifiques à l'IA[63].

Au-delà de la Chine et des États-Unis d'Amérique, ces deux dernières années ont vu plusieurs initiatives concrètes de la part des organisations de gestion collective et d'autres intermédiaires visant à intégrer la formation à l'IA dans leurs mandats de concession de licences, dont plusieurs n'ont été annoncées qu'à la fin de l'année 2025. En Suède, la STIM a lancé en septembre 2025 un contrat de licence collective associant la société de musique générée par IA Songfox et le fournisseur de services d'attribution Sureel, dont la technologie permet de remonter, à partir des créations générées par l'IA, jusqu'aux œuvres qui les ont influencées[64]. En Norvège, Kopinor a conclu en décembre 2025 un accord en vertu duquel la Bibliothèque nationale de Norvège est autorisée à utiliser des contenus de presse pour entraîner des

---

62 Orrin G. Hatch–Bob Goodlatte Music Modernization Act, Pub. L. No. 115-264, 132 Stat. 3676 (2018), codifié aux articles 115 et suivants du titre 17 du Code des États-Unis d'Amérique. Le Mechanical Licensing Collective est désigné en vertu de l'article 115.d)3) du titre 17 du Code des États-Unis d'Amérique.

63 Voir www.copyright.com/ai.

64 STIM, "STIM launches world's first collective AI music licence" (9 septembre 2025), communiqué par la CISAC, https://www.cisac.org/Newsroom/society-news/stim-launches-worlds-first-collective-ai-music-licence. La licence opt-in couvre à la fois l'entraînement des modèles et l'utilisation en aval des résultats générés par l'IA; la rémunération combine un paiement initial et une part des revenus du développeur, et le fournisseur de services d'attribution Sureel permet de remonter jusqu'aux œuvres qui ont influencé ces résultats.

modèles en norvégien, grâce à un financement de l'État norvégien[65]. Au Danemark, l'organisation de gestion collective des publications de presse danoises a conclu un partenariat de licence à adhésion volontaire (opt-in) avec ProRata AI pour l'utilisation des contenus de ses membres dans un moteur de recherche par extraction, également en décembre 2025, tandis qu'une autre organisation danoise, Tekst&Node, élabore actuellement un modèle de licence dans le cadre du dispositif danois de licences collectives étendues[66]. Aux Pays-Bas, l'association des éditeurs de presse privés, en collaboration avec l'agence de presse ANP, travaille depuis juillet 2025 avec des organismes publics à la mise au point d'un modèle de langage national, le GPT-NL[67]. En Espagne, le CEDRO a rejoint l'initiative "Really Simple Licensing" afin de développer un modèle de licence en matière d'IA destiné au marché espagnol[68]. Au Royaume-Uni, la Copyright Licensing Agency, en collaboration avec Publishers' Licensing Services et l'Authors' Licensing and Collecting Society, a mis au point une licence collective spécifique à l'IA générative, couvrant l'entraînement, le réglage fin et la génération augmentée par récupération[69]. Par ailleurs, les organisations de gestion des droits de reproduction ont intégré des droits de réutilisation par l'IA dans leurs licences de droit d'auteur numérique existantes ou sont en train de mettre au point des solutions comparables : au Japon,

---

65 "Kopinor signs historic agreement on newspaper content for AI in Norway" (19 décembre 2025), communiqué par l'IFRRO, https://ifrro.org/page/article-detail/ifrro-member-kopinor-signs-historic-agreement-on-newspaper-content-for-ai-in-norway/. Conclu par Kopinor au nom de l'Association norvégienne des entreprises du secteur des médias, cet accord permet à la Bibliothèque nationale de Norvège d'accéder à des contenus de presse datant de plus d'un an afin de former des modèles en norvégien; il est financé par l'État norvégien à hauteur de 45 millions de couronnes norvégiennes par an. Il se distingue de l'accord "Bookshelf" conclu précédemment concernant la diffusion numérique des livres.

66 ProRata AI and the Danish Press Publications' Collective Management Organisation (DPCMO), "ProRata Partners with Danish Publishers Group DPCMO" (10 décembre 2025), https://www.prnewswire.com/news-releases/prorata-partners-with-danish-publishers-group-dpcmo-to-launch-the-first-decentralized-sovereign-ai-answer-engine-302637182.html. L'Organisation de gestion collective des publications de la presse danoise (DPCMO), qui affirme représenter 99% du secteur danois de l'information, a conclu un partenariat de concession de licences collectives "opt-in" permettant l'utilisation des contenus de ses membres dans le moteur de recherche par extraction de ProRata, selon un modèle de partage des recettes. Une autre organisation danoise (Tekst&Node) met actuellement au point un modèle de licence pour l'IA dans le cadre du dispositif danois de licences collectives étendues, qui permet aux titulaires de droits de concéder des licences pour certaines utilisations d'œuvres littéraires à des fins d'exploration de textes et de données ainsi que pour des applications d'IA, sous réserve de possibilités de retrait et de l'approbation du Ministère de la culture. Voir Terese Foged, "[DK] First extended collective licence approval for text and data mining and the use of works by AI" (IRIS Legal Observations of the European Audiovisual Observatory, 2026) <https://merlin.obs.coe.int/article/10505> consulté le 3 juillet 2026.

67 NDP Nieuwsmedia et TNO, "Large dataset news organizations for Dutch AI language model GPT-NL" (17 juillet 2025), https://www.ndpnieuwsmedia.nl/2025/07/17/large-dataset-news-organizations-for-dutch-ai-language-model-gpt-nl/. Les membres de NDP Nieuwsmedia, l'association des éditeurs de presse privés néerlandais, en collaboration avec l'agence de presse ANP, mettent à disposition une partie de leurs archives pour l'entraînement de GPT-NL, un modèle de langage néerlandais à but non lucratif développé par TNO, NFI et SURF; les éditeurs participants recevront une rémunération lors de la mise sur le marché du modèle.

68 CEDRO, communiqué de presse concernant son accord de collaboration avec le collectif RSL (janvier 2026), https://www.cedro.org/sala-de-prensa/noticias. Le CEDRO, l'organisation espagnole de gestion des droits de reprographie, a rejoint le collectif RSL en tant que partenaire pour l'Espagne et participe au comité technique du RSL, où il s'efforce d'adapter la norme de licence déchiffrable par machine au marché espagnol afin que ses membres puissent définir les conditions d'utilisation de leurs œuvres dans le cadre de l'entraînement de l'IA. Le 16 décembre 2025, le CEDRO a publié séparément des recommandations concernant la cession des droits liés à l'entraînement de l'IA.

69 Copyright Licensing Agency, "CLA announces development of Generative AI Training Licence" (2025), https://cla.co.uk/development-of-cla-generative-ai-licence/; Publishers' Licensing Services, "PLS and ALCS agree to development of pioneering CLA generative AI licence" (avril 2025), https://www.pls.org.uk/news-events-policy/news/pls-and-alcs-agree-to-development-of-pioneering-cla-generative-ai-licence/. La CLA, en collaboration avec PLS et ALCS, met en place une licence collective à adhésion volontaire pour l'utilisation de textes dans le cadre de l'IA générative, couvrant l'entraînement, le réglage fin et la génération augmentée par récupération; les titulaires de droits y participent en enregistrant leurs œuvres.

la JAC a ajouté ces droits à sa licence de droit d'auteur numérique en avril 2025[70]; la Copyright Agency en Australie a mis au point des solutions comparables[71]; et en Allemagne, la VG Wort a étendu sa licence de droit d'auteur numérique à la réutilisation par l'IA, tandis que la société de gestion collective GEMA a proposé séparément un modèle dédié qui n'avait pas encore abouti à un accord conclu à la date de la présente étude[72]. Ces évolutions restent inégales et, dans plusieurs cas, expérimentales, mais elles montrent que la voie des licences collectives fait l'objet d'essais concrets plutôt que de simples débats.

La licence collective étendue est un modèle de licence né dans les pays nordiques dans les années 1960 et qui a été adopté dans plusieurs autres pays. Ce dispositif résout le problème du risque résiduel en étendant la portée de la licence d'une organisation de gestion collective aux titulaires de droits non membres, dans le cadre d'un répertoire défini, et prévoit un droit de retrait. Ce modèle a été évoqué dans les débats politiques européens sur l'entraînement de l'IA comme un moyen de régler les questions de droits sur les grands corpus textuels et visuels dont le développeur a besoin, mais pour lesquels la concession de licences individuelles n'est pas envisageable. Une étude du Parlement européen de 2025 sur l'IA générative et le droit d'auteur y faisait expressément référence[73]. Aux États-Unis d'Amérique, le Bureau du droit d'auteur, dans la troisième partie de son rapport sur le droit d'auteur et l'intelligence artificielle, a identifié la licence collective étendue comme une option intermédiaire ciblée envisageable pour l'entraînement de l'IA, distincte de la licence obligatoire[74]. Cette même proposition se retrouve, avec quelques variantes, dans la littérature universitaire consacrée à la rémunération : les travaux du professeur Senftleben sur l'IA générative et la rémunération des auteurs proposent un système dans lequel les organisations de gestion collective percevraient une rémunération pour l'utilisation de services d'IA générative dont les résultats se substituent aux œuvres d'auteurs identifiables ou leur font concurrence, et répartiraient les recettes entre leurs membres[75].

Outre les exemples déjà cités, deux autres exemples montrant comment les mécanismes existants de gestion collective volontaire pourraient être adaptés à l'entraînement de l'IA, en s'inscrivant dans les structures institutionnelles existantes plutôt qu'en attendant une nouvelle législation, illustrent cette possibilité. Au Brésil, la gestion des droits de représentation ou d'exécution publique d'œuvres musicales est centralisée au sein d'un organisme unique,

---

70 Japan Academic Association for Copyright Clearance (JAC) et RightsDirect Japan, "JAC and RightsDirect Japan Announce the Availability of AI Re-Use Rights for Digital Copyright License" (9 avril 2025), https://www.businesswire.com/news/home/20250409032666/en/. La JAC a ajouté des droits de réutilisation de l'IA à sa licence de droit d'auteur numérique, une licence à usage interne à l'échelle de l'entreprise destinée aux sociétés dont le siège social est situé au Japon, qui s'appuie sur un répertoire d'œuvres provenant d'éditeurs et de sociétés savantes du monde entier.

71 Copyright Agency, "AI and Copyright in Australia" and accompanying Q&A, https://www.copyright.com.au/membership/ai-and-copyright-in-australia/. Le bureau du droit d'auteur (Copyright Agency), qui fait partie du groupe de référence sur le droit d'auteur et l'IA mis en place par le Gouvernement australien, indique qu'il élabore actuellement des solutions de concession de licences pour l'utilisation de contenus protégés dans le cadre de l'IA générative en Australie.

72 Concernant VG Wort, la licence de droit d'auteur numérique gérée par RightsDirect (une filiale de CCC) a été étendue à la réutilisation interne par l'IA; voir https://www.rightsdirect.com/license/vg-wort-digital-license-2/. La société de gestion collective GEMA a proposé en 2024 un modèle de licence pour l'IA générative; aucun accord n'avait encore été conclu à la date de la présente étude, et la GEMA a parallèlement engagé des poursuites pour contrefaçon à l'encontre de développeurs d'IA (voir *GEMA* c*. OpenAI*, *supra*). Voir GEMA, "AI and music: GEMA calls for music creators to share in AI providers' revenues and presents first licensing model" (26 septembre 2024) <https://www.cisac.org/Newsroom/society-news/ai-and-music-gema-calls-music-creators-share-ai-providers-revenues-and> consulté le 3 juillet 2026; GEMA, "Suno AI and Open AI: GEMA sues for fair compensation" (20 janvier 2025) <https://www.gema.de/en/news/ai-and-music/ai-lawsuit> consulté le 3 juillet 2026.

73 Parlement européen, Direction générale des politiques internes, Étude intitulée "Generative AI and Copyright: Training, Creation, Regulation", IUST_STU(2025)774095_EN (2025), recommandant la licence légale comme cadre optimal pour l'entraînement de l'IA dans l'UE.

74 Bureau américain du droit d'auteur, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (2025).

75 M.R.F. Senftleben, "Generative AI and Author Remuneration", 54 IIC 1535 (2023).

l'ECAD, qui dispose d'un cadre juridique et judiciaire lui permettant de centraliser la perception et la répartition des redevances de représentation ou d'exécution publique sur les œuvres musicales et les phonogrammes, tant au titre du droit d'auteur que des droits connexes[76]. Dans le cadre de la loi actuelle sur le droit d'auteur et en l'absence de modification législative, il serait juridiquement possible qu'un droit spécifique concernant l'utilisation d'œuvres et de phonogrammes pour l'entraînement de modèles d'IA soit exercé collectivement selon un modèle similaire à celui actuellement en vigueur pour le droit de représentation ou d'exécution publique d'œuvres musicales. Cela impliquerait que les titulaires de droits adhèrent à une organisation de gestion collective et que celle-ci prévoie dans ses statuts la représentation du titulaire des droits en ce qui concerne d'autres droits, pas uniquement pour les droits de représentation ou d'exécution publique d'œuvres musicales; cela nécessiterait également l'autorisation préalable du Ministère de la culture, l'organisme fédéral chargé de la supervision des organisations de gestion collective. La loi brésilienne sur le droit d'auteur, en son article 97, autorise les organisations de gestion collective à étendre leur champ d'action à la concession de licences portant sur les droits liés à la formation et à la mise en œuvre du modèle (le droit de reproduction, la communication au public ou la mise à disposition et, selon la position adoptée par le Ministère de la culture, le droit d'adaptation lors de la phase de résultats), à condition que les autorisations requises soient obtenues. L'obligation de perception unitaire ne s'applique toutefois qu'aux droits de représentation ou d'exécution publique d'œuvres musicales; ainsi, même si l'ECAD ou ses filiales devaient centraliser la représentation d'un droit lié à l'entraînement de l'IA au sein de l'organisation, cela n'empêcherait pas la création d'autres entités collectives poursuivant le même objectif.

Au Kenya, le règlement de 2020 relatif au droit d'auteur (gestion collective) régit l'enregistrement, le fonctionnement et la supervision des organisations de gestion collective[77]. Ce cadre a un caractère général et n'est pas spécifiquement adapté aux utilisations informatiques, mais il offre une base institutionnelle sur laquelle pourrait s'appuyer la mise en place d'un système de licences collectives volontaires pour l'entraînement de l'IA. Le projet de loi de 2026 sur le droit d'auteur et les droits connexes, publié en mars 2026 par le Conseil kényan du droit d'auteur afin de recueillir les commentaires du public – la période de consultation prenant fin à la fin de ce même mois –, contient un article (article 34) autorisant l'analyse de textes et de données à des fins de recherche non commerciale; toutefois, le projet de loi ne fournit pas de définition légale claire du terme "non commercial", n'introduit pas de mécanisme de retrait pour les titulaires de droits et ne contient pas de dispositions spécifiques concernant l'utilisation d'œuvres protégées par le droit d'auteur pour l'entraînement de l'IA, ni sur la protection par le droit d'auteur des œuvres générées par l'IA. La situation au Kenya illustre une tendance plus générale observée dans les ressorts juridiques où le cadre du droit d'auteur fait l'objet d'une adaptation progressive : l'infrastructure existante des organisations de gestion collective offre un moyen potentiel de concession de licences collectives pour l'entraînement de l'IA, mais cette adaptation nécessite des mesures législatives ou contractuelles concrètes qui n'ont pas encore été prises.

3. Licences communautaires et de gestion responsable des données linguistiques primaires

Les mécanismes de licence décrits ci-dessus s'appliquent à des données d'entraînement qui sont déjà définitives et publiées au moment où le développeur y a accès. Une catégorie distincte de données d'entraînement voit le jour d'une manière différente. Il n'est pas possible de créer des modèles de parole et de langage pour la plupart des langues africaines à partir de

---

[76] Concernant la structure du système brésilien de gestion collective et la centralisation de la perception des redevances de représentation ou d'exécution publique au sein de l'ECAD, voir la loi brésilienne sur le droit d'auteur (loi n° 9 610/1998), articles 97 à 100, et le décret n° 9 574/2018. La constitutionnalité de ce modèle a été confirmée par le Tribunal fédéral suprême dans la décision ADI 2.054 (2003) et dans l'arrêt conjoint rendu dans les ADI 5062 et 5065 (2016).

[77] Règlement de 2020 relatif au droit d'auteur (gestion collective) (Kenya). Le dispositif relatif aux licences est défini aux articles 33 et 34 ainsi qu'à la partie VII de la loi sur le droit d'auteur (chap. 130).

corpus textuels à l'échelle du Web, car il existe très peu de textes de ce type. Ces données sont en effet collectées sous forme de données primaires : enregistrements audio et transcriptions recueillis auprès des locuteurs dans le cadre de travaux de terrain, de sessions d'enregistrement communautaires et de partenariats avec des diffuseurs. Le droit d'auteur prend effet au moment de la fixation de l'œuvre, et non au moment de sa mise en ligne. L'enregistrement est protégé en tant qu'enregistrement sonore, la transcription en tant qu'œuvre littéraire, et le locuteur peut détenir des droits d'artiste interprète ou exécutant[78]. Les questions d'autorisation et de rémunération se posent donc dès le moment de la collecte, avant même que ne se pose la question de la fouille de textes et de données.

Cette catégorie de données soulève également une autre dimension en matière de propriété intellectuelle. Les enregistrements de la tradition orale, des contes populaires et des proverbes constituent à la fois des œuvres protégées par le droit d'auteur, dont les droits appartiennent à la personne qui les a fixées, et des expressions culturelles sur lesquelles les communautés détiennent des droits collectifs en vertu d'une législation *sui generis.* Au Kenya, la loi n° 33 de 2016 sur la protection des savoirs traditionnels et des expressions culturelles, qui donne effet aux articles 11, 40 et 69.1)c) de la Constitution du Kenya de 2010, reconnaît aux communautés la qualité de titulaires de droits sur leurs expressions culturelles, exige le consentement préalable en connaissance de cause et la conclusion d'accords d'utilisation autorisés pour leur exploitation, et prévoit le versement d'une compensation aux communautés; l'article 11.3)a) de la Constitution oblige le Parlement à veiller à ce que les communautés perçoivent des redevances pour l'utilisation de leurs cultures et de leur patrimoine culturel[79]. Lorsque des données linguistiques de ce type sont intégrées à un corpus d'entraînement, la chaîne de titularité des droits d'auteur et les droits de la communauté s'appliquent en parallèle, et une licence accordée par la personne ayant réalisé l'enregistrement ne régit pas en soi la situation juridique de la communauté.

Une pratique distincte en matière d'octroi de licences s'est développée en réponse à cette situation. Les licences ouvertes standard, et en particulier les mises au domaine public, intègrent ces données dans des chaînes de traitement sans conférer de pouvoir de gouvernance ni de retour économique à la communauté qui les fournit. Les licences axées sur la communauté réinterprètent cette approche tout en préservant l'ouverture. La licence de données ouvertes Nwulite Obodo prévoit des conditions d'utilisation à plusieurs niveaux, différenciées en fonction du contexte géographique et de développement de l'utilisateur, et intègre des obligations de partage des bénéfices; elle a été appliquée à des ensembles de données vocales africaines, notamment l'ensemble de données vocales DhoNam Dholuo développé en collaboration avec l'université de Maseno[80]. La licence Esethu, élaborée pour l'ensemble de données vocales Vuk'uzenzele isiXhosa, autorise l'utilisation à des fins de recherche et l'utilisation commerciale par des entités africaines, sans frais, tout en exigeant le paiement de droits de licence de la part des autres utilisateurs commerciaux, les recettes étant

---

78 Loi sur le droit d'auteur (chap. 130), législation kényane, article 2 (définissant, entre autres, les notions d'"enregistrement sonore" et d'"œuvre littéraire") et dispositions relatives aux droits des artistes interprètes ou exécutants. En ce qui concerne les droits des artistes interprètes ou exécutants au niveau international, voir les articles 10 et 14 du WPPT, examinés dans la section B.1 ci-dessus.

79 Loi n° 33 de 2016 sur la protection des savoirs traditionnels et des expressions culturelles (Kenya); Constitution du Kenya de 2010, articles 11, 40.5) et 69.1)c).

80 Nwulite Obodo Open Data Licence, Version 1.0 (Data Science Law Lab, University of Pretoria, 2025) <https://datasciencelawlab.africa/nwulite-obodo-open-data-license/> consulté le 31 juillet 2026; C.I. Okorie and M. Omino, "Addressing Inequitable Openness in Licences for Sharing African Data and Datasets Through the Nwulite Obodo Open Data Licence" (2025) 7(3) Law, Technology and Humans 94. L'ensemble de données vocales DhoNam Dholuo (Maseno University Centre for African AI) est publié sous cette licence : <https://mozilladatacollective.com/datasets/cmjepxo6t08nmmk07iauvua6v> consulté le 31 juillet 2026.

réinvesties dans la création de nouvelles données linguistiques[81]. La licence Kaitiakitanga de Te Hiku Media, qui régit les données relatives à la langue maorie, reflète la même approche dans une autre région[82]. Ces instruments constituent un troisième mécanisme, parallèlement aux licences individuelles et collectives, par lequel s'organisent l'autorisation et la rémunération liées à l'utilisation de données linguistiques primaires dans l'entraînement des IA; ils illustrent comment un cadre réglementaire national régissant les biens culturels collectifs peut étayer les pratiques contractuelles. Au Kenya, cette disposition entrera également en interaction avec l'exception relative à l'analyse de textes et de données proposée dans le projet de loi de 2026 sur le droit d'auteur et les droits connexes, évoquée dans la sous-section précédente.

4. Exceptions TDM réglementaires et clauses générales

La législation de l'Union européenne en matière de droit d'auteur prévoit des exceptions explicites relatives à la fouille de textes et de données, qui s'appliquent aux opérations utilisées dans le cadre de l'entraînement des IA, bien que leur application et leur pertinence dans ce contexte restent contestées. Les articles 3 et 4 de la directive (UE) 2019/790 sur le droit d'auteur et les droits voisins dans le marché unique numérique prévoient deux exceptions obligatoires en matière de fouille de textes et de données. L'article 3 prévoit une exception à des fins de recherche pour les organismes de recherche et les institutions du patrimoine culturel, sans possibilité de dérogation. L'article 4 prévoit une exception générale applicable à tout bénéficiaire, sous réserve d'une opposition du titulaire des droits "de manière appropriée, notamment par des procédés lisibles par machine pour les contenus mis à la disposition du public en ligne"[83]. La loi sur l'IA intègre ce cadre par référence, obligeant les fournisseurs de modèles d'IA à usage général à mettre en place une politique visant à se conformer à la législation de l'Union en matière de droit d'auteur, à identifier et à respecter les réservations de droits exprimées conformément à l'article 4.3) de la directive sur le droit d'auteur et les droits voisins dans le marché unique numérique, ainsi qu'à publier un résumé suffisamment détaillé du contenu utilisé pour l'entraînement[84].

Deux décisions judiciaires allemandes ont abordé la question de l'application de ces exceptions à l'entraînement des IA. Dans l'affaire *Kneschke* c*. LAION e.V.*, le Landgericht de Hambourg a jugé en septembre 2024 que l'utilisation d'une photographie dans la constitution du jeu de données LAION-5B, utilisé pour l'entraînement de modèles de génération d'images, relevait de la transposition allemande de l'exception prévue à l'article 3 de la directive (UE) 2019/790 concernant la fouille de textes et de données à des fins de recherche scientifique. Le tribunal a considéré LAION comme un organisme de recherche scientifique au sens de cette disposition et a estimé que l'analyse informatique en cause constituait une fouille de textes et de données à des fins scientifiques. Il a également indiqué, à titre accessoire, que l'exception générale TDM prévue à l'article 4 ne s'appliquerait en tout état de cause pas, car la réservation de droits figurant dans les conditions d'utilisation de la plateforme source était, de l'avis du tribunal, lisible

---

81 Lelapa AI, Way With Words, and Data Science for Social Impact (University of Pretoria), "The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (2025) <https://aclanthology.org/2025.acl-long.1487/> consulté le 31 juillet 2026.

82 Te Hiku Media, Kaitiakitanga License <https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license> consulté le 31 juillet 2026.

83 Directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique, JO L 130, p. 92, article 3 (exception obligatoire relative à la fouille de textes et de données à des fins de recherche scientifique par des organismes de recherche et des institutions du patrimoine culturel) et article 4 (exception générale, avec possibilité de dérogation, relative à la fouille de textes et de données d'œuvres accessibles de manière licite).

84 Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (loi sur l'intelligence artificielle), JO L, 2024/1689, 12 juillet 2024, article 53.1)c) et d), et considérants 105 et 106.

par une machine compte tenu des capacités actuelles des grands modèles linguistiques[85]. En appel, dans l'affaire *Kneschke* c*. LAION e.V.*, l'Oberlandesgericht de Hambourg a confirmé la décision rendue en première instance, tout en approfondissant l'analyse de l'exigence de lisibilité par machine prévue à l'article 4.3)[86]. La cour d'appel a estimé que la lisibilité par machine exigeait non seulement que la réservation puisse être saisie par une machine, mais également qu'elle puisse être interprétée par un processus automatisé de telle sorte que le contenu visé par la réservation ne soit pas traité. En résumé, la réservation doit être à la fois lisible et exploitable par une machine. La cour a en outre précisé que la lisibilité par machine n'est pas une caractéristique statique, mais qu'elle doit être appréciée au regard des capacités technologiques existant au moment de l'utilisation litigieuse; en l'espèce, cela correspondait à l'état de la technique au cours du second semestre 2021. Le raisonnement de la cour d'appel laisse en suspens la question de savoir ce qui constituera la lisibilité par machine en 2025 ou au-delà, lorsque des modèles capables d'extraire du sens à partir de textes en langage naturel non structurés seront largement disponibles[87].

Environ un mois auparavant, dans l'affaire *GEMA* c*. OpenAI*, le Landgericht München I était parvenu à une conclusion différente sur une question connexe mais distincte[88]. La société allemande de gestion collective GEMA avait intenté une action en justice contre OpenAI, affirmant que ChatGPT avait mémorisé les paroles de neuf chansons allemandes célèbres et pouvait être amené à les reproduire. Le tribunal de Munich a considéré que tant l'intégration des paroles dans le corpus d'entraînement que leur mémorisation ultérieure au sein du modèle constituaient des actes de reproduction au sens de la transposition allemande de la directive InfoSoc; il a rejeté l'application de l'exception générale TDM prévue à l'article 44 b de la loi allemande sur le droit d'auteur, au motif que les modèles génératifs mémorisent de manière permanente les paroles et que cela va au-delà de l'analyse d'informations textuelles et basées sur des modèles que cette exception était censée autoriser; et il a considéré les résultats restitués comme la preuve que les paroles étaient fixées au sein du modèle d'une manière comparable à celle d'une bibliothèque numérique. Le tribunal a accordé une mesure injonctive, ordonné la communication d'informations et accordé des dommages-intérêts[89]. Cette décision fait l'objet d'un recours. Certains y ont vu le premier arrêt européen majeur en matière d'entraînement des IA en faveur des titulaires de droits; d'autres l'ont critiqué, estimant qu'il réduisait à un seul acte de reproduction ce qui, sur le plan technique, constitue une succession

---

85 *Kneschke* c*. LAION e.V.*, Tribunal régional de Hambourg, 27 septembre 2024 (310 O 227/23) (jugement selon lequel la constitution de l'ensemble de données LAION-5B destiné à l'entraînement de modèles de génération d'images relevait de la transposition allemande de l'exception prévue à l'article 3 de la directive (UE) 2019/790 relatif à la fouille de textes et de données à des fins de recherche scientifique (§ 60 d de la loi allemande sur le droit d'auteur (UrhG)); indiquant également, à titre d'observation incidente, que l'exception générale prévue à l'article 4 de la directive (UE) 2019/790 n'aurait pas été applicable, au motif que la réservation de droits figurant dans les conditions d'utilisation de la plateforme source était lisible par machine au sens pertinent). En appel, voir *Kneschke* c*. LAION e.V.*, OLG de Hambourg, 10 décembre 2025 (5 U 104/24), commenté ci-dessous.

86 *Kneschke* c*. LAION e.V.*, Cour d'appel de Hambourg, 10 décembre 2025 (5 U 104/24).

87 La Cour d'appel de Hambourg a précisé que la lisibilité par machine exige que la réservation "puisse être interprétée par une machine de telle sorte […] que, dans le cadre d'un processus automatisé, le contenu visé par la réservation[de droits] ne soit pas traité" (traduction des auteurs). Voir aussi P. Keller, "LAION Round 2 : Machine-Readable but Still Not Actionable – The Lack of Progress on TDM Opt-Outs", Kluwer Copyright Blog (17 et 18 décembre 2025), Part 1.

88 *GEMA* c*. OpenAI*, Tribunal régional de Munich I, 11 novembre 2025 (42 O 14139/24), jugement définitif. Le plaignant est la société allemande de gestion collective GEMA; les défendeurs sont OpenAI OpCo LLC et OpenAI Ireland Limited. Le tribunal a ordonné des mesures injonctives, la communication d'informations et le versement de dommages-intérêts.

89 Ibid. Le tribunal a considéré que tant l'intégration des paroles dans le corpus d'entraînement que leur mémorisation ultérieure par le modèle constituaient des actes de reproduction au sens de l'article 16 de l'UrhG, qui transpose l'article 2 de la directive 2001/29/CE.

d'opérations distinctes, et qu'il appliquait à tort l'exception TDM en confondant les preuves de mémorisation au stade de la sortie avec la réalité technique du processus d'entraînement[90].

Au-delà de la jurisprudence, un débat théorique oppose les partisans et les détracteurs quant à savoir si l'exception prévue à l'article 4 de la directive (UE) 2019/790 s'applique réellement à l'entraînement des IA. Une opinion largement répandue dans la littérature est que l'entraînement de l'IA constitue une forme de fouille de données relevant des articles 3 et 4 de la directive (UE) 2019/790. Ce point de vue invoque souvent le fait que ces articles sont mentionnés dans la loi sur l'IA. Un point de vue contraire a été largement défendu par plusieurs spécialistes européens du droit, dont le professeur Tim Dornis, qui soutient, dans un article publié dans la revue *European Intellectual Property Review* et dans un avis juridique plus complet rédigé en collaboration avec l'informaticien Stober, qu'un examen approfondi des fondements technologiques de l'entraînement de l'IA générative révèle que les opérations effectuées vont au-delà de ce que l'exception relative à la fouille de textes et de données était censée couvrir. L'argument se présente globalement comme suit : l'exception TDM autorise l'extraction d'informations sémantiques, factuelles et statistiques à partir d'œuvres; en revanche, l'entraînement d'un modèle d'IA générative extrait et encode également des caractéristiques syntaxiques et expressives, d'une manière qui est ensuite utilisée pour produire des résultats se substituant aux œuvres originales; et l'atteinte qui en résulte aux intérêts légitimes des titulaires de droits dépasse le champ d'application de ce que l'exception a été conçue pour autoriser[91].

La professeure Eleonora Rosati a développé un raisonnement similaire, mais en partant d'un point de vue différent. Elle estime que la fouille de textes et de données peut faire partie des processus d'entraînement de l'IA, mais qu'elle n'est ni synonyme d'entraînement de l'IA, ni ne recouvre l'intégralité de ce que cet entraînement implique; qu'il est trop réducteur de limiter le débat européen à l'article 4 de la directive (UE) 2019/790, compte tenu de la multiplicité des actes protégés par le droit d'auteur que l'entraînement de l'IA peut impliquer; et que, en tout état de cause, le triple critère prévu à l'article 7.2) de la directive (UE) 2019/790 et à l'article 5.5) de la directive InfoSoc peut avoir pour effet d'exclure l'entraînement de l'IA du champ d'application de l'exception, au motif que l'utilisation est contraire à l'exploitation normale des œuvres et porte atteinte de manière déraisonnable aux intérêts légitimes des titulaires de droits. Si cela est exact, la question des licences reste un élément indispensable de l'analyse[92]. Le professeur Matthias Leistner a quant à lui fait valoir que le mécanisme de dérogation prévu à l'article 4.3) est susceptible de devenir le principal enjeu du droit européen du droit d'auteur et qu'il est actuellement inapplicable dans la pratique, une position qui a été reprise dans le débat politique européen[93].

Le professeur Senftleben soutient au contraire que l'article 4 s'applique non seulement dans le cadre de l'entraînement des IA, mais qu'il peut également être directement pertinent pour

---

90 Pour une analyse critique sur ce point, voir G. Frosio, "Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI", Kluwer Copyright Blog (10 et 11 décembre 2025), parties 1 et 2, qui soutient que le tribunal de Munich confond les preuves de mémorisation au stade de la production avec la réalité technique de l'entraînement, et qu'il serait plus judicieux de considérer les copies techniques intermédiaires licites comme relevant de l'article 5.1), de la directive InfoSoc ou des exceptions relatives à l'exploitation de bases de données.

91 T. W. Dornis, "The Training of Generative AI Is Not Text and Data Mining", 47 European Intellectual Property Review 65 (2025). Voir aussi T.W. Dornis and S. Stober, *Urheberrecht und Training generativer KI-Modelle – technologische und juristische Grundlagen*, Initiative Urheberrecht (Nomos, septembre 2024).

92 E. Rosati, "Is text and data mining synonymous with AI training?", 19 Journal of Intellectual Property Law and Practice 851 (2024); E. Rosati, "The German LAION decision: A problematic understanding of the scope of the TDM copyright exceptions and the transition from TDM to AI training", IPKat (7 October 2024); E. Rosati, Copyright Exceptions and Fair Use Defences for AI Training Done for 'Research' and 'Learning', or the Inescapable Licensing Horizon, SSRN Working Paper No. 5331405 (2025).

93 M. Leistner, "TDM und KI-Training in der Europäischen Union", GRUR 2024, 1665 (dans lequel l'auteur fait valoir que le mécanisme de dérogation est susceptible de devenir le principal enjeu du droit d'auteur de l'Union européenne et qu'il est actuellement inapplicable dans la pratique).

déterminer ce qui constitue une exploitation normale d'une œuvre dans le contexte de l'IA, ce qui serait pertinent pour l'application du triple critère[94].

Dans son interprétation de la loi sur l'IA, la Commission européenne a considéré que l'article 4 de la directive (UE) 2019/790 s'appliquait à l'entraînement des IA, en ce sens que son approche met l'accent sur la mise en œuvre du mécanisme de dérogation plutôt que sur l'exclusion de l'entraînement des IA du champ d'application de l'exception. L'infrastructure de dérogation est abordée dans la troisième partie de la présente étude, dans le cadre de la transparence et de l'application de la réglementation[95].

Le Parlement coréen a examiné, sans toutefois l'adopter, une exception relative à la fouille de textes et de données; des projets de loi allant dans ce sens sont à l'étude à l'Assemblée nationale depuis 2021[96]. En l'absence de disposition réglementaire, la question de l'entraînement doit être appréciée au regard de la clause générale d'usage loyal prévue à l'article 35-5 de la loi sur le droit d'auteur, disposition inspirée de l'article 107 de la loi américaine sur le droit d'auteur et introduite à la suite de l'accord de libre-échange conclu entre la République de Corée et les États-Unis. Le débat national suit de près les considérations exposées dans cette section. Les partisans d'une exception réglementaire mettent en avant la certitude qu'apporterait une formulation réglementaire explicite; ceux qui s'opposent à une exception réglementaire rétorquent que le libellé réglementaire crée ses propres ambiguïtés d'interprétation, comme le montre le litige allemand sur la lisibilité par machine évoqué plus haut, que la formulation en 2024 par la Cour suprême de Corée des critères régissant l'article 35-5 donne à la clause générale un contenu applicable, et qu'une exemption large pourrait profiter de manière disproportionnée aux développeurs étrangers tout en détournant la valeur des titulaires de droits sur les contenus en langue coréenne[97].

En matière d'entraînement, cette clause générale a été concrétisée par des directives administratives. En février 2026, le Ministère de la culture, des sports et du tourisme et la Commission coréenne des droits d'auteur ont publié conjointement le "Guide sur l'applicabilité de la doctrine de l'usage loyal à l'entraînement des modèles d'IA générative" (mentionné dans la première partie de la présente étude), qui applique les quatre critères de l'article 35-5 aux différentes étapes du processus d'entraînement décrites dans la première partie de la présente étude. Le Guide précise que la manière dont les données d'entraînement ont été obtenues est un élément pertinent : ainsi, le moissonnage non autorisé, le contournement des mesures techniques de protection, le contournement des barrières d'accès payant et le non-respect des réservations exprimées par le biais du protocole d'exclusion des "robots" constituent des éléments qui vont à l'encontre d'une conclusion d'usage loyal; des exemples illustratifs sont fournis pour les deux cas de figure[98].

---

94 Senftleben, M. Text and Data Mining, Generative AI, and the Copyright Three-Step Test. IIC 57, 67, at 79 (2026). https://doi.org/10.1007/s40319-026-01680-2

95 Loi sur l'intelligence artificielle, article 53.1)c) (obligation de mettre en place une politique visant à se conformer au droit de l'Union en matière de droit d'auteur et à identifier et à respecter, y compris au moyen de technologies de pointe, une réservation de droits exprimée conformément à l'article 4.3) de la directive (UE) 2019/790).

96 À commencer par le projet de loi complet portant modification de la loi sur le droit d'auteur, présenté à l'Assemblée nationale le 15 janvier 2021, qui prévoyait une exception relative à la fouille de textes et de données; ce projet de loi est tombé en désuétude avec la fin de la vingt et unième législature, et les propositions qui lui ont succédé sont toujours à l'étude.

97 Cour suprême de Corée, arrêt n° 2021Da272001 (2024), qui définit les critères applicables en vertu de l'article 35-5 de la loi sur le droit d'auteur.

98 Guide coréen sur l'usage loyal, note 10 ci-dessus.

5. Systèmes de rémunération

a) Types de systèmes de rémunération

Lorsque l'utilisation d'œuvres protégées dans le cadre de l'entraînement des IA est autorisée, que ce soit en vertu d'une exception, d'une licence ou des deux, la question de la rémunération peut se poser. Plusieurs modèles sont à l'étude, s'inspirant de l'expérience plus générale du droit d'auteur en matière d'adaptation aux nouveaux modes d'exploitation.

Le premier est un système de redevances de licence, négociées soit individuellement, soit collectivement, versées par le développeur au moment de la constitution du corpus ou de l'entraînement. L'avantage de ce modèle réside dans le fait que le titulaire des droits est rémunéré pour l'utilisation de l'œuvre à des fins d'entraînement, sans autre condition; son inconvénient, dans le cas d'un octroi de licences individuelles, est qu'il ne prend pas en compte la longue succession de titulaires de droits individuels pour lesquels une négociation directe n'est pas envisageable.

Le deuxième est un système fondé sur une redevance, dans lequel une redevance est imposée aux développeurs d'IA ou à certaines catégories d'utilisateurs de services d'IA générative, et dont le produit est réparti entre les titulaires de droits par l'intermédiaire d'une organisation de gestion collective ou d'un organisme équivalent. L'étude du Parlement européen mentionnée ci-dessus à la section B.3 recommande la mise en place, au sein de l'Union européenne, d'un régime de licence réglementaire pour l'entraînement des IA, qui présenterait les caractéristiques d'un système fondé sur une redevance et assorti d'une gestion collective obligatoire. Des propositions similaires ont été formulées dans des juridictions situées en dehors de l'Union européenne, souvent par analogie avec les redevances de copie privée qui existent dans de nombreux pays. Des questions ont été soulevées quant à la compatibilité de ces systèmes avec le triple critère prévu à l'article 9.2) de la Convention de Berne, l'article 10 du WCT et l'article 13 de l'Accord sur les ADPIC, ainsi que, dans la mesure où des droits autres que le droit de reproduction sont concernés, avec les conditions spécifiques prévues aux articles 11*bis.*2) et 13.1) de la Convention de Berne.

Le troisième système, le plus récent et le plus distinct sur le plan conceptuel, est celui de la rémunération liée aux résultats. Dans cette optique, la rémunération ne porte pas sur l'utilisation à des fins d'entraînement en soi, mais sur la production, par le modèle déployé, de résultats qui se substituent aux œuvres d'auteurs identifiables ou qui leur font concurrence. Ce modèle établit un lien direct entre la rémunération et la possibilité de substitution de l'œuvre du titulaire des droits sur le marché concerné. Sa mise en œuvre semblerait toutefois nécessiter des mécanismes techniques permettant d'identifier les résultats de substitution et de les relier à des œuvres identifiables dans les données d'entraînement.

Deux autres questions se posent. Lorsque le nombre d'œuvres identifiées est très élevé, la rémunération par œuvre peut être si faible qu'elle en devient symbolique. Il convient par ailleurs de clarifier la base de calcul : dans la plupart des propositions, la rémunération ne reposerait pas sur la démonstration qu'une production donnée reproduit ou adapte une œuvre spécifique, mais sur la déclaration par le développeur des modalités d'utilisation des œuvres au sein du système, une approche qui présuppose les mécanismes de transparence évoqués à la section C de la présente partie[99]. De plus, les auteurs peuvent être affectés par l'effet de substitution, que leurs travaux aient été utilisés ou non pour l'entraînement. Comme indiqué plus haut, avec les technologies de filtrage actuelles, il est peu probable, dans la pratique, que les résultats obtenus enfreignent une œuvre particulière protégée par le droit d'auteur; le préjudice envisagé est donc de nature systémique plutôt que lié à une violation typique du droit d'auteur. Une reproduction

[99] Dans le domaine musical, l'intégrité des métadonnées au niveau de l'œuvre et au niveau du titulaire des droits associées au matériel d'entraînement a été identifiée comme une condition préalable indispensable tant pour l'attribution que pour tout partage des recettes issues des résultats générés : de Berardinis *et al.*, "RAIM : Operationalising Trustworthy AI in Music Generation", cité dans la partie I, note de bas de page n° 17, de la présente étude.

réalisée au cours de la phase d'entraînement, ou subsistant dans le modèle, n'implique pas que les résultats constituent une contrefaçon : l'analyse de la contrefaçon pour un résultat donné repose sur la ressemblance de ce résultat avec une œuvre protégée.

L'ampleur du problème de substitution est de plus en plus documentée dans le secteur musical : en avril 2026, Deezer a indiqué que les morceaux entièrement générés par l'IA représentaient environ 44% de ses mises en ligne quotidiennes, soit environ 75 000 morceaux par jour, tandis que Spotify a signalé avoir supprimé plus de 75 millions de morceaux indésirables au cours des 12 derniers mois[100], et une étude sectorielle prévoit qu'une part substantielle des revenus des créateurs de musique pourrait être supplantée par du contenu généré par l'IA d'ici 2028[101]. En réponse, le professeur Senftleben a fait valoir que le test de substitution pourrait servir de cadre conceptuel permettant de distinguer l'utilisation légitime de l'IA pour soutenir la créativité humaine, qui ne donnerait lieu à aucune rémunération, de l'utilisation de l'IA pour se substituer à la créativité humaine, qui, elle, donnerait lieu à une rémunération. On peut considérer qu'une approche similaire sous-tend la version sénatoriale du projet de loi brésilien n° 2338/2023, qui associe un mécanisme de dérogation à un droit à rémunération que le titulaire des droits peut exercer à titre individuel ou collectif[102].

b) Modalités et base de rémunération

Au Brésil, l'effet de substitution des créations synthétiques sur les œuvres humaines est observable dans certains secteurs, notamment celui de la conception sonore d'ambiance, où les utilisateurs ont cherché à diffuser de la musique entièrement générée par des systèmes d'IA dans le but d'échapper aux redevances d'exécution publique. Les affaires en cours contre l'ECAD sur cette question permettront, en temps voulu, d'établir des orientations judiciaires quant à savoir si l'exécution publique de musique générée par l'IA dans des locaux commerciaux donne lieu à une obligation de rémunération. Cet exemple illustre l'interaction entre la rémunération liée à l'entraînement de l'IA et celle liée aux résultats générés par l'IA.

Deux questions concernant un éventuel système de rémunération ont été soulevées au cours des discussions en cours. La première question est de savoir si la rémunération doit être liée à l'entraînement, à l'effet de substitution (c'est-à-dire à l'un ou l'autre, ou aux deux). L'étude du Parlement européen et plusieurs propositions universitaires tentent de combiner ces deux approches; le projet de loi brésilien n° 2338/2023, dans sa forme actuelle, doit plutôt être interprété comme liant la rémunération à l'utilisation à des fins d'entraînement : l'article 65 prévoit une rémunération pour l'utilisation de contenus protégés dans le cadre de la fouille de données ainsi que dans l'entraînement et le développement de systèmes d'IA, de sorte que le projet de loi établit un mécanisme de rémunération lié à l'acte d'entraînement qui ne s'étend pas à l'exploitation des résultats[103]. La deuxième question est de savoir si le système de

---

100 Deezer, "AI-generated tracks now represent 44% of all new uploaded music" (19 April 2026) <https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/> consulté le 3 juillet 2026; Spotify, "Spotify Strengthens AI Protections for Artists, Songwriters, and Producers" (25 septembre 2025) <https://newsroom.spotify.com/2025-09-25/spotify-strengthens-ai-protections/> consulté le 3 juillet 2026.

101 Confédération internationale des sociétés d'auteurs et compositeurs (CISAC), étude réalisée par PMP Strategy (décembre 2024), selon laquelle environ 24% des revenus des créateurs de musique pourraient être menacés d'ici 2028 en raison de l'IA générative, ce chiffre étant encore plus élevé pour les revenus numériques en particulier.

102 Projet de loi n° 2338/2023 du Sénat, tel qu'approuvé par le Sénat fédéral brésilien le 10 décembre 2024 et transmis à la Chambre des députés. À la date de la présente étude, le projet de loi est examiné par une commission spéciale de la Chambre des députés; la version qui pourrait finalement être soumise à l'approbation présidentielle n'est pas encore définitive.

103 Voici le texte de l'article 65 du projet de loi n° 2338/2023, tel qu'approuvé par le Sénat fédéral et adopté dans le cadre de la présente étude. Le projet de loi a également été interprété comme combinant une approche fondée sur la consommation avec une prise en compte de la substitution au niveau de la production, notamment parce que les critères relatifs à la rémunération font référence aux effets sur le marché. Selon l'interprétation retenue dans le texte, la rémunération s'applique à la fouille de données, à l'entraînement et au développement.

rémunération présuppose un droit exclusif reconnu qui a fait l'objet d'une concession de licence ou d'une renonciation, ou s'il convient plutôt de le considérer comme un droit *sui generis* à une rémunération équitable, s'inspirant peut-être de la rémunération au titre de la copie privée qui existe dans de nombreuses juridictions[104]. Comme indiqué ci-dessus, la législation brésilienne en vigueur pourrait s'appliquer à la phase de production par le biais du droit de reproduction, du droit d'adaptation ou du droit de communication au public, même si le projet de loi n° 2338/2023 n'était pas adopté, et, dans ce cas, l'autorisation préalable et expresse des titulaires de droits serait requise[105]. La copie privée, en revanche, constitue une forme de compensation et non une licence, car les actes de copie à titre privé sont considérés comme ne pouvant, par nature, faire l'objet d'une licence pour diverses raisons, notamment la protection de la vie privée des utilisateurs.

Au Brésil, une autre observation concerne la conception des rémunérations : l'effet de substitution des créations synthétiques sur les œuvres humaines est observable dans certains secteurs concrets, notamment celui de la conception sonore d'ambiance, où les utilisateurs ont cherché à diffuser de la musique entièrement générée par des systèmes d'IA dans le but d'échapper aux redevances d'exécution publique. Les affaires en cours contre l'ECAD sur cette question permettront, en temps voulu, d'établir des orientations judiciaires quant à savoir si la diffusion publique de musique générée par l'IA dans des locaux commerciaux entraîne une obligation de rémunérer les membres de l'ECAD. Cet exemple illustre l'interaction entre la rémunération liée à l'entraînement de l'IA et celle liée aux résultats de l'IA; les choix effectués dans un domaine peuvent avoir une incidence sur la conception de l'autre[106].

Enfin, dans les juridictions où le cadre réglementaire est encore en cours d'élaboration, la question de la rémunération reste pour l'instant en suspens. Le Guide de 2026 sur l'applicabilité de la doctrine de l'usage loyal à l'entraînement des modèles d'IA générative, publié conjointement par le Ministère de la culture, des sports et du tourisme et la Commission coréenne du droit d'auteur, et mentionné dans la première partie, traite principalement de la question de savoir quelles opérations d'entraînement peuvent être menées sans licence; il laisse la question correspondante de la rémunération à traiter dans le cadre du processus de réforme plus large actuellement en cours[107]. Une grande partie du débat national part toutefois du principe que l'entraînement commercial des modèles d'IA devrait donner lieu à une rémunération des créateurs; la question qui reste donc en suspens est celle du mécanisme de rémunération. Les situations au Kenya, au Brésil et en Chine sont, chacune à leur manière, comparables : dans chacun de ces pays, le processus législatif est en cours, les options politiques sont ouvertes et le modèle qui en résultera n'est pas encore défini. La troisième partie de cette étude, consacrée à l'application de la loi, revient sur les conséquences concrètes de cette situation d'incertitude juridique pour les titulaires de droits qui cherchent à faire valoir leurs prétentions.

---

104 Une analyse approfondie des régimes *sui generis* envisageables et des utilisations qu'ils couvriraient dépasse le cadre de la présente étude.

105 Voici le texte de l'article 65 du projet de loi n° 2338/2023, tel qu'approuvé par le Sénat fédéral et adopté dans le cadre de la présente étude. Le projet de loi a également été interprété comme combinant une approche fondée sur la consommation avec une prise en compte de la substitution au niveau de la production, notamment parce que les critères relatifs à la rémunération font référence aux effets sur le marché. Selon l'interprétation retenue dans le texte, la rémunération s'applique à la fouille de données, à l'entraînement et au développement.

106 Projet de loi n° 2338/2023, articles 62 et suivants (dans la version transmise à la Chambre des députés). La clause d'exclusion relative aux secrets d'affaires restreint la portée de l'obligation de transparence et fait l'objet d'un débat animé au sein du processus législatif.

107 Ministère de la culture, des sports et du tourisme et Commission coréenne des droits d'auteur, *Guide sur l'applicabilité de la doctrine de l'usage loyal à l'entraînement des modèles d'IA générative* (février 2026); voir la partie I, note de bas de page n° 10, de la présente étude.

TROISIÈME PARTIE. APPLICATION

Les opérations techniques décrites dans la première partie relèvent souvent de droits exclusifs reconnus par la législation nationale sur le droit d'auteur et par les traités internationaux relatifs au droit d'auteur; les accords de licence examinés dans la deuxième partie proposent des mécanismes permettant d'obtenir l'autorisation d'exercer ces droits lorsque aucune exception ou limitation ne s'applique. La troisième partie traite du cas où aucune de ces deux situations ne s'est produite : lorsqu'un titulaire de droits estime que son œuvre a été utilisée pour l'entraînement d'un modèle d'IA générative sans autorisation et sans bénéficier d'une exception applicable, et qu'il cherche à faire valoir ses droits devant les tribunaux ou dans le cadre de procédures administratives.

Jusqu'à récemment, il n'existait pratiquement aucune jurisprudence à prendre en considération. Ce n'est plus le cas. Au 22 août 2026, 137 actions en justice liées à l'utilisation d'œuvres protégées par des systèmes d'IA générative avaient été intentées contre des développeurs d'IA aux États-Unis d'Amérique, et une trentaine d'autres actions avaient été recensées dans d'autres juridictions[108]. Les tribunaux du Brésil, du Canada, de la Chine, de l'Allemagne, du Royaume-Uni, des États-Unis d'Amérique et d'autres pays ont rendu des décisions qui portent sur plusieurs aspects de l'analyse : sur la question de savoir si le modèle entraîné est lui-même une copie des œuvres sur lesquelles il a été entraîné ou s'il contient des copies des œuvres; sur la question de savoir si les exceptions prévues par le droit européen s'étendent à l'entraînement de l'IA; sur la question de savoir si les tribunaux nationaux sont compétents pour statuer sur un entraînement mené à l'étranger; sur la question de savoir si les résultats qui se substituent aux œuvres du corpus d'entraînement font intervenir le droit de reproduction ou le droit de mise à disposition du titulaire des droits; et sur la question de savoir si les résultats sont eux-mêmes protégés par le droit d'auteur (une question qui sort du champ de la présente étude, mentionnée ici uniquement à titre de contexte).

Cette partie s'articule autour de trois questions principales. La section A tire les enseignements de l'état actuel de la jurisprudence, organisée par thème abordé plutôt que par juridiction. La section B aborde la question structurelle de l'application de la loi à l'encontre des développeurs d'IA, y compris les aspects transfrontières qui compliquent cette application. La section C recense les principaux défis pratiques auxquels sont confrontés les titulaires de droits qui cherchent à faire valoir leurs droits, notamment le manque de transparence, la multiplication des normes techniques relatives aux options de dérogation et la difficulté opérationnelle liée à la collecte de preuves.

SECTION A. ENSEIGNEMENTS TIRÉS DE LA JURISPRUDENCE ACTUELLE

La jurisprudence accumulée au cours des 18 derniers mois est fragmentaire, varie d'une juridiction à l'autre et n'a pas encore fait l'objet d'une consolidation doctrinale au sein d'un seul et même système juridique. Elle n'a pas encore répondu à toutes les questions. Elle met néanmoins en lumière au moins quatre questions distinctes, sur chacune desquelles les tribunaux de différentes juridictions ont commencé à s'accorder ou, dans certains cas, s'opposent. Ces questions sont abordées tour à tour ci-dessous.

1. La question du "modèle en tant que copie"

La première question est de savoir si, au regard du droit d'auteur, le modèle d'IA entraîné contient lui-même des copies des œuvres sur lesquelles il a été entraîné. Cela diffère de la question précédente, à savoir si des reproductions sont effectuées lors de la construction du corpus et de l'entraînement. L'aspect technique de cette question de la mémorisation a été

---

[108] D'après le suivi des litiges tenu par le professeur Edward Lee sur le site chatgptiseatingtheworld.com, la carte des États-Unis d'Amérique faisait état de 156 actions en justice au 22 août 2026, auxquelles s'ajoutaient environ 30 autres actions enregistrées dans d'autres juridictions. Ce chiffre est sujet à évolution et correspond à la date de la présente étude. Voir également la carte des litiges mise à jour par la Copyright Alliance.

abordé dans la première partie de la présente étude. Trois décisions rendues au cours des 18 derniers mois apportent des réponses très différentes.

Dans l'affaire *GEMA* c*. OpenAI*, évoquée ci-dessus dans la section B.5 de la deuxième partie, le Landgericht München I a jugé en novembre 2025 que l'entraînement d'un modèle linguistique à partir de paroles de chansons, ainsi que la mémorisation ultérieure de ces paroles par le modèle, constituaient des actes de reproduction au sens de la transposition allemande de la directive "InfoSoc", et le tribunal a rejeté l'application de l'exception prévue à l'article 4 de la CDSM[109].

Sur une question similaire concernant la notion de modèle en tant que copie et portant sur une technologie différente, la Haute Cour de justice d'Angleterre et du Pays de Galles était parvenue à une conclusion différente quelques jours auparavant. Dans l'affaire *Getty Images (US) Inc.* c*. Stability AI Ltd*., la juge Joanna Smith DBE a estimé, le 4 novembre 2025, que les poids de Stable Diffusion, un modèle de diffusion latente utilisé pour la génération d'images, ne constituaient pas des copies contrefaites des photographies sur lesquelles le modèle avait été entraîné[110]. Le demandeur a renoncé à ses demandes relatives à l'atteinte primaire et au droit sur les bases de données à l'issue du procès, reconnaissant que les éléments de preuve ne permettaient pas d'établir que l'entraînement avait eu lieu au Royaume-Uni; le litige portait donc sur l'atteinte secondaire, ce qui a conduit le tribunal à déterminer si le modèle constituait un article qui était lui-même une copie illicite. Le tribunal a estimé que les preuves techniques présentées lors du procès démontraient que les poids du modèle de diffusion ne stockaient pas les données d'entraînement et que, en l'absence de preuve de mémorisation au sens pertinent du terme, le modèle ne constituait pas une copie illicite[111]. Le tribunal a ajouté, à titre d'observation incidente, que si elle avait conclu autrement, Stability AI aurait été tenue responsable au titre du régime de l'atteinte secondaire, car le développeur savait ou avait des raisons de croire que le corpus d'entraînement contenait des œuvres extraites du Web sans autorisation[112].

Une troisième décision, le jugement rendu par le tribunal de l'Internet de Hangzhou dans l'affaire dite "Hangzhou Ultraman", porte sur un état de fait similaire, mais distinct. Dans cette affaire, le défendeur était une plateforme hébergeant des modèles d'IA open-source auxquels les utilisateurs avaient accès; certains utilisateurs avaient entraîné ces modèles à partir du manga japonais "Ultraman", protégé par le droit d'auteur, et avaient mis en ligne les modèles ainsi entraînés à l'intention d'autres utilisateurs. Le titulaire du droit d'auteur a intenté une action contre la plateforme, et le tribunal a jugé cette dernière responsable de l'atteinte par complicité. Dans une opinion incidente, le tribunal a estimé que l'acte d'entraînement en lui-même pouvait être considéré comme relevant de l'usage loyal au regard du droit chinois, au motif que l'utilisation d'œuvres protégées par le droit d'auteur dans le cadre de l'entraînement d'un modèle d'IA générative a pour objectif d'apprendre et d'analyser des traits linguistiques, des caractéristiques stylistiques et d'autres éléments, plutôt que de reproduire l'expression créative originale. Le tribunal n'a pas eu à se prononcer sur ce point et son raisonnement

---

109 *GEMA* c*. OpenAI*, Tribunal régional de Munich I, 11 novembre 2025 (42 O 14139/24), jugement définitif. Cette décision est également abordée dans la section B.5 de la deuxième partie.

110 *Getty Images (US) Inc.* c*. Stability AI Ltd.*, [2025] EWHC 2863 (Ch) (juge Smith, 4 novembre 2025). La décision porte sur des revendications relatives à la fois au droit d'auteur et à la marque; la présente analyse ne porte que sur le droit d'auteur. L'autorisation d'interjeter appel de la décision relative à l'atteinte secondaire a été accordée en décembre 2025 ([2025] EWHC 3343 (Ch)); l'appel était en instance à la date d'établissement de la présente étude.

111 Concernant l'architecture des modèles de diffusion, voir la section B.1 de la première partie, ainsi que les références qui y sont citées.

112 Pour plus d'informations, voir E. Derclaye, "*Getty Images* c*. Stability AI* : un arrêt historique qui souligne la nécessité pour le Gouvernement britannique de modifier sa législation sur le droit d'auteur", Kluwer Copyright Blog (20 novembre 2025), disponible en ligne : https://legalblogs.wolterskluwer.com/copyright-blog/getty-images-v-stability-ai-a-landmark-judgment-reinforcing-the-need-for-the-uk-government-to-amend-its-copyright-laws/.

n'avait pas d'incidence sur l'issue de l'affaire, mais dans son opinion incidente, il a considéré que l'entraînement pouvait potentiellement relever de l'usage loyal en vertu du droit chinois[113].

Ces trois décisions ne sont pas tout à fait cohérentes d'un point de vue purement technique, car elles concernent différents types de modèles (un modèle linguistique de type "Transformer" à Munich, un modèle de diffusion latente à Londres, une plateforme hébergeant des modèles de génération d'images affinés par les utilisateurs à Hangzhou) et différentes approches procédurales. Les tribunaux de Munich et de Londres se sont tous deux penchés sur la littérature informatique consacrée à la mémorisation dans les grands modèles génératifs. Cela inclut notamment les travaux menés par les professeurs Cooper, Grimmelmann, Lemley et d'autres sur l'extraction de contenus mémorisés à partir de modèles à poids libre[114]. Cette même jurisprudence a été interprétée par le tribunal de Munich comme établissant que l'entraînement équivaut à une reproduction à grande échelle, et par le tribunal de Londres comme venant étayer la conclusion selon laquelle les poids des modèles de diffusion ne constituent pas, en règle générale, des copies illicites. Cette comparaison montre que les données empiriques disponibles sur la mémorisation ne suffisent pas à elles seules à déterminer pleinement le résultat de l'analyse juridique. La question de savoir si la capacité démontrée d'un modèle à reproduire des séquences issues de ses données d'entraînement doit être analysée comme la preuve que le modèle est une copie du corpus d'entraînement, ou comme un défaut de comportement devant être corrigé par des filtres de sortie et des techniques de "désapprentissage", relève d'une qualification juridique sur laquelle différents tribunaux sont parvenus à des conclusions divergentes, sur la base de dossiers probatoires qui, bien que différents quant à leur nature et aux éléments de preuve effectivement présentés à chaque tribunal, s'appuyaient sur une littérature scientifique sous-jacente similaire[115]. Comme indiqué dans la première partie, la question de savoir si un modèle contient une copie d'œuvres protégées ou d'objets visés par des droits connexes doit être tranchée au cas par cas.

La question du "modèle en tant que copie" a été soulevée sous une forme différente, distincte sur le plan procédural, aux États-Unis d'Amérique, où certains plaignants ont fait valoir non seulement que le processus d'entraînement produisait des copies, mais aussi que le modèle entraîné constituait en soi une œuvre dérivée portant atteinte à des droits. Dans l'affaire *Kadrey* c*. Meta Platforms, Inc.*, les auteurs ont fait valoir que les modèles LLaMA ne pouvaient pas fonctionner sans les informations expressives extraites de leurs livres, que ces informations étaient conservées dans les poids du modèle et que les modèles constituaient donc des œuvres dérivées non autorisées, une qualification qui engagerait la responsabilité du modèle lui-même sans qu'il soit nécessaire de prouver l'existence d'une correspondance entre un résultat particulier et les œuvres sur lesquelles le modèle avait été entraîné. Le tribunal a refusé d'admettre cette thèse au stade de la procédure écrite. Le défendeur a finalement obtenu gain de cause en invoquant le principe de l'usage loyal, abordé dans la section 4 ci-dessous, après avoir fait valoir que les poids constituaient des paramètres statistiques dérivés des données d'entraînement plutôt que des expressions stockées[116].

---

113 Tribunal d'Internet de Hangzhou, (2024) Zhe 0192 Min Chu n° 1587 (première instance); Tribunal populaire intermédiaire de Hangzhou (province du Zhejiang), (2024) Zhe 01 Min Zhong n° 10332 (deuxième instance).

114 A. Feder Cooper et J. Grimmelmann, "The Files are in the Computer: Copyright, Memorization, and Generative AI", document de travail SSRN n° 4803118 (2024). En ce qui concerne les recherches empiriques sur la mémorisation dans les modèles linguistiques, voir également l'étude de Cooper *et al.* citée dans la première partie, note de bas de page n° 18, de la présente étude. Voir également M. A. Lemley et A. F. Cooper, "Probabilistic 'Copies' in Generative AI Models", cité dans la première partie, note de bas de page n° 25, de la présente étude.

115 Pour une analyse critique du raisonnement de la GEMA sur ce point, voir G. Frosio, "Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI", Kluwer Copyright Blog (10 et 11 décembre 2025), parties 1 et 2.

116 Pour connaître la position des développeurs en dehors du contexte judiciaire, voir Google, "A Policy Agenda for Responsible Progress in Artificial Intelligence" (livre blanc sur la réglementation de l'IA), disponible à l'adresse https://blog.google/company-news/outreach-and-initiatives/public-policy/white-paper-ai-regulation/, consulté le 7 juillet 2026.

Une théorie similaire a été avancée, concernant un modèle de diffusion latent, dans l'affaire *Andersen* c*. Stability AI Ltd.*, où les plaignants ont qualifié le modèle de moyen alternatif de stockage de copies des images d'entraînement; cette qualification s'est heurtée à des difficultés en matière de preuve en l'absence d'un résultat correspondant étroitement à une œuvre protégée, une difficulté qui fait écho à la conclusion de la juridiction anglaise dans l'affaire Getty selon laquelle les poids d'un modèle de diffusion ne stockent pas les données d'entraînement. L'expérience des États-Unis d'Amérique suggère que le fait de prouver que des œuvres protégées ont été utilisées lors de l'entraînement ne suffit pas, à lui seul, à établir que le modèle entraîné est une œuvre dérivée de ces œuvres[117]. À la date de rédaction du présent document, aucune décision rendue en appel n'avait encore été publiée concernant l'une ou l'autre des questions susmentionnées.

2. L'exception relative à l'exploration de textes et de données

La deuxième question, qui se pose au sein de l'Union européenne et dans les ressorts juridiques dont la législation en matière de droit d'auteur a été influencée par le cadre européen (par exemple, les pays de l'EEE), est de savoir si les exceptions relatives à l'exploration de textes et de données prévues aux articles 3 et 4 de la directive CDSM s'étendent à l'entraînement des modèles d'IA générative. Le débat doctrinal de fond est abordé dans la section B.5 de la deuxième partie de la présente étude. La jurisprudence allemande en la matière est résumée ici par souci d'exhaustivité, l'accent étant mis sur les enseignements qu'elle apporte concernant l'exécution des demandes des titulaires de droits.

Les deux décisions allemandes rendues dans l'affaire *Kneschke* c*. LAION e.V.* concernant l'application de l'exception relative à l'exploration de textes et de données à l'entraînement des IA (en première instance, Landgericht de Hambourg, 27 septembre 2024, et en appel, Oberlandesgericht de Hambourg, 10 décembre 2025) sont examinées dans la section B.5 de la deuxième partie, ci-dessus.

Il en découle deux enseignements pour la mise en œuvre. Premièrement, dans les ressorts juridiques où l'exception relative à l'exploration de textes et de données s'applique à l'entraînement de l'IA, la stratégie de défense des droits du titulaire est conditionnée par le mécanisme de retrait. Lorsqu'un refus valable a été exprimé d'une manière permettant à un processus automatisé de le détecter et d'agir en conséquence, le titulaire des droits peut faire valoir ses droits au moyen des procédures habituelles en matière d'atteinte. En l'absence d'une telle clause de retrait, ou lorsque la réserve ne satisfait pas aux exigences applicables, l'exception relative à l'exploration de textes et de données peut s'appliquer, sous réserve de ses autres conditions, limitant ainsi toute action pour atteinte fondée sur le droit de reproduction. À cet égard, la capacité du titulaire des droits à se réserver ces derniers dépend en partie de la forme technique du signal de retrait.

Deuxièmement, la norme relative à la lisibilité par machine n'est pas figée. La Cour d'appel de Hambourg a expressément jugé que la lisibilité par machine devait être appréciée au regard des capacités technologiques existant au moment de l'utilisation litigieuse. Au vu des éléments du cas d'espèce, cela correspondait à l'état de la technique au second semestre 2021, avant que ne se généralisent les grands modèles linguistiques capables d'extraire du sens à partir de textes en langage naturel non structurés. La décision indique qu'une clause de retrait qui n'était pas lisible par machine en 2021 peut l'être aujourd'hui, et qu'une clause de retrait qui n'est pas lisible par machine en 2025 pourrait le devenir à une date ultérieure. Sur le plan pratique, cela

---

[117] Concernant la théorie des œuvres dérivées aux États-Unis d'Amérique, voir l'affaire *Kadrey* c*. Meta Platforms*, *Inc.*, n° 3:23-cv-03417-VC (N.D. Californie, juge Chhabria, 20 novembre 2023) (rejetant, au stade des conclusions, l'argument selon lequel les modèles LLaMA constituent en eux-mêmes des œuvres dérivées portant atteinte à des droits, tout en autorisant la poursuite de l'action directe pour atteinte fondée sur la copie à des fins d'entraînement); et *Andersen* c. *Stability AI Ltd*., n° 3:23-cv-00201-WHO (N.D. Californie, juge Orrick), ordonnances du 30 octobre 2023 et du 12 août 2024 (traitant, entre autres, de l'argument selon lequel le modèle stocke des copies compressées des images d'entraînement). La décision relative à l'usage loyal rendue dans l'affaire Kadrey est examinée dans la section A.4 ci-dessous.

signifie que les titulaires de droits pourraient devoir réexaminer et, dans de nombreux cas, reformuler leurs clauses de retrait à mesure que la technologie évolue; pour les développeurs en IA, cela signifie que la question de savoir si l'opération était licite au moment où elle a été effectuée n'est pas tranchée par une modification ultérieure des capacités concernées[118].

3. Compétence en matière d'entraînement hors du territoire national

Une troisième question, dont l'importance pratique ne cesse de croître, concerne la compétence des tribunaux nationaux à l'égard des opérations d'entraînement menées sur des infrastructures situées dans un autre pays. La principale décision canadienne sur cette question, rendue dans l'affaire *Toronto Star Newspapers Ltd.* c*. OpenAI Inc.*, a été prononcée le 7 novembre 2025 par le juge Kimmel de la Cour supérieure de justice de l'Ontario[119]. Une alliance regroupant plusieurs grands groupes de presse canadiens avait intenté une action en justice contre OpenAI, alléguant que le défendeur avait utilisé leurs contenus pour entraîner ChatGPT, en violation de la loi canadienne sur le droit d'auteur, des mesures techniques de protection de ces contenus et des conditions d'utilisation des sites Web concernés. OpenAI a demandé le rejet de l'affaire pour défaut de compétence, faisant valoir que les opérations d'entraînement avaient été effectuées sur des serveurs situés aux États-Unis d'Amérique et qu'elles ne relevaient donc pas du champ d'application territorial de la loi canadienne sur le droit d'auteur. Le tribunal a rejeté cet argument, estimant qu'il existait un lien réel et substantiel entre la procédure et l'Ontario et que la question de savoir si les actes allégués relevaient du droit canadien du droit d'auteur relevait du fond plutôt que de l'analyse de la compétence. La procédure est passée à la phase d'instruction au début de l'année 2026; les défendeurs ont interjeté appel de la décision relative à la compétence, et cet appel était encore en instance à la date d'établissement de la présente étude.

La portée du critère du "lien réel et substantiel" en matière d'entraînement de l'IA n'est pas encore clairement établie. Des commentaires universitaires sur l'arrêt "Toronto Star" ont souligné que ce critère a été élaboré en droit canadien dans le cadre du droit de communication au public, dans des affaires telles que *SOCAN* c*. CAIP* et *SOCAN* c*. ESA*, et que sa transposition au droit de reproduction n'est pas nécessairement aisée[120]. Le droit de reproduction est traditionnellement attribué, à des fins juridictionnelles, au lieu où l'acte de reproduction ou de fixation a lieu. La relation entre le champ d'application territorial du droit de reproduction et le principe du traitement national prévu à l'article 5 de la Convention de Berne est abordée dans la section B.2 ci-dessous.

D'autres évolutions observées au Canada illustrent la diversité des réponses procédurales possibles. Dans l'affaire *CanLII* c*. Caseway AI*, CanLII avait intenté une action en justice contre la start-up Caseway AI, basée à Vancouver, en 2024, pour le téléchargement massif de sa base de données de décisions judiciaires dans le but d'entraîner un assistant IA spécialisé dans l'information juridique. CanLII a fait valoir que, bien que les décisions sous-jacentes relèvent du domaine public, la structure organisée, les métadonnées et la sélection contenues dans sa base de données constituent une compilation protégée par le droit d'auteur. Les parties ont

---

118 P. Keller, "LAIN Round 2 : lisible par machine, mais toujours pas exploitable – L'absence de progrès concernant les demandes d'exercice du droit de retrait à l'égard de l'exploration de textes et de données", Kluwer Copyright Blog (17 et 18 décembre 2025), parties 1 et 2.

119 *Toronto Star Newspapers Ltd.* c*. OpenAI Inc.*, 2025 ONSC 6217 (J. Kimmel, 7 novembre 2025). Les plaignants sont le Toronto Star, le Globe and Mail, Postmedia et la Société canadienne de radiodiffusion. La requête en irrecevabilité pour défaut de compétence reposait sur l'argument selon lequel les actes d'entraînement allégués avaient été effectués sur des serveurs situés aux États-Unis d'Amérique et se trouvaient donc hors du champ d'application territorial de la loi canadienne sur le droit d'auteur.

120 *Société des compositeurs, auteurs et éditeurs de musique du Canada* c*. Association canadienne des fournisseurs d'accès à l'Internet*, 2004 CSC 45 ("*SOCAN* c*. CAIP*"); voir également *Société des compositeurs, auteurs et éditeurs de musique du Canada* c*. Entertainment Software Association*, 2022 CSC 30. L'observation doctrinale selon laquelle ce critère a été élaboré dans le cadre du droit de communication au public et pourrait ne pas être facilement transposable au droit de reproduction a été avancée dans les commentaires universitaires consacrés à l'arrêt "Toronto Star".

conclu un accord à l'amiable en mars 2026, selon des modalités confidentielles; cet accord ayant été conclu, l'affaire n'a donné lieu à aucune décision judiciaire sur le fond de ces questions. Des recours collectifs sont également en cours : en Colombie-Britannique, l'auteur J.B. MacKinnon a intenté des recours collectifs contre Meta, Nvidia, Anthropic et d'autres sociétés, alléguant que celles-ci ont utilisé des copies illicites d'œuvres (issues de l'ensemble de données Books3) pour entraîner leurs modèles; au Québec, Anne Robillard a déposé une demande d'autorisation de recours collectifs contre Meta et OpenAI, réclamant des dommages-intérêts légaux pouvant atteindre 20 000 dollars canadiens par œuvre, soit le plafond légal prévu en cas d'atteinte à des fins commerciales. Cette même contrainte structurelle s'est avérée déterminante dans l'affaire Getty au Royaume-Uni, évoquée à la section A.1 ci-dessus, où les principaux chefs d'accusation d'atteinte à des droits ont été abandonnés dès lors que les éléments de preuve n'ont pas permis d'établir que l'entraînement avait eu lieu au Royaume-Uni[121].

En Inde, la Haute Cour de Delhi a confirmé sa compétence sur un entraînement effectué sur des serveurs situés aux États-Unis d'Amérique dans l'affaire *ANI Media Pvt. Ltd.* c*. OpenAI OPCO LLC*, estimant, à première vue, que le stockage des œuvres à l'étranger ne constituait que la "dernière étape d'une chaîne d'événements" commençant par un accès depuis l'Inde, et que l'objection du développeur était "indéfendable, car elle conduirait à un contournement de la loi indienne sur le droit d'auteur"[122]; le fond de cette décision est examiné dans la section A.4 ci-dessous. Au Brésil, la question de l'entraînement extraterritorial a été soulevée dans le cadre de la procédure engagée par le journal *Folha de São Paulo* contre OpenAI en août 2025, dans laquelle le journal affirme que le défendeur a utilisé ses contenus journalistiques pour entraîner ChatGPT et fournir des services à des utilisateurs brésiliens sans autorisation[123]. La procédure était soumise au secret judiciaire; les parties ont trouvé un accord à l'amiable en 2026 et aucune décision sur le fond n'a été rendue. En Argentine, Editorial Perfil, un important éditeur de presse, a intenté une action en justice contre OpenAI et Microsoft en août 2026 devant les tribunaux fédéraux civils et commerciaux, invoquant l'utilisation non autorisée de son contenu journalistique dans le cadre de l'entraînement et de l'exploitation de services d'IA générative[124]. De manière plus générale, un autre argument, qui n'a pas encore été examiné par les tribunaux brésiliens dans le contexte de l'IA, a été avancé par analogie avec la jurisprudence de la Cour suprême fédérale concernant les plateformes numériques

---

121 L'affaire *CanLII* c*. Caseway AI* s'est conclue par un accord à l'amiable confidentiel en mars 2026. Les recours collectifs intentés par J.B. Les actions intentées par MacKinnon en Colombie-Britannique contre Meta, Nvidia, Anthropic et Databricks, ainsi que celles intentées par Anne Robillard au Québec contre Meta et OpenAI, en sont au stade de la certification à la date de la présente étude. En ce qui concerne la peine maximale prévue par la loi en cas d'atteinte à des fins commerciales, voir la loi sur le droit d'auteur, L.R.C. 1985, chap. C-42, article 38.1, paragraphe 1, alinéa a).

122 *ANI Media Pvt. Ltd.* c*. OpenAI OPCO LLC*, Haute Cour de Delhi, CS(COMM) 1028/2024 (juge Bansal, arrêt du 24 juillet 2026), 2026 SCC OnLine Del 5291, paragraphe 51 pour les passages cités dans le texte. La décision a été rendue à la suite de la demande d'injonction provisoire déposée par le demandeur; l'affaire est toujours en cours et, à la date de la présente étude, un pourvoi était en instance devant une chambre collégiale de la Haute Cour.

123 *Folha de São Paulo* c*. OpenAI*, affaire n° 1107237-96.2025.8.26.0100 (TJSP). L'affaire a été portée devant les tribunaux en août 2025 et est traitée dans le cadre du secret judiciaire; aucune décision publique n'avait été rendue à la date de la présente étude.

124 *Editorial Perfil S.A.* c*. OpenAI, L.L.C.*, Microsoft Corporation et Microsoft de Argentina S.R.L. (Tribunaux fédéraux civils et commerciaux, Argentine, déposée le 24 août 2026). Au moment de la rédaction du présent article, peu d'informations publiques étaient disponibles concernant cette plainte, mais voir l'éditorial de Perfil, "Perfil poursuit OpenAI et Microsoft pour utilisation non autorisée de contenus journalistiques" (24 août 2026) <https://www.perfil.com/noticias/judiciales/perfil-demanda-a-openai-y-microsoft-por-el-uso-de-contenidos-periodisticos-sin-autorizacion.phtml> consulté le 29 septembre 2026. Voir également EFE, "Le quotidien argentin Perfil poursuit OpenAI et Microsoft pour utilisation abusive de ses contenus" (25 août 2026) <https://www.swissinfo.ch/spa/el-diario-argentino-perfil-demanda-a-openai-y-microsoft-por-uso-indebido-de-sus-contenidos/91953813> consulté le 29 septembre 2026.

étrangères opérant au Brésil[125]. Selon cet argument, la fourniture de services d'IA au Brésil soumet l'entraînement sous-jacent à la législation brésilienne sur le droit d'auteur, car la production de résultats qui en découle par des utilisateurs brésiliens constitue une prolongation de l'acte d'entraînement; de ce point de vue, si l'entraînement a eu lieu de manière illicite, la fourniture du service qui en dépend relève de la législation brésilienne sur le droit d'auteur en vertu du principe du traitement national prévu à l'article 5 de la Convention de Berne[126]. Cet argument revient, en substance, à une adaptation au contexte technologique – où la fixation pertinente intervient en amont de l'effet perçu par l'utilisateur – du principe selon lequel une violation qui trouve son origine à l'étranger, mais produit ses effets dans la juridiction du for peut faire l'objet d'une action en justice dans cette juridiction.

4. La question de la substitution

Une quatrième question porte sur la qualification juridique des résultats qui se substituent au matériel protégé sur lequel le modèle a été entraîné ou qui lui font concurrence. Cette question est essentielle, car elle détermine si le préjudice économique allégué par les titulaires de droits peut être pris en considération dans le cadre des catégories existantes du droit d'auteur, qui s'appliquent aux actes de reproduction et de communication au public, ou s'il nécessite au contraire un dispositif distinct, abordé dans la deuxième partie, de rémunération des effets de substitution, pour lesquels aucune production individuelle ne donne lieu à une action en justice. Les principes techniques sont exposés dans la section C.3 de la première partie.

Aux États-Unis d'Amérique, plusieurs tribunaux se sont déjà penchés sur cette question. Dans l'affaire *The New York Times Company* c. *Microsoft Corp.*, désormais regroupée sous l'intitulé *In re* OpenAI, Inc. Copyright Infringement Litigation, le défendeur a demandé, dès le début de la procédure, le rejet des chefs d'accusation d'atteinte directe au droit d'auteur, au motif que le demandeur avait utilisé ce que le défendeur qualifiait de "consignes élaborées et manipulées" pour provoquer la "régurgitation" qu'il dénonçait. En avril 2025, le tribunal fédéral de première instance du district sud de New York a refusé de rejeter ces demandes pour ce motif, et l'affaire est passée à la phase d'instruction. La phase d'instruction a été marquée par un litige concernant la communication des journaux d'utilisation destinés aux consommateurs. Dans une série d'ordonnances qui a abouti, le 5 janvier 2026, à la confirmation par le juge Stein d'une ordonnance antérieure rendue par le juge Wang, le tribunal a ordonné la production d'un échantillon aléatoire et anonymisé de 20 millions de journaux d'interactions d'utilisateurs de ChatGPT[127]. Cette décision déplace le centre de gravité de la preuve : il ne s'agit plus de déterminer si la régurgitation peut être provoquée par des incitations contradictoires, mais de savoir à quelle fréquence elle se produit dans le cadre d'une utilisation normale par le consommateur. Le tribunal a considéré que les éléments de preuve généraux et non contradictoires relatifs aux résultats générés par les utilisateurs étaient pertinents pour l'analyse de l'usage loyal, notamment pour déterminer si les résultats produits par le modèle constituaient des substituts commerciaux aux œuvres des plaignants.

Par la suite, dans deux décisions rendues en juin 2025 par le district nord de Californie, à deux jours d'intervalle, dans les affaires *Bartz* c. *Anthropic PBC* et *Kadrey* c. *Meta Platforms, Inc.*, deux tribunaux fédéraux de première instance sont parvenus à la même conclusion sur la question principale, mais en s'appuyant sur un raisonnement sensiblement différent. Dans

---

125 STF, RE 1037396 et RE 1057258 (affaires jointes). La Cour fédérale suprême a estimé que les plateformes pouvaient être tenues pour responsables des contenus publiés par les utilisateurs si elles ne donnaient pas suite à une notification, que les plateformes fournissant des services au Brésil devaient se conformer à la législation brésilienne et que ces entreprises devaient disposer d'un représentant légal dans le pays.

126 Convention de Berne pour la protection des œuvres littéraires et artistiques, article 5.1) (traitement national); voir également l'article 5.2) (indépendance de la protection).

127 *Affaire OpenAI, Inc. – Litige pour violation du droit d'auteur*, n° 1:25– md-03143-SHS-OTW (S.D.N.Y.). La procédure MDL regroupe plusieurs actions en justice intentées contre OpenAI par des organes de presse, des auteurs et d'autres plaignants. Concernant l'ordonnance de communication des 20 millions de journaux de ChatGPT, voir les références figurant dans la première partie, note de bas de page n° 20, de la présente étude.

l'affaire Bartz, le juge Alsup a estimé que l'entraînement d'un grand modèle linguistique à partir d'ouvrages acquis légalement constituait une transformation extrêmement importante et relevait de l'usage loyal au sens de l'article 107 de la loi sur le droit d'auteur; il a simultanément jugé que le téléchargement et la conservation de copies pirates de ces ouvrages provenant de LibGen et de PiLiMi ne relevaient pas de l'usage loyal, et le recours collectif a été certifié le 17 juillet 2025 uniquement en ce qui concerne les allégations de piratage. Les parties sont par la suite parvenues à un accord à l'amiable d'un montant de 1,5 milliard de dollars É.-U., qui a été définitivement approuvé en juillet 2026[128]. Dans l'affaire Kadrey, le juge Chhabria a lui aussi rendu un jugement sommaire en faveur du défendeur sur la question de l'usage loyal[129], mais son raisonnement a mis davantage l'accent sur le quatrième critère de l'usage loyal (préjudice au marché) et moins sur le premier (usage transformatif), reprenant ainsi la méthodologie de la Cour de cassation dans l'affaire *Andy Warhol Foundation* c*. Goldsmith.* Dans les commentaires ultérieurs, cette affaire a été interprétée comme reposant principalement sur l'incapacité du demandeur à apporter la preuve d'une substitution sur le marché, ce qui laisse entendre qu'un dossier probatoire différent aurait pu aboutir à une issue différente[130]. Ces deux décisions n'établissent pas de principe général selon lequel l'entraînement de l'IA relève ou non de l'usage loyal.

Dans l'affaire *Thomson Reuters Enterprise Centre GmbH* c*. Ross Intelligence Inc.*, la Cour d'appel des États-Unis d'Amérique pour le troisième circuit, confirmant un jugement sommaire partiel rendu en appel interlocutoire, a rendu la première décision d'appel américaine concernant l'usage loyal et l'entraînement des IA. Le défendeur avait utilisé les "sommaires" du demandeur, c'est-à-dire des résumés rédactionnels de points de droit préparés pour la plateforme de recherche Westlaw, afin d'entraîner un outil de recherche juridique non génératif fondé sur l'intelligence artificielle. En ce qui concerne le premier critère, le tribunal a estimé que l'utilisation était, au mieux, peu transformatrice, dans la mesure où elle servait essentiellement le même objectif de recherche juridique que l'utilisation des sommaires par le demandeur; cette conclusion, combinée au caractère commercial de l'utilisation, a fait pencher le premier critère en défaveur de l'usage loyal. Elle a distingué les cas de copie intermédiaire au motif que, dans ces cas-là, la copie était nécessaire pour accéder à des éléments fonctionnels non protégés du logiciel, alors qu'aucune nécessité comparable ne justifiait la copie des sommaires. Le deuxième facteur jouait en faveur du défendeur, mais, comme l'a fait remarquer le tribunal, il revêt rarement une importance significative; le troisième facteur allait à l'encontre de l'usage loyal, car des sommaires avaient été repris dans leur intégralité sans justification transformatrice. Concernant le quatrième facteur, le tribunal a constaté un préjudice sur trois plans : d'une part, sur la valeur des sommaires jurisprudentiels en tant qu'attrait pour les abonnements à la plateforme du demandeur; d'autre part, sur le marché initial des plateformes de recherche juridique, sur lequel le défendeur était en concurrence directe; et, à l'issue de son analyse de l'usage loyal, à un marché dérivé potentiel de concession de licences sur les sommaires de jurisprudence en tant que données d'entraînement pour l'IA, que le tribunal a qualifié de marché en plein essor, en observant que le demandeur utilisait ses propres sommaires de jurisprudence pour entraîner ses propres produits d'IA, que sa décision de ne pas accorder de licences à des tiers ne réfutait pas l'existence de ce marché, et que la copie

---

128 *Bartz* c*. Anthropic PBC*, n° 3:24– cv – 05417 (N.D. Californie, juge Alsup, 23 juin 2025) (accordant en partie le jugement sommaire en faveur du défendeur au titre de l'usage loyal à des fins d'entraînement sur des œuvres acquises légalement, et rejetant le jugement sommaire en faveur du défendeur en ce qui concerne le téléchargement et la conservation de copies pirates d'œuvres provenant de LibGen et PiLiMi). Le groupe a été certifié uniquement en ce qui concerne les demandes d'indemnisation liées au piratage. Un accord à l'amiable d'un montant de 1,5 milliard de dollars É.-U. a reçu une approbation préliminaire le 25 septembre 2025; l'audience finale sur l'équité de l'accord s'est tenue le 14 mai 2026 devant le juge Martínez-Olguín, qui a mis l'affaire en délibéré. À la date de la présente étude, l'autorisation définitive n'avait pas encore été accordée. L'accord a été définitivement approuvé par le tribunal de grande instance en juillet 2026, et l'affaire a été classée.

129 *Kadrey* c*. Meta Platforms, Inc.*, n° 3:23-cv-03417-VC (N.D. Californie, juge Chhabria, 25 juin 2025).

130 Concernant le contraste entre les raisonnements des deux juges, voir (parmi de nombreux commentaires) l'analyse publiée par Wiggin et Dana le 18 septembre 2025, disponible sur wiggin.com.

non autorisée privait le demandeur de la possibilité d'y pénétrer. Le tribunal a établi une distinction entre les modèles génératifs en cause dans les affaires Bartz et In re OpenAI, de sorte que l'appréciation du caractère transformatif dans ces affaires relatives aux modèles génératifs pourrait différer. Cette décision confirme néanmoins, en appel, qu'un marché émergent de concession de licences pour des contenus que les créateurs élaboreraient généralement eux-mêmes ou chargeraient des tiers d'élaborer constitue un élément à prendre en considération au titre du quatrième critère[131]. La vague d'accords à l'amiable dans le secteur de la musique n'a pas mis fin aux litiges dans ce domaine. En janvier 2026, un groupe d'éditeurs de musique a intenté une nouvelle action contre le développeur dans le cadre du litige Bartz, lui reprochant l'utilisation de plus de vingt mille œuvres musicales et réclamant des dommages-intérêts légaux supérieurs à 3 milliards de dollars É.-U.; parallèlement, des sociétés de gestion collective en Allemagne et au Danemark ont poursuivi leurs poursuites contre un autre développeur de logiciels de génération musicale[132].

En dehors des États-Unis d'Amérique, les tribunaux se sont également penchés sur cette question. L'affaire *Folha de São Paulo* c*. OpenAI*, évoquée plus haut, a soulevé cette question jusqu'à ce que les parties parviennent à un accord. Un exemple qui est toujours en cours au moment où nous écrivons ces lignes est l'affaire portée devant le tribunal de l'Internet de Beijing, dans laquelle quatre illustrateurs ont poursuivi Xiaohongshu en justice, alléguant que la plateforme avait utilisé leurs œuvres originales pour entraîner, sans autorisation, son modèle d'IA de génération d'images "Trik", et que ce modèle était capable de générer des images substantiellement similaires aux œuvres des plaignants[133].

En janvier 2025, en République de Corée, les trois principaux diffuseurs terrestres, KBS, MBC et SBS, ont poursuivi Naver devant le tribunal de grande instance de Séoul, alléguant que leurs contenus d'actualité avaient été utilisés sans autorisation pour entraîner la famille de modèles linguistiques HyperCLOVA de Naver. Il s'agit de la première procédure coréenne dans laquelle la question de l'entraînement est abordée de front et, compte tenu de l'origine commune de l'article 35-5 de la loi coréenne sur le droit d'auteur et de l'article 107 de la loi américaine sur le droit d'auteur, la jurisprudence américaine relative à l'usage loyal, examinée ci-dessus, fait

---

131 *Thomson Reuters Enterprise Centre GmbH* c*. Ross Intelligence Inc.*, n° 1:20-cv-613 (Tribunal fédéral de première instance du Delaware, juge Bibas, 11 février 2025) (accordant un jugement sommaire partiel au demandeur concernant l'atteinte directe et rejetant le moyen de défense fondé sur l'usage loyal). La décision du tribunal de première instance a été confirmée en appel interlocutoire : n° 25-2153 (3e Cir., audience du 11 juin 2026, décision du 29 septembre 2026) (arrêt rédigé par la juge Montgomery-Reeves). Dans une note de bas de page, la cour d'appel a fait référence à la déclaration d'intérêt déposée par le Ministère de la Justice des États-Unis d'Amérique dans l'affaire In re OpenAI, Inc. Litige pour atteinte au droit d'auteur, n° 1:25– md-3 143 (S.D.N.Y., 1er septembre 2026), dans laquelle elle faisait valoir que l'entraînement d'un grand modèle linguistique génératif constituait une utilisation transformatrice, et distinguait le contexte génératif de l'affaire dont elle était saisie.

132 Universal Music Publishing Group, Concord et ABKCO ont intenté une deuxième action en justice contre Anthropic devant le tribunal fédéral de première instance du district nord de Californie le 28 janvier 2026, invoquant une atteinte à plus de 20 000 œuvres. Cette action est distincte de la première action intentée en 2023 par les éditeurs contre Anthropic concernant des paroles de chansons, qui avait été introduite devant le tribunal du district central du Tennessee avant d'être transférée au tribunal du district nord de Californie. Voir Music Business Worldwide, "UMG, Concord et ABKCO poursuivent Anthropic pour 3 milliards de dollars É.-U. dans ce qui pourrait être la plus grande affaire de droit d'auteur hors recours collectif" (28 janvier 2026) <https://www.musicbusinessworldwide.com/umg-concord-and-abkco-sue-anthropic-for-3bn-in-what-could-be-single-largest-non-class-act/> consulté le 3 juillet 2026; concernant l'affaire de 2023, Rolling Stone, "Des éditeurs de musique poursuivent une plateforme d'IA soutenue par Amazon pour atteinte au droit d'auteur" (18 octobre 2023) <https://www.rollingstone.com/music/music-news/umg-concord-abkco-sue-anthropic-ai-company-copyright-infringement-1234856761/> consulté le 3 juillet 2026; Music Business Worldwide, "Universal, Concord et ABKCO demandent au tribunal une injonction pour empêcher Anthropic AI d'utiliser leurs paroles" (16 novembre 2023) <https://www.musicbusinessworldwide.com/universal-concord-and-abkco-ask-court-for-injunction-to-stop-anthropic-ai-from-using-their-lyrics/> consulté le 3 juillet 2026. La GEMA (Allemagne) et la Koda (Danemark), entre autres, ont poursuivi leurs actions en justice contre Suno malgré les accords conclus par les maisons de disques avec ce développeur et avec Udio fin 2025.

133 Tribunal de l'Internet de Beijing (affaire déposée en novembre 2023; en instance à la date de la présente étude); les quatre plaignants sont des illustrateurs qui affirment que Xiaohongshu a utilisé leurs œuvres originales pour entraîner, sans autorisation, le modèle d'IA de génération d'images Trik.

partie des arguments avancés devant le tribunal coréen[134]. Au Brésil, un doubleur a poursuivi en justice un centre commercial, affirmant que sa voix avait été utilisée sans autorisation dans des publicités; le centre commercial a rétorqué que la voix en question avait été générée par l'intelligence artificielle, et cet argument a été retenu par le tribunal de première instance. En 2024, toutefois, la Cour d'appel de São Paulo a infirmé cette décision, estimant que "l'utilisation de l'intelligence artificielle, en soi, n'élimine pas (bien au contraire, elle l'aggrave) le risque d'atteinte aux droits de tiers", étant donné que ces outils sont entraînés à l'aide de contenus extraits du Web par moissonnage de données, sans l'autorisation des titulaires des droits. Le tribunal a également estimé qu'il convenait de recueillir des éléments de preuve supplémentaires, et l'affaire a été renvoyée devant la juridiction inférieure; à la date de la présente étude, celle-ci n'avait pas encore rendu de décision définitive sur le fond[135].

En Inde, en revanche, la question de l'entraînement a déjà fait l'objet de l'examen judiciaire le plus approfondi à ce jour en dehors des États-Unis d'Amérique et de l'Allemagne. Dans l'affaire *ANI Media Pvt. Ltd.* c. *OpenAI OPCO LLC*, une agence de presse a allégué que ses œuvres avaient été utilisées sans autorisation tant pour l'entraînement des modèles linguistiques du défendeur que pour la génération de contenus. Se prononçant sur la demande d'injonction provisoire présentée par l'agence, et après avoir confirmé sa compétence territoriale sur les entraînements effectués sur des serveurs situés aux États-Unis d'Amérique (voir la section A.3 ci-dessus), la Haute Cour de Delhi a estimé que l'entraînement relevait du droit de reproduction, puisque l'article 14.a)i) de la loi de 1957 sur le droit d'auteur étend ce droit au stockage d'une œuvre "sur tout support par des moyens électroniques" et que la suppression des copies après l'entraînement n'annule pas le stockage. Le tribunal a toutefois estimé, à première vue, que ce stockage relevait de la disposition relative à l'utilisation loyale prévue à l'article 52.1)a), qui couvre les utilisations à des fins "d'usage privé ou personnel, y compris la recherche". Le terme "recherche" a été interprété dans un esprit d'actualisation : "Le processus d'entraînement des grands modèles de langage qui sous-tendent ChatGPT peut être considéré comme une forme de recherche visant à générer de nouvelles connaissances", et "les activités de recherche ne peuvent se limiter aux seules actions menées par les êtres humains", car "en fin de compte, la recherche est menée à la demande des humains et dans leur intérêt". Le caractère commercial du défendeur n'a pas non plus permis de faire sortir cette utilisation du champ d'application de la disposition : le tribunal a qualifié cette utilisation de "purement privée", les exemplaires destinés à l'entraînement restant dans un environnement fermé accessible aux modèles plutôt qu'au public, et l'article 52.1)a) ne prévoyant aucune exclusion des utilisations commerciales.

En ce qui concerne le critère d'équité, le tribunal a estimé que l'utilisation des œuvres se limitait à l'entraînement des modèles, que le dossier ne contenait aucune preuve de substitution sur le marché ou de perte de revenus au-delà de simples allégations, et que les avantages de cette technologie pour le public pesaient dans la balance; il a également noté que le demandeur aurait pu bloquer les robots d'indexation du défendeur, mais qu'il ne l'avait pas fait. L'injonction provisoire a été rejetée au motif supplémentaire que tout préjudice, s'il venait à être établi lors du procès, pourrait faire l'objet d'une indemnisation sous forme de dommages-intérêts, le tribunal ayant fait remarquer que l'obligation d'obtenir des licences auprès de chaque source d'information comparable rendrait l'élaboration de grands modèles de langage économiquement non viable. Il s'agit d'une décision provisoire : l'affaire sera jugée en première instance et, à la date de la présente étude, un pourvoi était en instance devant une chambre

---

134 *KBS, MBC et SBS* c. *Naver Corp.*, Tribunal de grande instance de Séoul (dépôt de la plainte le 13 janvier 2025). Les actions en justice sont intentées en vertu de la loi sur le droit d'auteur et de la loi sur la prévention de la concurrence déloyale; les plaignants réclament des dommages-intérêts et une injonction visant à empêcher toute utilisation future. Ces mêmes diffuseurs ont intenté une action similaire contre OpenAI devant le même tribunal en février 2026.

135 Tribunal de justice de São Paulo, appel civil n° 1119021-41.2023.8.26.0100 (6e chambre de droit privé, juge rapporteur Costa Netto, arrêt du 31 octobre 2024), présenté lors du Forum des juges en matière de propriété intellectuelle de l'OMPI de 2025; voir <https://www.wipo.int/wipolex/en/judgments/details/3217>, consulté le 19 septembre 2026.

collégiale. Son intérêt comparatif n'en reste pas moins considérable. Alors que les tribunaux allemands se sont demandé si une exception relative à l'exploration de textes et de données à des fins spécifiques s'appliquait à l'entraînement de l'IA, et que les tribunaux américains se sont interrogés sur la question de savoir si cet entraînement relevait de l'usage loyal, la Haute Cour de Delhi a abordé la même question à travers une clause d'utilisation loyale énumérant "l'usage privé ou personnel, y compris la recherche" et y a répondu, à ce stade, en faveur du développeur, en s'appuyant sur un raisonnement (l'absence de préjudice démontré pour le marché, l'existence d'une option de retrait non utilisée, les avantages publics de la technologie), qui s'apparentent aux facteurs pris en compte dans les décisions examinées ci-dessus aux États-Unis d'Amérique.

SECTION B. APPLICATION DES DROITS FACE AUX DÉVELOPPEURS D'IA

La jurisprudence passée en revue dans la section A illustre les évolutions de fond intervenues au cours des 18 derniers mois. La section B aborde la question structurelle de savoir comment les titulaires de droits peuvent faire valoir leurs prétentions à l'encontre des développeurs d'IA, compte tenu du caractère transfrontière de cette technologie et des asymétries entre les parties généralement impliquées dans ce type de litiges.

1. Caractéristiques structurelles du domaine de l'application des droits

Trois caractéristiques structurelles reviennent régulièrement. La première est la persistance et la gravité du problème transfrontière mis en évidence à la section A.3 ci-dessus. L'entraînement d'un modèle d'IA générative s'effectue généralement sur une infrastructure située dans un autre ressort juridique que celui du titulaire des droits; le modèle est ensuite déployé et utilisé simultanément dans de nombreux ressorts juridiques; les résultats sont générés dans le ressort juridique de chaque utilisateur. Chacune de ces étapes peut relever du droit d'un pays différent. À ce jour, la jurisprudence n'a pas encore apporté de réponse uniforme à la question de savoir comment ces différents régimes juridiques nationaux interagissent, et les développeurs peuvent se trouver confrontés à des obligations incohérentes selon les ressorts juridiques dans lesquels ils proposent leurs services.

La deuxième concerne le rôle du principe du traitement national prévu à l'article 5.1) de la Convention de Berne. En vertu de ce principe, les auteurs bénéficient, dans les pays de l'Union autres que le pays d'origine de l'œuvre, des droits que les législations de ces pays accordent à leurs propres ressortissants, ainsi que des droits spécialement accordés par la Convention. Ce principe relève de la non-discrimination et ne détermine pas en soi le champ d'application territorial du droit national.

La troisième est le recours à des mécanismes procéduraux permettant le regroupement des demandes, dont les procédures judiciaires multidistricts aux États-Unis d'Amérique constituent le principal exemple. Le regroupement, au sein du district sud de New York, des actions intentées contre OpenAI par des organes de presse, des auteurs et d'autres plaignants au sein d'une seule et même procédure judiciaire multidistricts a permis de centraliser des procédures qui, sans cela, auraient été dispersées entre différentes actions individuelles. De même, l'action de groupe Bartz a regroupé des plaintes concernant des copies pirates d'environ 482 000 livres et a abouti à un accord à l'amiable d'une ampleur considérable, comme indiqué plus haut. Dans les ressorts juridiques où les mécanismes d'action de groupe (ou "action de masse") ou les procédures judiciaires multidistricts n'existent pas ou sont moins développés, le coût d'une action individuelle peut s'avérer prohibitif, en particulier pour les auteurs individuels et les petits titulaires de droits. Les actions de groupe au Québec décrites à la section A.3 constituent une tentative de mettre en œuvre ce même mécanisme dans un cadre procédural différent; les recours collectifs de la Colombie-Britannique présentent un tableau comparable. Dans les ressorts juridiques dépourvus de tels mécanismes, la capacité concrète des titulaires de droits individuels à faire valoir leurs droits face à des promoteurs disposant de moyens importants est considérablement limitée.

2. Asymétries entre les parties

Au-delà de ces caractéristiques, la jurisprudence met en évidence un certain nombre d'asymétries qui compliquent l'application des droits dans la pratique. Il s'agit tout d'abord de l'asymétrie d'information : les développeurs détiennent les documents techniques (manifestes des ensembles de données, journaux d'exploration, configurations d'entraînement) à partir desquels la composition du corpus d'entraînement peut en principe être établie, même lorsque ces documents sont incomplets; les titulaires de droits ne disposent d'aucun accès comparable, sauf lorsque l'extraction ou d'autres démonstrations empiriques fournissent des preuves indirectes. C'est là l'objet de la réflexion sur la transparence présentée dans la section C.1 ci-dessous. La deuxième est l'asymétrie des ressources : les développeurs, en particulier ceux qui développent des grands modèles frontières, disposent de ressources supérieures à celles de la plupart des titulaires de droits. L'organisation d'une action longue pour l'application des droits a un coût considérable compte tenu des expertises techniques désormais requises dans ce type d'affaires. La troisième est l'incertitude sur le plan de la doctrine : lorsque le droit n'est pas clairement établi, la jurisprudence publiée peut être influencée par des effets de sélection, car les parties règlent certains litiges à l'amiable et en portent d'autres devant les tribunaux, tandis que la capacité des titulaires de droits à engager des actions à titre de test est limitée par les coûts de la procédure. La jurisprudence qui en résulte doit donc être analysée en tenant compte de ces effets de sélection.

SECTION C. DÉFIS DANS LE DOMAINE DE L'APPLICATION DES DROITS

Dans l'état actuel du droit et de la technologie, les titulaires de droits qui cherchent à faire valoir leurs droits à l'encontre des développeurs d'IA sont confrontés à trois défis majeurs. Ceux-ci sont abordés les uns après les autres.

1. Transparence

Le premier défi réside dans le manque de transparence concernant le processus d'entraînement. À quelques rares exceptions près, les développeurs d'IA n'ont pas divulgué la composition de leurs corpus d'entraînement; lorsque des informations sommaires ont été fournies, celles-ci étaient d'un niveau de généralité tel qu'elles ne permettent pas à un titulaire de droits de déterminer si son œuvre a été utilisée. Cette opacité est renforcée par des revendications de secrets d'affaires : début 2026, par exemple, le développeur d'un système de génération de musique a demandé que même la taille de son ensemble de données d'entraînement reste confidentielle dans le cadre d'un litige intenté par de grandes maisons de disques, invoquant un préjudice concurrentiel[136]. L'une des raisons tient au fait que la composition du corpus d'entraînement peut constituer un atout commercial, et que la divulgation exhaustive de son contenu pourrait nuire à la position concurrentielle du développeur. Il en résulte une asymétrie d'information susceptible de limiter la capacité concrète des titulaires de droits à identifier les atteintes potentielles et à engager des actions pour faire appliquer leurs droits.

L'Union européenne a réglé cette question en instaurant une obligation légale expresse de divulgation. L'article 53.1)d) du règlement sur l'IA impose aux fournisseurs de modèles d'IA à usage général d'élaborer et de mettre à la disposition du public un "résumé suffisamment détaillé" du contenu utilisé pour l'entraînement du modèle, conformément à un modèle fourni par le Bureau de l'IA de l'UE[137]. La mise en œuvre dépend en partie du niveau de détail jugé suffisant au regard de l'objectif de la disposition. Cette disposition est en outre assortie d'une reconnaissance explicite de la nécessité de respecter les secrets d'affaires et les informations

---

136 Music Business Worldwide, "Suno moves to keep size of its AI training data sealed in UMG and Sony's copyright case, citing competitive harm" (2026) <https://www.musicbusinessworldwide.com/suno-moves-to-keep-size-of-its-ai-training-data-sealed-in-umg-and-sonys-copyright-case-citing-competitive-harm/> consulté le 7 juillet 2026.

137 Règlement (UE) 2024/1 689, article 53.1)d), mentionné dans la partie II, note de bas de page n° 82 de la présente étude.

commerciales confidentielles, et la mise en œuvre de cette exigence fait désormais l'objet de travaux en cours en matière de politique et de normalisation[138]. Cette tension entre l'obligation de divulgation et l'exception relative au secret d'affaires fait l'objet d'un débat animé. La portée opérationnelle finale de cette obligation dépendra probablement à la fois de l'application du modèle publié par le Bureau de l'IA en juillet 2025 et de l'interprétation que les autorités judiciaires nationales et autres donneront à cette exception.

D'autres ressorts juridiques ont abordé la question de la transparence à l'aide de différents instruments. En Chine, un régime complet d'obligations d'étiquetage s'applique aux contenus générés par l'IA[139]. Le régime chinois d'étiquetage ne vise pas principalement la protection du droit d'auteur, mais la prévention de la fraude, de la tromperie et de la confusion du public; néanmoins, cela a eu pour effet concret, pour les titulaires de droits, de permettre un certain niveau de raisonnement à rebours, puisque l'étiquetage des contenus générés par l'IA permet d'identifier des résultats particuliers comme étant générés par l'IA et de déterminer, à partir de ces résultats, si des œuvres protégées sont susceptibles d'avoir été utilisées lors de l'entraînement.

En République de Corée, la loi-cadre relative au développement de l'intelligence artificielle et à la mise en place d'un cadre de confiance, en vigueur depuis le 22 janvier 2026, impose d'informer au préalable les utilisateurs des services d'IA générative et exige l'étiquetage des contenus produits par l'IA[140]. Les obligations d'étiquetage, à l'instar de celles en vigueur en Chine, ne visent pas en premier lieu à protéger le droit d'auteur; elles peuvent toutefois permettre de tirer le même type de raisonnement à rebours à partir des résultats. Les questions relatives au droit d'auteur sont régies par la loi sur le droit d'auteur plutôt que par la législation-cadre, une répartition entre ces deux instruments qui reflète la relation, au sein de l'Union européenne, entre la loi sur l'IA et l'acquis en matière de droit d'auteur.

Au Brésil, la réponse législative figure dans le projet de loi n° 2338/2023 qui, dans sa version actuelle, oblige les développeurs d'IA à divulguer l'utilisation de contenus protégés dans le cadre du développement de leurs systèmes par la publication d'un résumé; cette obligation est soumise à une réserve visant à garantir le respect des secrets commerciaux et industriels[141]. L'incidence concrète de l'obligation de divulgation dépendra de l'étendue de l'interprétation qui sera donnée à l'exception relative au secret d'affaires. Cette même observation s'applique, mutatis mutandis, au cadre européen.

Deux observations portent sur la transparence en tant que question liée à l'application des droits. Premièrement, la transparence peut faciliter l'application des droits, mais elle n'est pas suffisante en soi. La publication d'un résumé, aussi détaillé soit-il, ne constitue pas en soi un motif d'action; le titulaire des droits doit néanmoins établir que son œuvre a été utilisée, que cette utilisation a porté atteinte à ses droits exclusifs et qu'aucune exception ne s'applique.

---

138 Règlement sur l'IA, considérant 107, qui précise que l'obligation de publier un résumé doit être formulée de manière à respecter les secrets d'affaires et les informations commerciales confidentielles, sans en exiger la divulgation.

139 Dispositions relatives à la gestion des services de recommandation par algorithme dans le cadre des services d'information sur l'Internet (République populaire de Chine), en vigueur depuis le 1er mars 2022; Dispositions relatives à la gestion de la synthèse profonde dans le cadre des services d'information sur l'Internet (République populaire de Chine), en vigueur depuis le 10 janvier 2023; Mesures provisoires relatives à la gestion des services d'intelligence artificielle générative (République populaire de Chine), en vigueur depuis le 15 août 2023; Mesures relatives à l'étiquetage des contenus générés et synthétisés par l'intelligence artificielle (République populaire de Chine), en vigueur depuis le 1er septembre 2025; et la norme nationale obligatoire du même titre approuvée par l'Administration d'État chargée de la régulation du marché et l'Administration de la normalisation en 2025.

140 Framework Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trust (loi-cadre relative au développement de l'intelligence artificielle et à l'établissement des fondements de la confiance) (République de Corée), promulguée en janvier 2025 et en vigueur depuis le 22 janvier 2026.

141 Projet de loi n° 2338/2023 (Brésil), article 62 (dans la version transmise à la Chambre des députés); voir la partie II, note de bas de page n° 100 de la présente étude pour connaître l'état d'avancement législatif de ce projet de loi.

L'obligation de transparence réduit l'asymétrie d'information, mais ne la supprime pas. Deuxièmement, la conception de l'obligation de transparence est étroitement liée à celle du mécanisme de réserve. Un titulaire de droits ayant exercé son droit de réserve et dont l'œuvre est identifiée dans le résumé de transparence comme ayant été incluse dans le corpus d'entraînement peut disposer d'un fondement plus solide pour demander l'application de ses droits, sous réserve du droit applicable et des autres conditions de l'exception concernée; la question pratique est de savoir si le résumé sera suffisamment précis pour permettre cette identification.

2. Attaques visant les données, filtrage et prolifération des normes de réserve

Le deuxième défi concerne les techniques mises en œuvre par les titulaires de droits, les développeurs et les intermédiaires pour prendre en considération les aspects techniques de l'entraînement et du raisonnement qui ont une incidence sur l'application des droits. On peut distinguer quatre catégories distinctes de techniques.

La première catégorie est celle des attaques par extraction, un concept développé dans la littérature spécialisée en sécurité et désormais invoqué dans le cadre de procédures judiciaires. Les attaques par extraction sont des techniques visant à amener un modèle entraîné à reproduire le contenu de ses données d'entraînement à l'aide de commandes soigneusement élaborées. Les travaux menés dans ce domaine, passés en revue dans la première partie de cette étude, ont été cités dans la littérature sur la sécurité comme preuve que l'extractibilité peut être une propriété structurelle des grands modèles génératifs, plutôt que le simple résultat d'une stratégie particulière de formulation de commandes[142]. Du point de vue des titulaires de droits, les attaques par extraction constituent une technique de collecte de preuves : elles peuvent permettre de démontrer que certaines œuvres ont été utilisées pour l'entraînement d'un modèle ou qu'elles peuvent être extraites de celui-ci. Du point de vue des développeurs, ces mêmes attaques constituent une vulnérabilité qu'il convient d'atténuer par la déduplication des données d'entraînement, le filtrage des résultats et les techniques de "désapprentissage". La question de principe visant à déterminer si la manière dont un résultat mémorisé est obtenu devrait influer sur l'analyse de fond en matière de droit d'auteur est actuellement examinée par les tribunaux de plusieurs ressorts juridiques, comme expliqué dans la partie I, section C.2, de la présente étude.

La deuxième est la famille des filtres de sortie et d'entrée. Les développeurs ont de plus en plus souvent recours à des filtres qui empêchent le modèle de reproduire certains contenus identifiés (filtres de sortie) ou qui excluent certains contenus du corpus d'entraînement dès la phase de préparation des données (filtres d'entrée)[143]. Le rôle de ces filtres dans l'analyse de fond n'est pas encore clairement établi. Aux États-Unis d'Amérique, les arrêts Bartz et Kadrey ont abordé la question du rôle des filtres sous différents angles, le juge Chhabria ayant notamment estimé que la question de la substitution de marché dépendait, en partie, de l'efficacité des filtres à empêcher le modèle de produire des résultats susceptibles de se substituer à des œuvres originales spécifiques. Au sein de l'Union européenne, les obligations de transparence prévues par le règlement sur l'IA concernent les politiques mises en place par les développeurs en matière de droit d'auteur; la mise en œuvre de ces obligations peut avoir une incidence sur la qualification juridique des pratiques de filtrage dans le cadre de la stratégie globale de conformité d'un développeur.

---

142 Concernant les attaques par extraction, voir la partie I, note de bas de page n° 18 (Cooper, Gokaslan, Ahmed, Cyphert, De Sa, Lemley, Ho et Liang) et note de bas de page n° 22 (Carlini *et al*. (2021) et Nasr *et al*. (2025)).

143 Le filtrage au stade de l'entrée peut consister à exclure les œuvres dont les métadonnées indiquent une réservation de droits, à dédupliquer le corpus afin de réduire le risque de mémorisation et à exclure les contenus provenant de sources dont on sait qu'elles ne sont pas visées par une licence. Le filtrage au stade des résultats peut consister à détecter et à bloquer les contenus qui correspondent étroitement à des œuvres protégées identifiées. L'importance relative accordée au filtrage au niveau de l'entrée et de la sortie varie selon les développeurs et les types de modèles.

La troisième catégorie regroupe les techniques d'empoisonnement des données, par lesquelles les titulaires de droits ou des tiers introduisent des données malveillantes, erronées ou altérées dans des ensembles de données afin de compromettre les performances de tout modèle entraîné sur ces ensembles. En Chine, l'Administration du cyberespace considère l'empoisonnement des données comme une menace pour la sécurité et a lancé une opération spéciale (intitulée "Qinglang – Correction du chaos dans les applications d'IA") pour lutter contre ce phénomène; la norme nationale recommandée relative aux exigences de sécurité fondamentales pour les services d'intelligence artificielle générative impose le filtrage des données d'entraînement afin d'éliminer les informations illégales et préjudiciables[144]. Aucun traitement réglementaire comparable en matière d'empoisonnement des données n'a encore vu le jour dans les autres ressorts juridiques étudiés. L'importance pratique de l'empoisonnement des données en tant que mécanisme d'autodéfense pour les titulaires de droits n'a encore fait l'objet d'aucune évaluation faisant autorité dans aucun ressort juridique; les questions techniques et juridiques qu'il soulève restent pour l'essentiel en suspens : le recours à des techniques d'empoisonnement à l'encontre des développeurs d'IA constitue-t-il une atteinte avérée au droit détenu par le développeur, ou bien un exercice légitime des droits propres au titulaire de ces droits?

La quatrième catégorie concerne tout particulièrement l'application de la législation européenne. Il s'agit de la question des normes relatives à la représentation sous forme lisible par machine des possibilités de réserve[145]. L'arrêt rendu par la Cour d'appel de Hambourg dans l'affaire *Kneschke* c*. LAION e.V.*, évoqué dans la partie II, section B.5, établit qu'une réserve valable au titre de l'article 4.3) de la directive européenne sur le droit d'auteur et les droits voisins doit être à la fois lisible par machine et exploitable par machine. Les modalités techniques de mise en œuvre de cette exigence n'ont pas encore été définies. Des efforts visant à établir un vocabulaire commun pour les préférences d'utilisation liées à l'IA sont en cours au sein du groupe de travail sur les préférences en matière d'IA de l'Internet Engineering Task Force, et se poursuivaient encore à la mi-2026; les projets sont encore en cours d'élaboration et ne définissent qu'un petit nombre de catégories d'utilisation (essentiellement, l'entraînement de l'IA et les utilisations liées à la recherche, une catégorie consacrée à la "génération augmentée par récupération" étant en cours de discussion), et aucune norme n'a encore été adoptée[146]. Une question connexe qui mériterait d'être examinée est celle de savoir dans quelle mesure l'obligation de réserve s'accorde avec l'interdiction des formalités pour l'existence ou l'exercice des droits protégés par la Convention de Berne. La réponse peut dépendre de la configuration concrète et des effets juridiques du mécanisme de réserve.

Plusieurs normes concurrentes ont vu le jour en parallèle, et elles se répartissent en deux groupes que l'on confond facilement. Le premier ne fait qu'exprimer une réservation de droits, c'est-à-dire une clause de réserve. Il comprend le protocole de réserve concernant l'exploration de textes et de données (TDMRep) et TDM·AI, tous deux spécifiquement destinés à permettre les options de réserve conformément à l'article 4.3) de la directive européenne sur le droit d'auteur et les droits voisins; le mécanisme "content-signals" de Cloudflare, une extension exclusive du protocole d'exclusion des robots (robots.txt); ainsi que des initiatives plus larges telles que C2PA et IPTC PLUS, qui offrent des fonctionnalités de réserve selon leurs propres termes. Des signaux de ce type sont déjà utilisés dans la pratique contre les

---

144 Administration chinoise du cyberespace, opération spéciale "Qinglang – Correction du chaos dans les applications d'IA"; *Technologie de cybersécurité – Exigences de sécurité de base pour les services d'intelligence artificielle générative*, norme nationale recommandée, en vigueur à compter du 1er novembre 2025, article 4.2.1 (filtrage du contenu des données d'entraînement).

145 Concernant l'arrêt LAION II et la multiplication des normes de réserve, voir la partie II, section B.5, et les renvois au commentaire de Keller.

146 Concernant le groupe de travail sur les préférences en matière d'IA de l'Internet Engineering Task Force, voir https://datatracker.ietf.org/wg/aipref/about/. Le protocole de réserve concernant l'exploration de textes et de données (TDMRep), TDM·AI, le protocole collectif "Really Simple Licensing" (RSL), le mécanisme "contentsignals.org" de Cloudflare, C2PA et IPTC PLUS comptent parmi les principales initiatives actives dans ce domaine.

robots d'indexation des moteurs de recherche conversationnels fondés sur l'IA : une part importante des principaux sites d'actualités bloquerait, semble-t-il, le robot d'exploration utilisé par un de ces moteurs[147]. Le deuxième groupe va plus loin, en permettant à un titulaire de droits d'intégrer des conditions de licence explicites et déchiffrables par machine au sein même de l'infrastructure Web. Par conséquent, la question n'est plus seulement de savoir si une œuvre peut être utilisée, mais à quelles conditions, notamment en termes de prix et de choix entre un paiement à l'exploration et un paiement à l'inférence. La norme "Really Simple Licensing" (RSL), une initiative soutenue par les éditeurs en ligne et les réseaux de diffusion de contenu américains, en est un excellent exemple; le modèle de paiement par exploration de Cloudflare en est un autre : il intervient au niveau même de la requête d'accès en renvoyant une réponse exigeant un paiement à un robot d'exploration non autorisé. Dans ce deuxième groupe, le signal reçu par un agent automatisé sert non seulement de mécanisme de sortie, mais aussi d'offre de licence. La mise en place d'agents d'intelligence artificielle pourrait faciliter l'automatisation des procédures de concession de licences dans ce domaine.

La conformité et le règlement des litiges sont compliqués par la coexistence de ces deux types de réserves dans le cadre d'un débat toujours en cours sur les normes. Un même prestataire peut intervenir dans les deux cas, comme le fait Cloudflare par le biais de ses "content-signals" et de ses mécanismes de paiement à l'exploration, de sorte qu'il n'est pas toujours évident, à première vue, de déterminer si un signal donné correspond à une simple réserve ou à une offre de licence. La Commission européenne a lancé une consultation sur les protocoles permettant de s'opposer à l'exploration de textes et de données, conformément au règlement sur l'IA et au code de bonnes pratiques pour l'IA à usage général, en vue d'établir une liste de protocoles généralement acceptés et déchiffrables par machine que les fournisseurs de modèles d'IA à usage général doivent respecter pour se conformer à leurs obligations au titre du règlement sur l'IA[148].

La fragmentation du cadre normatif constitue en soi un obstacle à l'application des droits. L'arrêt de la Cour d'appel de Hambourg indique que la lisibilité par machine d'une réserve doit être évaluée au regard des capacités technologiques existant au moment de l'utilisation litigieuse; toutefois, dans un domaine où coexistent plusieurs normes incompatibles entre elles, la question pratique de savoir quelles normes un développeur doit respecter pour se conformer à la loi ne trouve pas de réponse simple. Le risque mis en évidence dans le commentaire sur cet arrêt est que, en l'absence de normes communes et gérées de manière transparente, les réserves lisibles par machine pourraient devenir le fondement d'une nouvelle couche intermédiaire qui normaliserait l'accès au Web ouvert selon un modèle de paiement à l'exploration ou au clic. La concrétisation de ce risque dépendra des choix qui seront faits au sein des organismes de normalisation et dans le cadre de l'élaboration des politiques au cours des prochaines années.

Le mécanisme de réserve s'inscrit dans le cadre d'une infrastructure de gestion des droits à plusieurs niveaux. La première couche est constituée d'identifiants et de métadonnées. Elle contient des identifiants conformes aux normes internationales établies pour les œuvres et les contributeurs, ainsi que des métadonnées relatives à la gestion des droits et à la provenance, lisibles par machine, qui permettent de relier une œuvre à des informations concernant son statut en matière de droits. Des travaux sont en cours pour recenser les bonnes pratiques en matière de données sur le droit d'auteur interopérables, fiables et lisibles par machine, adaptées à l'environnement de l'IA, notamment au sein de l'équipe d'experts sur l'infrastructure

---

147 Une étude publiée en 2026 a révélé que de nombreux sites d'information de premier plan aux États-Unis d'Amérique et au Royaume-Uni bloquent le robot d'indexation de Perplexity, PerplexityBot, dans leurs fichiers robots.txt. Voir BuzzStream, "Most Major News Publishers Block AI Training & Retrieval Bots" (17 décembre 2025, publié en janvier 2026) <https://www.buzzstream.com/blog/publishers-block-ai-study/>, consulté le 3 juillet 2026.

148 Commission européenne, Consultation sur les protocoles permettant de s'opposer à l'exploration de textes et de données au titre du règlement sur l'IA et du code de bonnes pratiques pour l'IA à usage général (2025); voir https://digital-strategy.ec.europa.eu.

du droit d'auteur mis en place par le Ministère finlandais de l'éducation et de la culture[149], partant du principe qu'une identification fiable est une condition préalable tant à l'expression des préférences d'utilisation qu'au suivi a posteriori des utilisations. Ce volet s'appuie sur les dispositions du WCT, du WPPT et du Traité de Beijing sur l'information sur le régime des droits, mentionnés ci-dessus.

La deuxième couche est constituée de contrôles d'accès. Elle englobe les interfaces de programmation d'applications (API) et les mesures anti-exploration grâce auxquelles les titulaires de droits et les plateformes cherchent à garantir que seuls les agents autorisés ont accès à leur contenu, ainsi que les mesures techniques de protection utilisées dans les secteurs de la diffusion en continu et de l'édition pour empêcher toute extraction non autorisée. Cette couche s'appuie sur les dispositions du WCT, du WPPT et du Traité de Beijing relatives aux technologies de neutralisation mentionnées plus haut.

La troisième couche est celle de l'attribution, c'est-à-dire l'évaluation de la contribution d'œuvres particulières à une production donnée, en vue de l'attribution d'une rémunération le cas échéant. L'attribution, dans ce sens, doit être distinguée de la détection. Les technologies de détection visent à déterminer, avec un certain degré de probabilité, si un fichier a été créé par un être humain ou par une machine; elles n'identifient pas les œuvres sources et ne permettent pas, à elles seules, d'étayer une revendication de droit d'auteur. Les technologies d'attribution, en revanche, visent à estimer la part des œuvres identifiées dans un résultat généré et, de ce fait, reposent sur un certain degré de coopération et de transparence de la part du développeur. Les investissements dans ce domaine ont déjà commencé : un fournisseur de technologies d'attribution, qui participait déjà au système suédois de licences collectives évoqué dans la partie II, a été racheté par une grande maison de disques en 2026[150]. La couche d'attribution est directement liée aux modèles de rémunération au niveau des résultats mentionnés dans la partie II, section B.6. Certains de ces modèles ne reposent pas sur une atteinte au stade des résultats, mais sur l'effet de substitution des résultats de certains modèles.

Une autre famille d'outils dédiés à la traçabilité et au tatouage numérique, associée à des initiatives visant à étiqueter les contenus créés par des humains, porte sur l'authenticité et l'origine des résultats plutôt que sur l'utilisation d'œuvres protégées dans le cadre de l'entraînement; ces outils concernent principalement les questions relatives aux résultats, que la présente étude n'aborde pas, et ne sont mentionnés ici que pour situer la couche d'attribution au sein de l'infrastructure générale.

3. La jurisprudence dans son ensemble

L'ensemble de la jurisprudence examinée permet de formuler plusieurs observations transversales concernant l'application actuelle du droit d'auteur dans le cadre de l'entraînement des IA génératives.

Premièrement, les tribunaux ont, dans l'ensemble, fait preuve d'une certaine prudence et se sont généralement abstenus de rendre des décisions visant à définir de manière exhaustive le cadre doctrinal. Les décisions rendues à ce jour ont porté sur des technologies spécifiques (la diffusion latente dans l'affaire Getty, les modèles de langage dans l'affaire GEMA, les livres acquis légalement et illégalement dans l'affaire Bartz), sur des questions de procédure particulières (la requête en irrecevabilité dans le litige opposant le New York Times, le jugement sommaire dans les affaires Bartz et Kadrey) et sur des questions spécifiques (le modèle en tant que copie dans les affaires GEMA et Getty, l'usage loyal dans les affaires Bartz et Kadrey, notamment). Elles n'ont pas encore permis de mettre en place de cadre complet dans aucun

---

149 L'équipe d'experts sur l'infrastructure du droit d'auteur (CITF), créé sous l'égide du Ministère finlandais de l'éducation et de la culture, a publié en 2025 un premier rapport intitulé "Interoperable, trustworthy and machine-readable copyright data in the AI era".

150 Warner Music Group a racheté Sureel, fournisseur de technologies d'attribution, en juin 2026. La technologie de ce même fournisseur est utilisée dans le cadre du contrat de licence conclu avec la société STIM, évoqué dans la partie II, section B.3.

système juridique. La position allemande, bien qu'elle se reflète désormais dans trois décisions, n'est pas homogène en soi : la ligne LAION admet l'application à l'entraînement de l'exception relative à l'exploration de textes et de données, tandis que la ligne de l'affaire GEMA suit une approche différente. La position des États-Unis d'Amérique présente une structure similaire et est en attente d'un réexamen en appel, puis, à terme, d'un ou plusieurs arrêts de la Cour suprême des États-Unis d'Amérique.

Deuxièmement, la jurisprudence a néanmoins permis de dégager certains points de convergence. Il existe un consensus sur l'exigence, inscrite dans la Convention de Berne, selon laquelle un auteur humain est nécessaire pour conférer l'originalité requise à la protection des résultats générés par l'IA. On peut observer ce phénomène aux États-Unis d'Amérique, au Kenya[151], en Allemagne au niveau des juridictions de première instance, ainsi qu'au sein de l'Union européenne, où le critère d'originalité réaffirmé par la CJUE dans l'affaire Mio, qui portait sur la protection de meubles au titre du droit d'auteur, présuppose l'existence d'un auteur humain[152]. On observe également une convergence en matière d'application du droit de reproduction dans le cadre de la création d'ensembles de données à des fins d'entraînement (en Allemagne, en Chine et au Canada). On constate enfin une certaine convergence quant à la pertinence de la substitution de la production dans le cadre de l'analyse (aux États-Unis d'Amérique et, à l'avenir, dans l'Union européenne grâce à l'application du "test en trois étapes"). Parmi les points sur lesquels la jurisprudence reste divisée figurent notamment la question de savoir si les modèles entraînés contiennent des copies des œuvres utilisées pour l'entraînement, le réentraînement et le réglage fin, celle de savoir si et dans quelle mesure l'exception relative à l'exploration de textes et de données s'étend à l'entraînement de l'IA, ainsi que celle de la portée de la compétence transfrontière.

La troisième observation est que les défis pratiques identifiés dans les sections C.1 et C.2 ci-dessus, à savoir la transparence, la prolifération des normes de réserve, ainsi que les coûts et les asymétries d'information entre les développeurs et les titulaires de droits, sont susceptibles d'influencer l'évolution de la jurisprudence au cours des mois et des années à venir. Même lorsque les tribunaux tranchent les questions de fond relatives au droit d'auteur, la possibilité concrète pour les titulaires de droits de faire valoir leurs revendications dépendra des mécanismes leur permettant de vérifier l'utilisation de leurs œuvres, d'exprimer des réserves et de présenter des preuves sous une forme permettant aux tribunaux de statuer. Dans ce domaine, la conception technique et l'analyse juridique sont étroitement liées, et la mise en œuvre dépendra des choix effectués dans ces deux domaines.

---

151 *Aryeh Movement Limited* c. *Cynthia Beldina Akoth Okello*, COPTA/E001/2025 (Tribunal du droit d'auteur du Kenya) (statuant que l'article 22.3) de la loi sur le droit d'auteur (chap. 130) présuppose une intervention humaine dans le déploiement d'un "effort suffisant", de sorte qu'une œuvre générée de manière autonome par un système d'intelligence artificielle ne peut satisfaire au seuil légal à moins qu'une intervention humaine suffisante ne confère à l'œuvre son originalité).

152 *Mio et autres*, C-580/23, jointe à l'affaire *USM* c. *Konektra*, C-795/23 (CJUE, arrêt du 4 décembre 2025). Les affaires jointes portaient sur la protection au titre du droit d'auteur des œuvres d'art appliqué, et plus particulièrement des meubles (dans l'affaire C-580/23, les tables de salle à manger "Palais Royal"). La Cour a réaffirmé que l'objet ne peut être considéré comme une œuvre protégée que s'il présente un caractère original, en ce sens qu'il doit constituer une création intellectuelle propre à son auteur, en manifestant les choix libres et créatifs qui expriment la personnalité de celui-ci, et que les œuvres d'art appliqué sont soumises à la même exigence d'originalité que les autres œuvres. L'arrêt ne portait pas sur l'intelligence artificielle; l'exigence d'un auteur humain pour la protection des résultats issus de l'IA découle de ce critère d'originalité plutôt que d'une quelconque décision de la Cour portant spécifiquement sur l'IA. Les auteurs pourraient en outre se heurter à un nouvel obstacle dans le cadre des procédures d'application, à savoir celui de devoir prouver que leur œuvre n'a pas été créée par une intelligence artificielle, ou du moins qu'ils y ont apporté une contribution suffisante pour justifier la protection au titre du droit d'auteur.

CONCLUSION

Cette étude met en évidence une évolution rapide et inégale dans deux domaines : d'une part, les ressorts juridiques qui expérimentent différentes combinaisons d'exceptions, de mécanismes de concession de licences et de systèmes de rémunération; d'autre part, une jurisprudence qui a permis de clarifier ces questions mais qui n'a pas encore abouti à un cadre stable. Les États membres qui envisagent aujourd'hui des mesures législatives ou réglementaires le font dans un contexte plus évolué qu'il y a encore 12 mois; les points sur lesquels l'expérience et la jurisprudence convergent constituent une base partielle, tandis que ceux sur lesquels elles continuent de diverger nécessiteront des choix politiques que la présente étude s'est efforcée d'exposer, sans toutefois les trancher. Ces choix continueront d'être pris au niveau national et, lorsque les États membres le décideront, feront l'objet de discussions au niveau international au sein du Comité permanent du droit d'auteur et des droits connexes et dans le cadre plus large de l'OMPI.

L'étude montre que l'équilibre entre les intérêts des titulaires de droits et ceux des développeurs d'IA peut être établi différemment selon les États membres. Cinq questions plus précises se dégagent de cette analyse.

La première question est de savoir si les reproductions effectuées dans le cadre de la constitution d'un corpus et de l'entraînement d'un modèle doivent faire l'objet d'une autorisation ou si elles relèvent d'une exception ou limitation, qu'il s'agisse d'une disposition spécifique à l'IA, d'une disposition relative à l'exploration de textes et de données ou d'une clause générale, et à quelles conditions, notamment l'accès licite, la finalité de l'entraînement et le caractère commercial ou non commercial de l'utilisation.

La deuxième question concerne la conception de tout mécanisme de réservation dans lequel ils sont utilisés. Cela inclut la forme de la réserve, en tenant compte du fait que les réserves lisibles par machine supposent des compétences que les auteurs et artistes interprètes ou exécutants ne possèdent pas nécessairement. À l'inverse, un régime de réservation qui ne peut pas être traité automatiquement à grande échelle ne laisse aux développeurs aucune voie claire pour se conformer à la réglementation. Il convient également d'examiner la compatibilité d'un régime de réserve avec l'article 5.2) de la Convention de Berne.

La troisième question est de savoir si la rémunération doit être liée à l'utilisation, à la substitution de résultats, ou aux deux; si elle doit prendre la forme d'une licence commerciale fondée sur des droits exclusifs ou d'un droit à rémunération prévu par la loi; et, dans les deux cas, par l'intermédiaire de quelles institutions elle doit être versée. Il s'agit notamment de déterminer si les mandats de gestion collective s'étendent aux utilisations à des fins d'entraînement, si des formes élargies ou nouvelles de licences collectives sont nécessaires, et comment concevoir les chaînes contractuelles de manière à ce que la valeur revienne aux auteurs et aux artistes interprètes ou exécutants dont les œuvres sont utilisées.

La quatrième question porte sur la transparence : quels sont les obligations en matière de conservation des documents et de communication des supports d'entraînement nécessaires pour permettre l'exercice des droits et la gestion des rémunérations, et comment concilier ces obligations avec la protection légitime des secrets d'affaires et les préoccupations en matière de concurrence?

La cinquième et dernière question est de savoir si les mesures de correction restent efficaces et proportionnées lorsque la suppression d'œuvres d'un modèle entraîné est techniquement impossible, et comment la compétence doit être exercée lorsque l'entraînement a lieu à l'étranger mais que les services qui en résultent sont proposés sur le territoire national.

[Fin du document]