



TDM et IA : un aperçu du cadre juridique applicable

Lionel Maurel

Ingénieur de recherche

Direction des Données Ouverte de la Recherche (DDOR)

CNRS





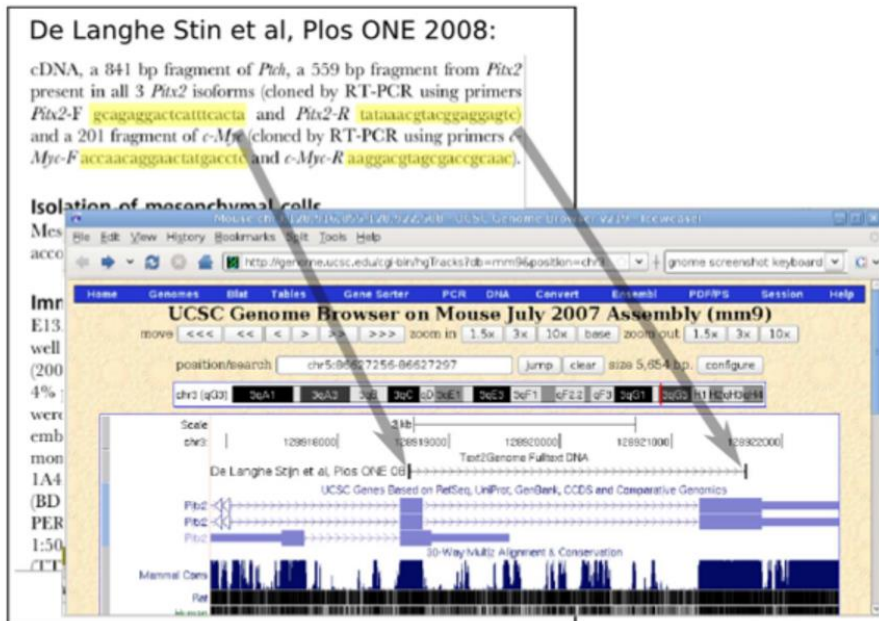
TDM et IA : un aperçu du cadre juridique applicable

Plan Général

- Les enjeux juridiques généraux du TDM ;
- La pratique du TDM à des fins de recherche ;
- Les enjeux spécifiques soulevés par l'intelligence artificielle ;
- Un aperçu de l'évolution du cadre juridique en dehors des activités de recherche.

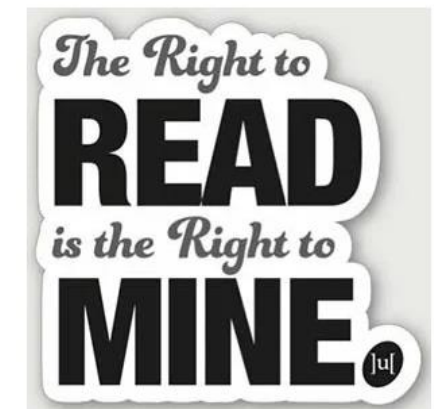
Les enjeux juridiques généraux autour du TDM

Définition : « On entend par fouille de textes et de données [...] la mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations. » [Directive 2019/790](#)



[Projet Text2Genome](#)

- Une pratique susceptible de se heurter aux **droits exclusifs de propriété intellectuelle** (reproduction, représentation)
- Impliquant donc la **négociation complexe d'autorisations** auprès des titulaires de droits (droit d'auteur, droit des producteurs de bases de données, etc).
- D'où la réflexion dans les années 2010 sur la mise en place d'une **exception en faveur du TDM**



La longue marche vers une exception TDM

- Dans les années 2010, **plusieurs pays pionniers** font évoluer leur législation pour introduire une exception TDM à des fins de recherche (Angleterre, Allemagne, Japon)
- Dans **d'autres systèmes juridiques plus souples**, les tribunaux avalisent les pratiques de TDM (Chine, Etats-unis) sur la base du « fair use » (usage équitable).



La République numérique

en actes

Projet de loi pour une République numérique

Gouvernement, le 26 septembre 2015

8221	1275	855	20
votes	contributions	participants	jours restants

Participer Partager

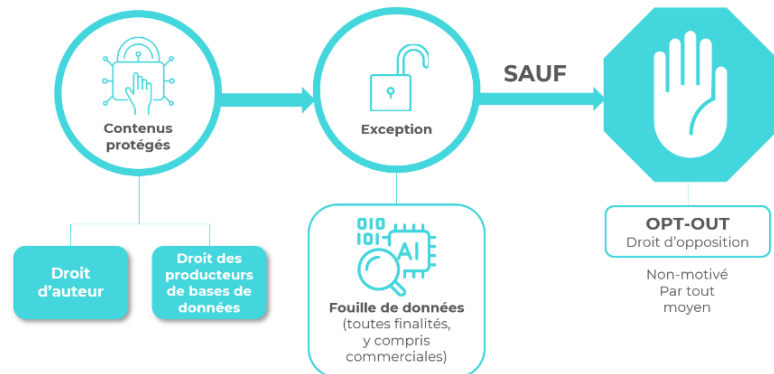
En France, la loi pour une République numérique établit en 2016 une exception TDM à des fins de recherche après de longs débats.

Mais les décrets d'application du texte ne paraîtront jamais...

Directive Copyright (2019) :

Deux exceptions en faveur du TDM

- Une exception pour la fouille de textes et de données **à des fins de recherche publique** ;
- Une **exception plus large** applicable au TDM pour toutes autres finalités, à condition que les titulaires de droits ne manifestent pas leur opposition (mécanisme **d'opt out**).



Une logique (inédite pour une exception),
fortement critiquée par les industries
culturelles et créatives...



Transposée en France par voie
d'ordonnance en 2021 (et pas besoin
cette fois de décret d'application).

Un texte qui ne contient pas le mot « intelligence
artificielle », mais qui a de nombreux liens avec elle.

L'exception TDM à des fins de recherche publique



https://doranum.fr/aspects-juridiques-ethiques/text-data-mining-tdm_10_13143_ywkr-5w34/

Une exception largement applicable :

- A **tous les objets protégeables par la propriété intellectuelle** (textes, images, vidéos, musiques, logiciels, bases de données, etc.) ;
- Du moment qu'on bénéficie d'**un accès licite** à ces contenus (par voie d'acquisition/abonnement ou en ligne sur Internet) ;
- Egalement ouverte aux **institutions culturelles** (bibliothèques, archives, musées) ;
- L'exception est **d'ordre public** (les clauses contractuelles contraires sont sans effet) ;
- L'exception est **gratuite**, sans compensation pour les ayants droit.

L'exception TDM à des fins de recherche publique

Des limites et des conditions à respecter :

- Finalité de recherche publique ;
- Accès licite seulement ;
- Respect de l'intégrité et de la sécurité des systèmes ;
- Cadre non-lucratif (mais partenariats public-privé de recherche ne sont pas exclus) ;
- Conservation des fichiers source possible (mais à condition qu'elle soit sécurisée) ;
- Réutilisation possible (par d'autres chercheurs), mais encadrée (pas d'exposition des fichiers source).



La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français

<https://www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche-une-pratique-confirmee-et-desormais-operationnelle-en-droit-francais/>



Mais débats continus sur l'applicabilité de ces exceptions TDM à l'intelligence artificielle ([rapport CSPLA 2025](#)).
+ CJUE saisie du sujet par un tribunal hongrois en 2025.

Principale limite : l'exception TDM couvre les questions de propriété intellectuelle, mais pas la protection des données personnelles

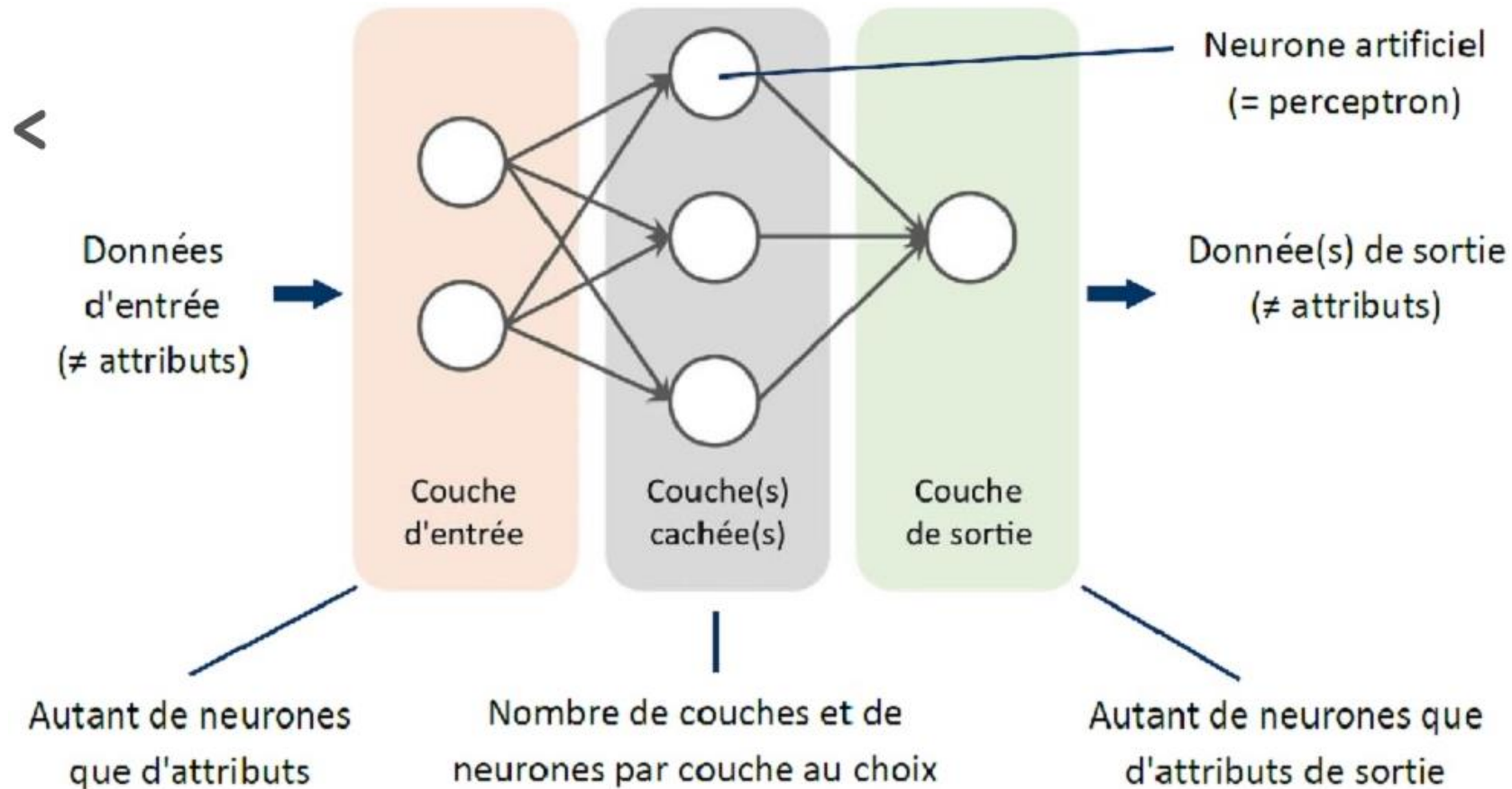


« Données à caractère personnel » :

toute information concernant une personne physique identifiée ou identifiable (personne concernée); est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale.

[RGPD \(2018\)](#)

Enjeux juridiques spécifiques à l'intelligence artificielle



Données
d'entrée/données
de sortie

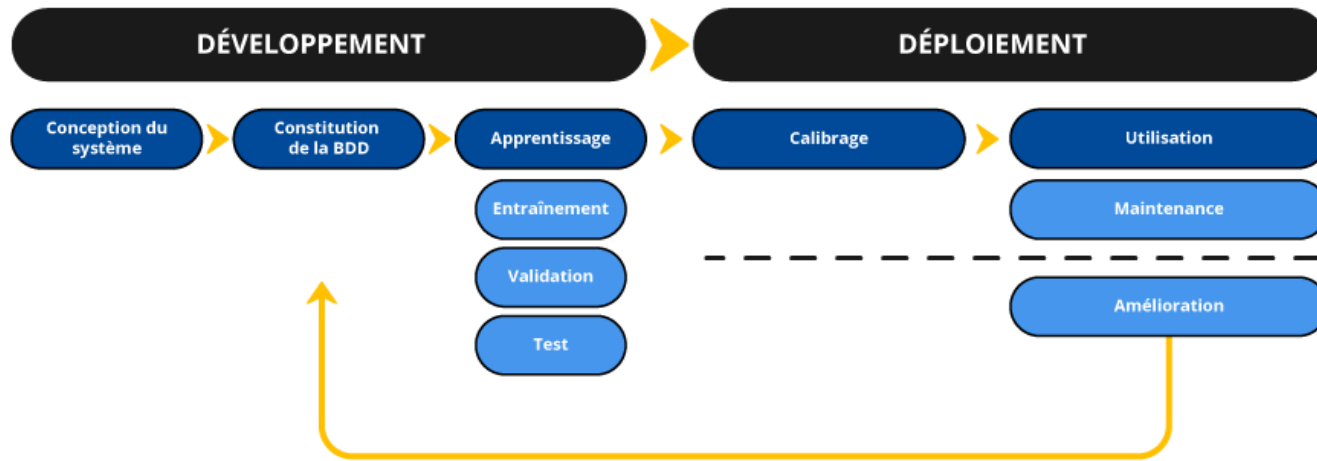
Phase de
développement /
Phase de
déploiement

Questions
spécifiques à
l'intelligence
artificielle
généraliste

« système d'IA » : un système basé sur une machine qui est conçu pour fonctionner avec différents niveaux d'autonomie et qui peut faire preuve d'adaptabilité après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer des environnements physiques ou virtuels. ([IA Act 2024](#))

Intelligence artificielle et protection des données personnelles

[CNIL : fiches pratiques
sur l'IA \(2024\).](#)



Notion centrale de
traitement de données
à caractère personnel.

- **Il est certain qu'aucune donnée personnelle n'est présente dans la base de données** : les fiches ne s'appliquent pas (même si certaines recommandations, de l'ordre de la bonne pratique, peuvent être pertinentes).
- **Il est certain que des données personnelles sont présentes** : les fiches s'appliquent. C'est le cas des systèmes d'IA développés à partir de vidéos ou d'images de personnes, d'enregistrements de voix, de données personnelles structurées, etc. À noter que les textes européens posent la règle selon laquelle les jeux de données comprenant des données personnelles et non personnelles, dits « mixtes », sont régis par le RGPD, si les deux types de données sont inextricablement liés.
- **Il est possible que des données personnelles soient présentes** : c'est un cas fréquent pour lequel la collecte de données personnelles n'est pas expressément souhaitée. Par exemple :
 - présence résiduelle de personnes ou de plaques d'immatriculation dans des images ;
 - occurrences de noms, prénoms, adresses, etc. dans des données textuelles de types commentaires ou prompt, etc.

Si des traitements de données personnelles sont opérés...

- Un (ou des) **responsables de traitement** doivent être identifiés ;
- Une **finalité de traitement** doit être définie ;
- De **grands principes** doivent être respectés (sécurisation, nécessité, minimisation, durée de conservation limitée, etc.) ;
- Une **base légale** doit être invoquée (qui n'est pas forcément le consentement des personnes : mission d'intérêt public, intérêt légitime du responsable de traitement, etc.) ;
- Les **droits des personnes** sur leurs données doivent être respectés (information, accès, rectification, opposition, etc.) ;
- Des **formalités** doivent être accomplis (accountability : inscription au registre des traitements de l'établissement).
- Une vigilance particulière sur les **transferts de données personnelles** hors de l'Union européenne.



Néanmoins les activités de recherche bénéficient de dérogations importantes dans le RGPD

- Dérogations possibles à l'exercice du **droit des personnes** (information, opposition) ;
- Dérogation sur la **durée de conservation** des données ;
- Dérogation concernant la **précision de la finalité** de traitement pour recherches exploratoires ;
- Possibilité de traitement des **données sensibles** ;
- Possibilité de s'appuyer sur **la mission d'intérêt public** comme base légale du traitement.

« La mission d'intérêt public peut être mobilisée, de manière générale, pour les traitements de données effectués par des laboratoires de recherche publics, dont les traitements de données sont nécessaires pour leur activité de recherche. » CNIL

The screenshot shows the CNIL website interface. At the top, the CNIL logo is displayed alongside its mission statement: 'PROTÉGER les données personnelles, ACCOMPAGNER l'innovation, PRÉSERVER les libertés individuelles'. Below this is a navigation bar with links for 'MA CONFORMITÉ AU RGPD', 'RESSOURCES', 'TEXTES OFFICIELS', and 'LA CNIL', along with a search icon. The main heading of the page is 'Recherche scientifique (hors santé)'. A paragraph of text states: 'La recherche scientifique doit s'accompagner de garanties fortes pour protéger les droits des personnes et respecter le RGPD et la loi Informatique et Libertés. Les contenus de cette page ne s'appliquent qu'en dehors du champ de la santé, pour lequel des règles spécifiques s'appliquent.' Below this, there are two columns: 'Les questions-réponses' and 'Les bases légales'. The 'Les questions-réponses' column contains the text: 'La recherche scientifique peut reposer sur des traitements de données personnelles. La CNIL rappelle les règles applicables pour protéger les droits et libertés des personnes concernées.' The 'Les bases légales' column contains the text: 'La CNIL détaille et précise les bases légales mobilisables pour la collecte et l'utilisation de données personnelles pour la recherche scientifique.'

<https://www.cnil.fr/fr/thematiques/recherche-scientifique-hors-sante>

Attention à la question de l'anonymisation

I. Do AI Models Fall Under Data Protection Laws?

In the view of the EDPB, the GDPR regularly applies to AI models trained with personal data.

Different types of AI models are distinguished as follows:

Model Type	Description	GDPR Applicability
Explicit Data-Providing Models	These models are specifically designed to provide or make available personal data about individuals whose data was used for training. Examples include: - Generative models fine-tuned with the personal data of an individual, such as voice recordings to mimic the individual's voice. - Models that reply to prompts with personal data, e.g., "Who is the president of Poland?"	Always applicable. The EDPB clearly states that these models inherently process personal data, so they cannot be considered anonymous.
Implicit Data-Embedding Models	These AI models are not intentionally designed to produce identifiable personal information. However, personal data from the training set can still be embedded in the model's parameters, namely represented through mathematical objects. Parameters in a model may reflect statistical relationships from the training data. This means it's possible to extract personal data—either accurately or inaccurately—by analyzing these relationships or querying the model. Information about training data might be extractable through targeted prompts. - For instance, by exploiting vulnerabilities in AI models, one can perform a membership inference attack to determine whether specific data was included in the training set. - Additionally, model inversion attacks can be used to reconstruct parts of the original training data. - At least, personal data can be obtained accidentally through interactions with an AI model, for example as part of an AI system.	Subject to case-by-case analysis. AI models trained on personal data can't always be considered anonymous. Instead, anonymity should be assessed on a case-by-case basis using specific criteria.

« L'**anonymisation** consiste à altérer la donnée de manière irréversible pour rendre impossible l'identification d'une personne.

En revanche, la **pseudonymisation** consiste à remplacer les données directement identifiantes (nom, prénom) par des données indirectement identifiantes (alias). Il s'agit d'un remplacement des données qui n'efface pas le caractère personnel des informations. De plus, ce processus est réversible. »

([Leto Legal](#))

Position complexe de l'EDPB (European Data Protection Board) sur les modèles d'IA « anonymes » ([décembre 2024](#))



[Fiche CNIL sur le statut d'un modèle d'IA au regard du RGPD](#)

Vigilance dans l'usage des outils Cloud d'intelligence artificielle...

1. Sécurité du système d'information

Nombre d'informations et de données ne doivent en aucun cas être transmises à des IAs hébergées hors de l'université¹ pour lesquelles la **confidentialité** ne peut être systématiquement garantie. Il ne faut jamais leur transmettre de données personnelles, ni de documents professionnels et a fortiori de documents confidentiels.

Seules des données relevant du **domaine public**, ou dont l'utilisateur est propriétaire, peuvent être communiquées.

2. Protection des données à caractère personnel

Le RGPD (Règlement Général pour la Protection des Données) s'applique aux traitements de données à caractère personnel mobilisant des outils d'IA.

À ce titre, il convient de prendre garde au dépôt de documents contenant des données personnelles, à la saisie directe de ces données menant à la divulgation ou la révélation d'informations préjudiciables à l'encontre de soi-même ou d'autres personnes et de se prémunir d'une éventuelle fuite de données personnelles dans ces outils. Une vigilance particulière doit être portée vis-à-vis de ceux qui ne respecteraient pas la réglementation (outils non proposés par l'université ou hors-Europe ou commerciaux ou non libres).

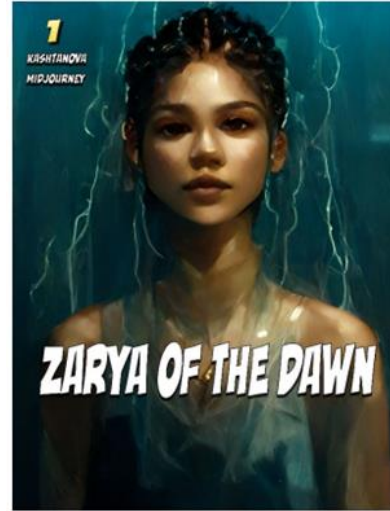
[Unistra. Charte des bons usages des moyens numériques de l'université \(2025\).](#)

IA et propriété intellectuelle

Quel statut pour les créations produites par (ou avec) l'intelligence artificielle génératives ?

Quel statut pour les prompts ?

- US Copyright Office, 21 février 2023, Kristina Kashtanova, « Zarya of the Dawn » ; ❌
- United States District Court, District of Columbia, Stephen Taler v. Shira Perlmuter, 18 août 2023 ; ❌
- US Copyright Review Board, Jason Allen, « Théâtre d'Opéra Spatial », 5 septembre 2023 ; ❌
- Tribunal municipal de Prague, 11 octobre 2023 ; ❌
- Beijing Internet Court, Li v. Liu, 27 novembre 2023 ; ✅
- US Copyright Office, 30 janvier 2025, A single Piece of American Cheese. ✅



Pour l'instant, les grands principes du droit d'auteur sont bousculés, mais globalement ils résistent...

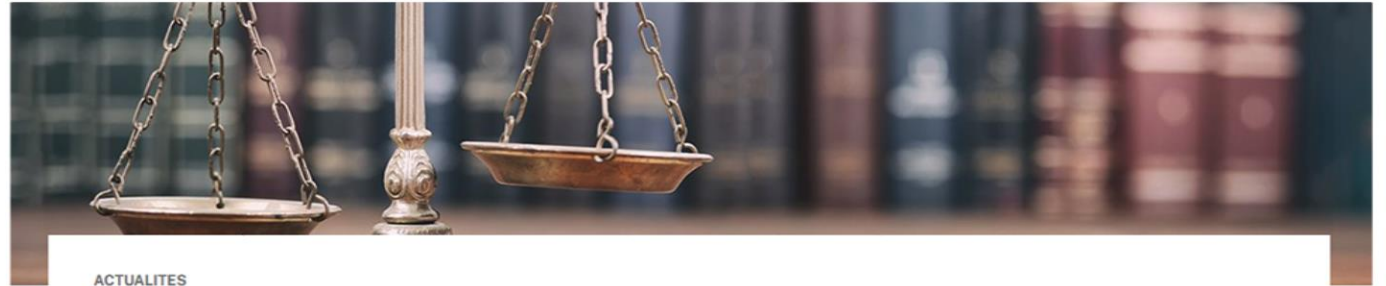
Les mêmes questions se posent pour les logiciels, les inventions brevetables, etc.

IA et propriété intellectuelle

De nombreux procès intentés dans le monde par des titulaires de droits contre les grands acteurs de l'IA.

Soit pour des usages d'œuvres protégées comme données d'entrée, soit pour des contrefaçons d'œuvres protégées dans les données de sortie.

Les premières décisions de justice US commencent à tomber...



ACTUALITES

IA et droits d'auteur : victoires fragiles pour Meta et Anthropic dans la guerre du « fair use »

Bien que les juges aient accepté l'argument de l'usage loyal de la part des fournisseurs dans leur affaire respective qui les confrontaient à des auteurs, cela ne signifie pas que d'autres procès auront la même décision. Il est également probable que la Cour suprême sera l'ultime instance de décision.

<https://www.lemagit.fr/actualites/366626854/IA-et-droits-dauteur-victoires-fragiles-pour-Meta-et-Anthropic-dans-la-guerre-du-fair-use>

En cas de confirmation de l'application du Fair Use aux IA génératives, quel impact sur l'Opt Out de la directive européenne ?



IA et redéfinition de l'Open Source ?

Proposition d'une Définition de l'IA Open Source
([OSAID](#)) par l'Open Source Initiative (2024)

What is Open Source AI

When we refer to a "system," we are speaking both broadly about a fully functional structure and its discrete structural elements. To be considered Open Source, the requirements are the same, whether applied to a system, a model, weights and parameters, or other structural elements.

An *Open Source AI* is an AI system made available under terms and in a way that grant the freedoms¹ to:

- Use the system for any purpose and without having to ask for permission.
- Study how the system works and inspect its components.
- Modify the system for any purpose, including to change its output.
- Share the system for others to use with or without modifications, for any purpose.

These freedoms apply both to a fully functional system and to discrete elements of a system. A precondition to exercising these freedoms is to have access to the preferred form to make modifications to the system.

Liberté d'utilisation : Utiliser le système d'IA pour n'importe quelle finalité, sans autorisation préalable.

Liberté d'inspection : Étudier le fonctionnement interne du système.

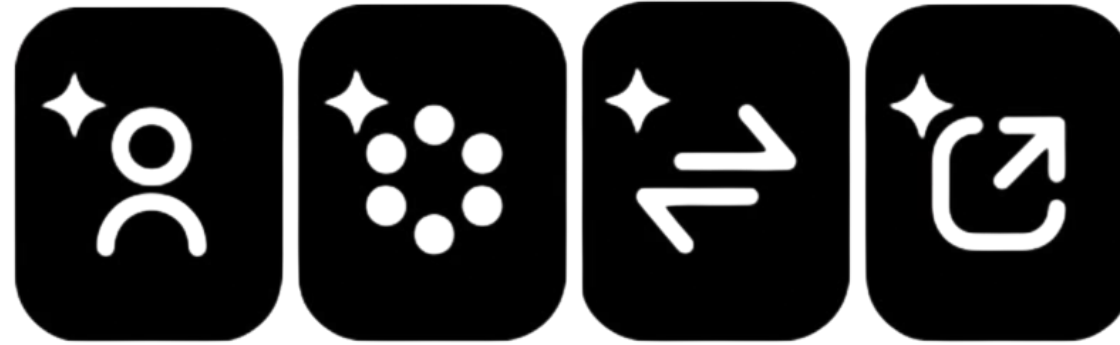
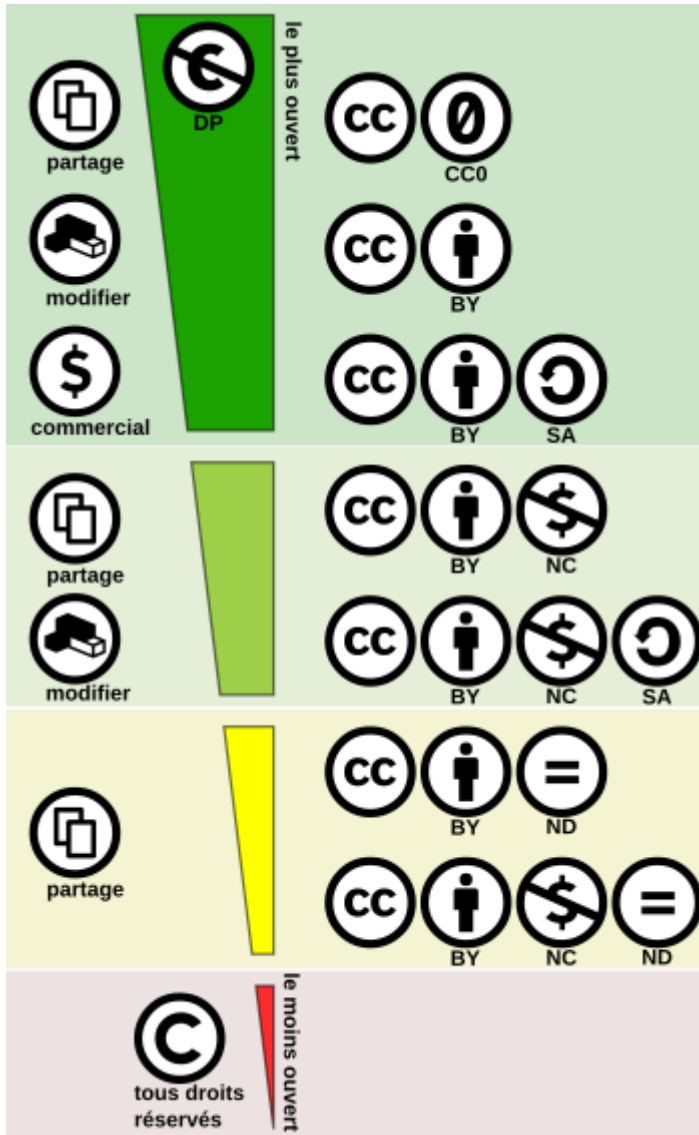
Liberté de modification : Modifier le système pour adapter ses réponses ou ses capacités à d'autres fins.

Liberté de partage : Distribuer le système ou ses versions modifiées à d'autres utilisateurs, à des fins diverses.

« La définition Open Source de l'IA demande ainsi un accès au code source complet du modèle (entraînement et exécution), ainsi qu'aux paramètres et poids du modèle et aux informations sur les données. Néanmoins, les données qui ont permis l'entraînement des modèles n'ont pas à être accessibles, ce qui a mené à de nombreux débats. » [Célya Gruson](#).

Différence entre Open Source et Open Weight
(seuls les paramètres appris durant l'entraînement sont rendus disponibles).

IA et évolution des Creative Commons (CC Signals 2025) ?



 **Credit:** You must give appropriate credit based on the method, means, and context of your use.

 **Direct Contribution:** You must provide monetary or in-kind support to the Declaring Party for their development and maintenance of the assets, based on a good faith valuation taking into account your use of the assets and your financial means.

 **Ecosystem Contribution:** You must provide monetary or in-kind support back to the ecosystem from which you are benefiting, based on a good faith valuation taking into account your use of the assets and your financial means.

 **Open:** The AI system used must be open. For example, AI systems must satisfy the Model Openness Framework (MOF) Class II, MOF Class I, or the Open Source AI Definition (OSAID).

<https://creativecommons.org/ai-and-the-commons/cc-signals/implementation/>

“La réciprocité à l’ère de l’intelligence artificielle consiste à favoriser une relation mutuellement bénéfique entre les créateurs des données et les concepteurs de modèles d’IA. Pour ces derniers, qui tirent un bénéfice disproportionné des biens communs, la réciprocité est une manière de rendre à ces communs, d’une façon adaptée à chaque communauté et à chaque contexte. »

Creative Commons

D'autres évolutions du cadre juridique européen



2022



2023



2023



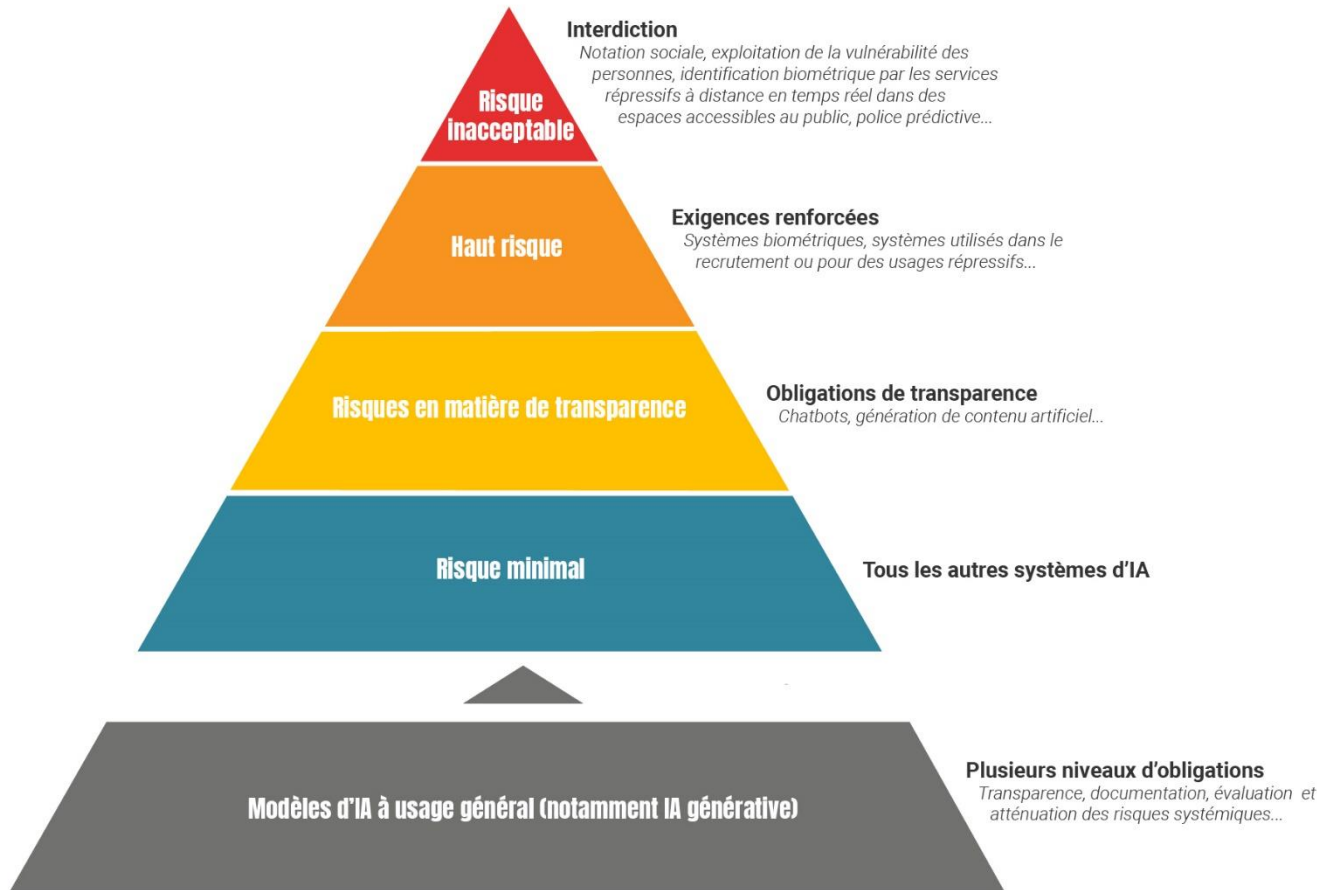
2024



2024

Quel apport de l'EU AI Act (2024) ?

Règlement IA - Approche par les risques



- Le texte vise à encadrer **le développement, la mise sur le marché et l'utilisation** de systèmes d'intelligence artificielle (IA), qui peuvent poser des risques pour la santé, la sécurité ou les droits fondamentaux.
- En fonction des niveaux de risques**, il prévoit pour les fournisseurs des obligations de transparence, de documentation, d'atténuation des risques systémiques, etc.
- Les fournisseurs doivent mettre **politique de conformité avec le droit d'auteur** en place une (respect de l'Opt-Out et transparence sur les données d'entraînement).

A noter : exclusion des activités de recherche et développement du champ de l'AI Act.

« Le présent règlement ne s'applique pas aux activités de recherche, d'essai ou de développement concernant les systèmes ou modèles d'IA avant leur mise sur le marché ou leur mise en service. » ([Article 2.8](#))