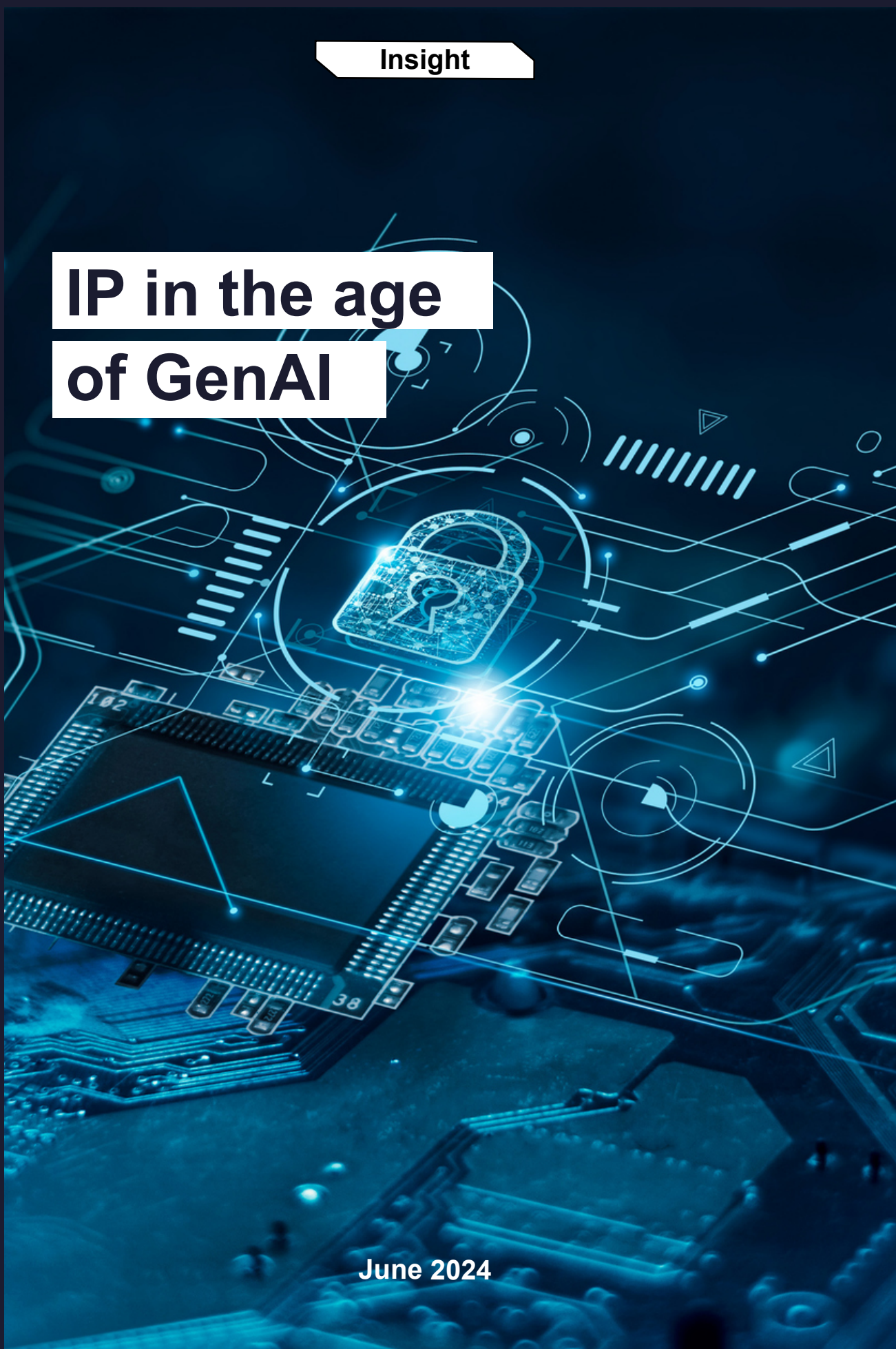
Insight

IP in the age of GenAI



June 2024

Executive Summary

1. **Effective IP rights** are a guarantee of the protection of creative works and a cornerstone of our creation model.

2. The emergence of generative IA raises the question of the **use of IP-protected content for AI training purposes**.

3. The exception to copyright for text and data mining created by the CDSM directive allows such use of protected content, while granted an **opt-out** right to rights holders.

4. Debate rose on the effectiveness of this opt-out right: the new AI Act strengthens this mechanism by creating **new transparency requirements** on content used for AI training purposes. Europe can now deal with these technological developments with a robust regulatory framework.

5. Many of today's concerns about the opt-out right will in fact be more resolved by technical and standardization works rather than by regulatory leverage. France and Europe must be pioneers in **drawing up the international standards** that will become tomorrow's norms.

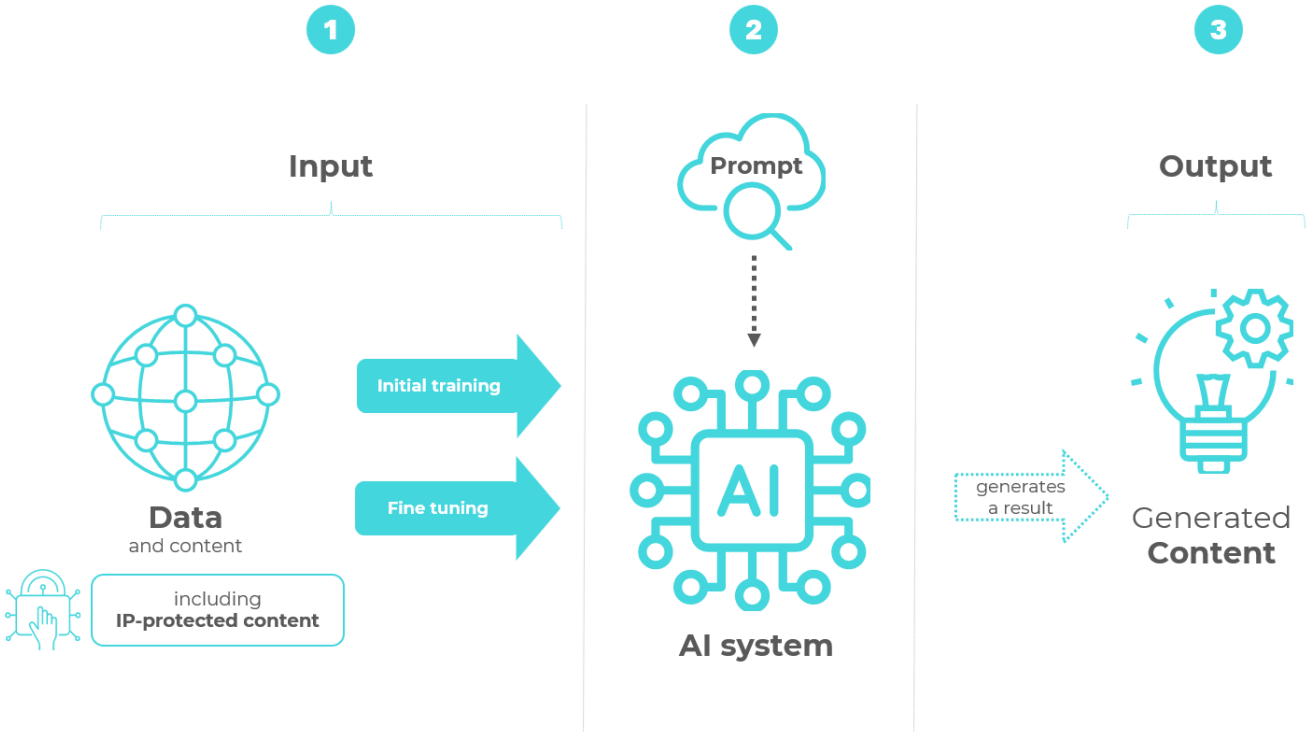
6. Beyond regulation, the question of the volume of data available in Europe for training AI models is central: it has important consequences, both for the **ability of AI startups to develop new models in Europe**, and for the **linguistic and cultural diversity of AI**.

Introduction

With the development of generative AI and foundation models trained thanks to self-supervised learning and large volumes of data, the interaction between AI and IP is at the center of discussions. The use of IP-protected content (text, images, sound, etc.) for AI training purposes raises the question of the adequacy of the protection currently granted to rights holders.

Even if today's discussions focus on legal issues, the stakes are high, both for our capacity to train AI-models and for the robustness of our cultural model. It is therefore necessary to both preserve creation and foster the development of competitive foundation models on European soil.

AI et IP-protected content: how do they interact?



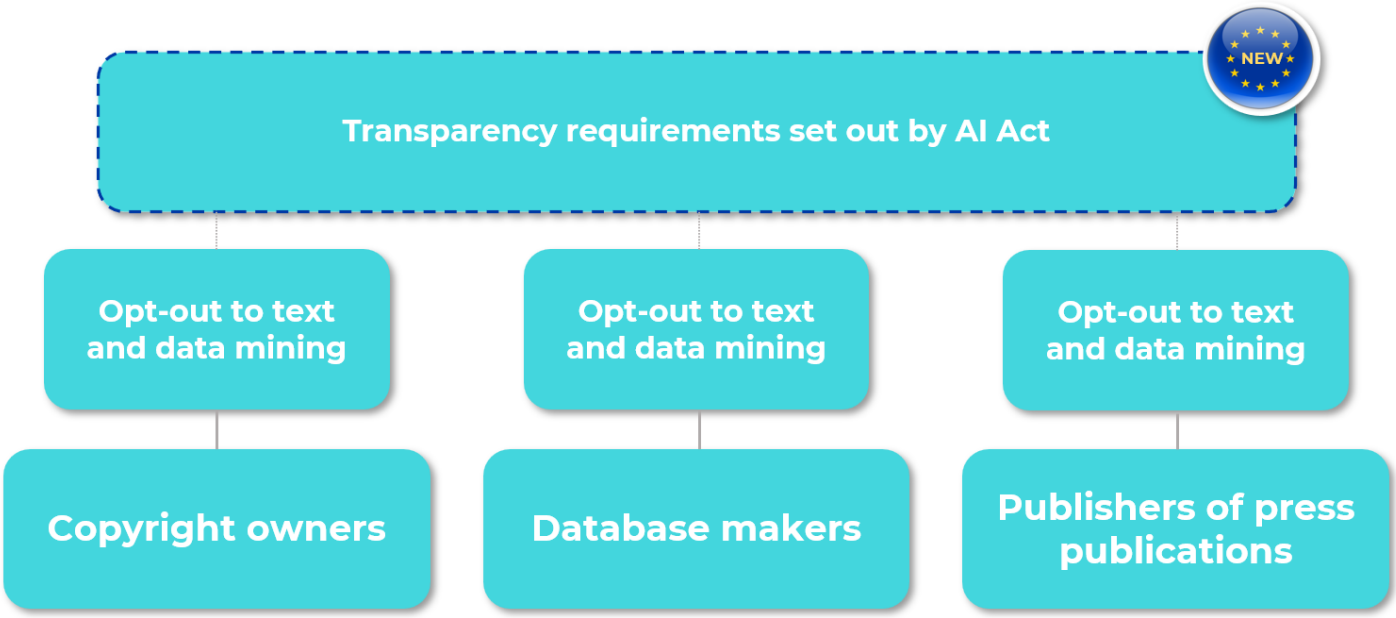
To be efficient, a generative AI model needs to be trained on large sets of pre-existing data and content. This data and content may be protected by intellectual property rights, such as **copyright** or ***sui generis* right to databases**.

The **input** can be used through different technical processes such as data scraping or text and data mining and at different stages, from initial model training to later phases, such as fine-tuning (optimizing a model by re-training it on specific data to adjust its parameters and performance).

When a prompt is entered by the user, the AI model generates an output from the learning previously performed.

While the distinction between *input* and *output* has theoretical advantages, it is in fact not totally hermetic: for example, the information contained in the prompt itself could in certain ways be also viewed as input.

What protection is there for IP right holders?



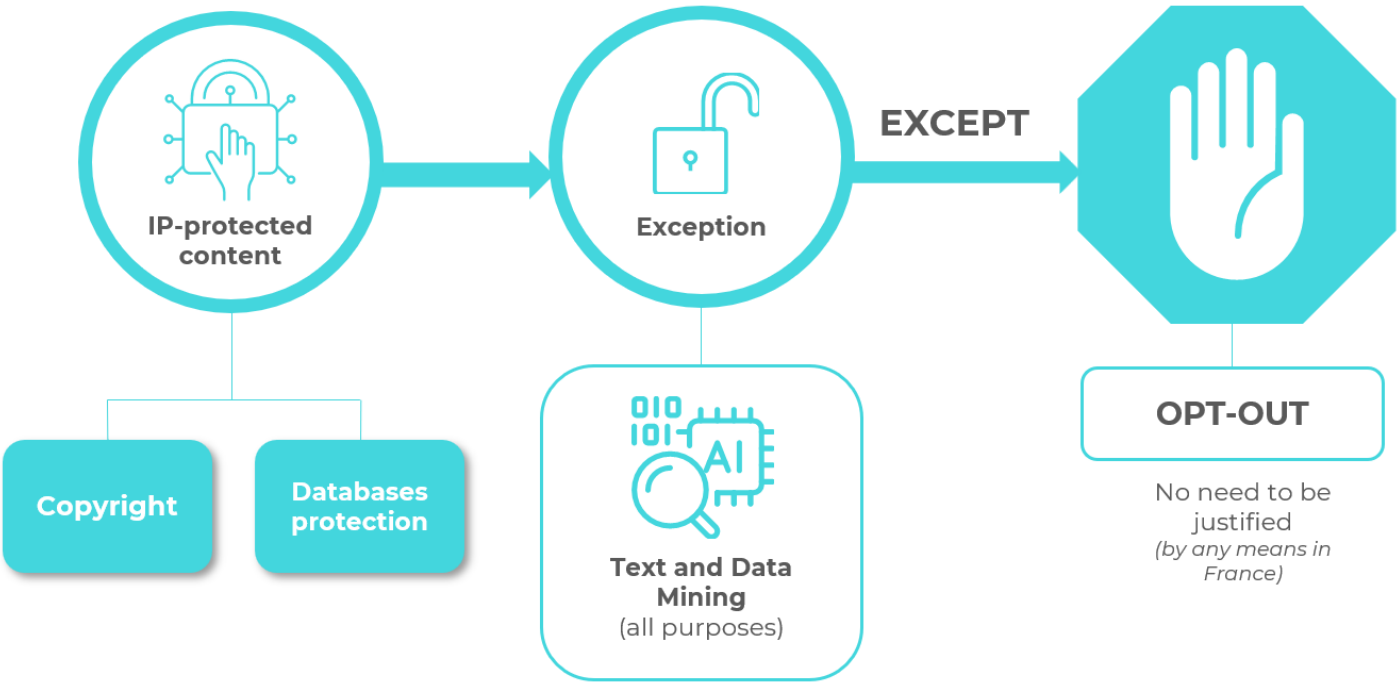
European law provides several mechanisms that enable IP rights holders to protect their content. This legal framework was deeply revised in 2019 by the directive on copyright and related rights in the Digital Single Market (the "**CDSM directive**") and adapted to technological developments, notably with the creation of the "**text and data mining**" characterization. This characterization is defined as "*any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations.*"

The CDSM directive sets out an **exception to copyright** (as well as the *sui generis* right to databases and the related right of publishers of press publications) **for text and data mining** for in specific cases:

- for the purposes of scientific research, without prior authorization from the rights holders. In French law, this exception does not apply if a company associated with the organization carrying out the mining has privileged access to the results.

- for any other purpose (including commercial), with lawful access to the content. In this case, the copyright owner (or database maker) may object to this exception and refuse to allow his or her content to be mined for learning purposes. Rights holders thus have an **opt-out right**.

The *opt-out* in the spotlight



The surge in the use of GenAI since 2023 has raised strong debate on the adequacy between technological developments and this mechanism, transposed into French Law in 2021 but relatively unused until now.

Two main issues are currently the focus of public debate:

- *How to exercise the opt-out right with certainty and ensure that it is taken into account by those carrying out text and data mining?*

- *How to check if IP-protected content is used for AI-training purposes in order to exercise the opt-out right?*

This raises two main topics: (i) the **methods for exercising and advertising the opt-out** and (ii) **transparency** regarding the datasets used for AI-training purposes.

What are the methods for exercising the opt-out?

The CDSM directive specifies that the opt-out shall be exercised "*in an appropriate manner, such as* ***machine-readable means***" in the case of content made publicly available online.

In French law, the methods for exercising the opt-out have been specified (Decree n° 2022-928 dated 23 June 2022):

- The opposition of the right holder **does not have to be justified**.
- The opposition of the right holder may be expressed **by any means**. As set out by the directive, the opposition may be expressed by machine-readable means; French law further specifies that these means may include metadata or terms and conditions of a website or a service.

Rights holders therefore have a **high degree of flexibility** in expressing their opposition to text and data mining.

Beyond the legal aspect, the main challenge is to ensure that the opt-out is indeed made known to AI companies. The **Robots Exclusion Protocol** ("**robots.txt**") is a file name used to indicate to robot search tools which portions of a website are allowed for scraping or mining. It is a simple, effective, and recognized first step in expressing the opt-out.

In practice in France, the main collective management organizations have been communicating widely since 2023 on their exercise of their opt-out, whether publicly (terms and conditions, press releases, etc.) or by sending information letters to most AI companies.

Nevertheless, the effectiveness of the opt-out must leave no room for doubt. Collective efforts must be carried out to **draw up common technical standards**, which tomorrow will be **standards recognized by all** regarding the exercise of opt-out. Progress has been made recently on this point, for example with the TDM Reservation Protocol (developed by the World Wide Web Consortium) or the "Do Not Train" tools (developed by Spawning.ai): such tools provide a web protocol capable of expressing the reservation of IP rights with regard to text and data mining. They could be viewed as "*machine-readable means*" (as provided for by the CDSM directive).

→ The main challenge is to draw up **harmonized and internationally recognized standards** to **guarantee a simple and effective opt-out** for all rights holders, whatever their size, nationality, or the type of work.

Transparency

The need for rights holders to determine whether their content is subject to text and data mining put transparency on datasets used for AI-training purposes on top of the negotiations of the European AI Act. To address this issue, the finalized version of the AI Act includes **new transparency requirements for general purpose AI models**.

These new transparency requirements, unprecedented in the world, is a new regulatory tool to guarantee the effectiveness of the opt-out as it offers rights holders greater knowledge as to how their content are used for AI-learning purposes. Pursuant to the AI Act, the future European AI Office will draw up the typology of information to be published.

FOCUS

Transparency on the datasets used: what does the AI Act say?

Providers of general purpose AI models shall make publicly available a **sufficiently detailed summary about the content used for training** of the model, according to a template provided by the AI Office.

This summary shall be generally comprehensive in its scope instead of technically detailed to facilitate parties to exercise and enforce their rights, for example:

- ✓ by listing the **main data collections or sets that went into training the model**, such as large private or public data bases or data archives.
- ✓ by providing a **narrative explanation** about **other data sources** used.

→ The AI Act states that the European AI Office will draw up the template of the information to be made publicly available by providers of general purpose AI models. These requirements shall be balanced: (i) **technically feasible**, (ii) within the reach of providers **without placing an excessive burden on them**, (iii) compatible with the **protection of trade secrets and confidential business information** [recital 107], and (iv) allowing rights holders to have **sufficient information to exercise their opt-out**.

→ An **intense dialogue** between AI companies and collective management organizations will be necessary to achieve these goals.

→ AI Act's transparency requirements strengthens the effectiveness of the opt-out right introduced by the CDSM directive. Together, these two mechanisms constitute a **robust body of law** that (i) **guarantees rights holders the effectiveness** of their opt-out right, (ii) **without hindering the development of GenAI models** in Europe.

→ In any case, every rights holder has **access to the courts** in the event of illicit or inappropriate use of their works. In addition, the CDSM directive makes the exception for text and data mining subject the "**three-step test**" if the exception conflicts with the normal exploitation of the works or prejudices the legitimate interests of the rights holders unreasonably.

Nevertheless, questions may remain as to the effectiveness of the opt-out, for example:

- *How to ensure that the opt-out is respected in all circumstances? Will there be in the future technical processes that "check" whether IP-protected content is being used for learning purposes despite the exercise of the opt-out?*
- *How can updates to catalogs of works and content for which opt-out has been exercised be made public "in real time"?*

These questions are important to guarantee the effectiveness of the rights enshrined in the CDSM directive: **they will be resolved more by technical work than by new regulatory changes**. Europe must be **pioneer in the development of such technical processes** and in the emergence of practices that will become tomorrow's international standards.

Beyond regulations, major challenges for innovation and culture

Europe must have a framework that both protects creation and fosters the development of AI on its soil. Beyond regulations, this debate involves major societal and economic issues, both for our capacity to innovate and for the strength of our cultural model. The growing exercise of the opt-out in Europe could **ultimately limit the volume of content available** in France and Europe for training AI models. This could have direct consequences for the European AI ecosystem:

- The exercise of the opt-out can foster the **constitution of a "market" for licenses to use IP-protected content**, with licenses being concluded between rights holders and AI models providers. This contractualization movement is legitimate and falls within the contractual freedom of each actor. However, this raises the question of the **ability of smaller AI startups and players**, with more limited resources, to operate in this market and **access relevant content** in sufficient volumes. Otherwise, **our ability to create "AI champions" in Europe could be diminished**.
- Less content available in Europe may **reduce the presence of the various European cultures and languages in the learning of the main foundation models**. It is therefore essential to have models developed in other European languages too, to ensure that AI keeps a strong level of **cultural diversity**. Otherwise, only English-developed models translated into other languages would be available on the market. As each language has its own patterns and ways of thinking, the exclusive use of English in foundation models could complicate the use of AI in cultural and educational fields. Initiatives led by French authorities on this issue are positive, such the call for tender "Digital commons for GenAi" (by Bpifrance), the "LANGU:IA" project (both aim to promote French and European content in the training of GenAI) or the Government-developed GenAI tool "Albert".