MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Report Task Force on the implementation of the European regulation establishing harmonized rules on artificial intelligence (« template »)

PRESENTED TO THE HIGH COUNCIL FOR LITERARY AND ARTISTIC PROPERTY

Task Force Chair : Alexandra Bensamoun
Rapporteur : Lionel Ferreira
With the support of Frédéric Pascal

Task Force Chair
Alexandra Bensamoun
University Professor (Law)
Qualified person at CSPLA

Rapporteur
Lionel Ferreira
Master of Requests at the Council of State

With the support of
Frédéric Pascal
University Professor (Applied Mathematics)
Qualified person at CSPLA

Report presented at the CSPLA plenary meeting of 9 December 2024

The Conseil supérieur de la propriété littéraire et artistique [Higher Council for Literary and Artistic Property] is responsible for advising the Minister of Culture on matters of literary and artistic property

The authors of this report are solely liable for its content

AI-generated cover image by MidJourney

Prompt: "*Abstract European flag made of glowing digital data and network connections, on a blue background with yellow stars, in a wide banner design. --ar 121:62 --v 6.1*"

Executive Summary

Collecting and exploiting quality data, particularly cultural data, is of strategic importance for providers of artificial intelligence (AI) templates. However, **quantity of human data** on the web is **declining**, and training an AI model on **synthetic data** leads to its **deterioration**. Paradoxically, despite these observations, data are the only “inputs” in the chain whose **commercial value** is being called into question.

Intended to create a framework favourable to innovation and protective of the rights and values of the European Union, the Artificial Intelligence Act (AI Act) of 13 June 2024 complements the landscape of standards in place, notably the General Data Protection Regulation (GDPR) of 27 April 2016 and the Directive on Copyright and Related Rights in the Digital Single Market of 17 April 2019, the latter two texts having been adopted before the emergence of "mainstream" generative AI.

In particular, article 53 of the AI Act creates a **transparency obligation** that requires providers of general-purpose AI, including where models are published under a free and open license, to put in place a **policy to comply with Union copyright and related rights legislation** (art 53, 1, c) and to make available to the public "**a sufficiently detailed** summary about the content used for training of the general-purpose AI model" (art 53, 1, d). This summary must conform to a template provided by the Artificial Intelligence Office, a department under the European Commission created by the AI Act. The first version is scheduled for January 2025.

The purpose of this task force is to **clarify the scope of the provisions** of article 53, 1, d, and to **propose a summary template** to undergird France's positions at the European level.

The scope of this flash task force is **copyright and related rights**, meaning that it does not address the issue of training based on personal data, the link with other areas of law, in particular competition law, or the diversity of data required to avoid bias and ensure the influence of French culture. These issues do, however, need to be examined in greater detail.

In the task force's view, the compliance policy required by article 53, 1, c of the AI Act and the provision of a sufficiently detailed summary required by article 53, 1, d **cannot be dissociated**. The compliance policy is the inverse of the detailed summary: what the latter says explicitly, the former says implicitly. Together, they form the two sides of the same obligation: the **obligation of transparency**. The summary template must therefore **incorporate the relevant elements of the compliance policy**, particularly those related to the opt-out clause provided for in Article 4 of the 2019 Directive on Copyright and Related Rights in the Digital Single Market.

The purpose of the summary is, as stated in the recitals of the AI Act, to "facilitate the effective implementation and enable the exercise of the data subjects’ rights and other remedies guaranteed under Union law". However, the content of the summary must not compromise **trade secrets**. The degree of detail of the summary must therefore be assessed in the light of this objective and taking this limitation into account. In this context, requirements must be assessed in relation to each other, necessitating **a purpose-driven and holistic interpretation** of the obligation. Indeed, European provisions must be "**effective**", as the CJEU regularly emphasizes.

This emphasis allows for **rigor in identifying the content used**. Contrary to what some AI model providers claim, the summary should not be a simple list of the "main" data sources. In particular, it is essential to require a list of domain names, and even dated URLs. As stated in Recital 107 of the AI Act, the summary must be "**comprehensive in its scope**".

However, technical information which by nature could compromise trade secrets, must be limited (same recital). It follows that a public summary must make it possible to **identify the potential use** of a protected work or content, but **not to detail how this content has been used**. Technical information on tokenization and the filtering process need not be included in this summary.

In other words, the precise list of ingredients can be made public, but not the recipe. Focusing solely on the term 'summary' to minimize the information regarding the key ingredients would amount to disregarding the legislative mandate. The lack of completeness (which justifies the use of the term "summary") therefore targets the recipe—the techniques—not the ingredients—the content.

The summary is therefore the first step on the road to ensuring that the interested party's rights are respected, but not the last. The obligation of transparency establishes a bridge towards respecting rights and creating a market that upholds the value chain. But there are still steps to be taken to build an ethical ecosystem. Practically speaking, how can one exercise and enforce their rights, which also requires access to technical information protected by trade secrecy? And all within the framework of **existing law**, since the task force cannot account for future legislative changes.

Two avenues are open at this stage.

The first involves a **direct dialogue** between rights holders (or their representatives) and AI suppliers, in which, where appropriate, information may be exchanged to enable negotiation. This can be achieved by including a point of contact in the summary to facilitate communication.

The second, as envisioned by the AI Act, involves an administrative authority (preliminarily the AI Office) handling **complaints**, without prejudice to potential litigation, with the aim of avoiding the judicialization of disputes (noting that over 30 lawsuits have been filed in the United States and that GEMA in Germany has initiated an infringement lawsuit against OpenAI). It would also create a **structured space for exchange** and potentially enable **mediation by facilitating dialogue regarding evidence**.

As things stand, it is almost impossible for rights holders to prove that their content has been used. The completeness required of the summary could support this, but it will largely depend on the template adopted by the AI Office. In addition, this requirement, combined with the inability to include information related to trade secrets (such as filtering methods) in the summary, means that some content mentioned in the summary may ultimately not have been used for training the model. The supplier would then have the opportunity to prove that they are complying with the law.

On the basis of these clarifications, the task force proposes adopting an **approach based on content type for the summary template, with an increasing level of detail depending on their sensitivity to legal rights**:

- For open content (public domain or use expressly authorized by the holder under a "free license"), general information is sufficient. On the other hand, if identifiers are available, it is important to mention them;
- For other data, it is essential to require details about the collection methods used to ensure that the data has been gathered in compliance with Union law, specifically indicating the legal basis for the collection. For data harvested from the Internet, URLs and harvesting dates must be disclosed. The training datasets used must be documented. In particular, unique identifiers should be mentioned when available.

The summary must also contain some essential information, such as the point of contact or the existence of any commercial or partnership agreements.

The task force emphasizes that, generally speaking, transparency is a prerequisite for the effectiveness of rights, and opacity inevitably leads to dysfunctional consequences in the market. Here, transparency is the prerequisite for the emergence of an ethical and competitive market, one that is respectful of the value chain and, as such, compensates rights-holding content.

Preliminary Remarks

This "flash task force" was active over a six-month period, working against a tight schedule, and including negotiations at the European level with the purpose of creating a "template" by the Artificial Intelligence (AI) Office. When relevant to better understanding positions, hearings were conducted with contributions by various stakeholders.

Contributors were **national** (French), **European**, and even **international**, representing a wide **range of interests** (rights holders from all sectors, AI suppliers, institutional players).

The positions, which are solely those of the task force, build on the reflections, particularly regarding transparency, carried out by the **Inter-Ministerial AI Commission**, which submitted its report, *AI: Our Ambition for France*, to the French president in March 2024[1]. In line with applicable law, the report also argues that the use of protected content for training AI models can only be done "in compliance with intellectual property rights" (see "Key recommendations").

By way of introduction, it is worth recalling that the right to intellectual property is a **fundamental right**, specifically mentioned in Article 17 of the Charter of Fundamental Rights of the European Union. In domestic law, intellectual property is linked to the right of ownership, which is constitutionally protected.

This right implies a **monopoly**, which translates into exclusive rights over an object, granting holders the power to say **yes or no**, i.e., to agree to or reject the use of their object by a third party, and where applicable to require remuneration. When an exception or limitation encroaches on this exclusivity, it can only do so in a **measured** way; otherwise, there is a risk that the principle becomes the exception and the infringement on the fundamental right becomes **disproportionate** as a result.

In this context, building a market is essential. The transparency obligation imposed by the European legislator is intended to serve as a **lever** for establishing this market by facilitating negotiation.

[1] https://www.elysee.fr/emmanuel-macron/2024/03/13/25-recommandations-pour-lia-en-france

The emergence of generative AI brings innovation and undoubtedly progress, but also risks, particularly economic and cultural. An MIT study published in August 2024 notes that copyrighted data obtained without authorization is frequently used to train AI models.[2] A survey of 348 experts in 68 countries carried out by the United Nations' AI Advisory Body further reveals that intellectual property violations rank high among the risks posed by AI and are a concern for more than half of the respondents.[3]

At a time when technological solutions and the legal framework are proving inadequate to the challenges posed by AI, and in particular the non-consensual use of works to train generative AI models, the European Union has adopted an AI regulation, imposing a logic of transparency on training data, consisting in the implementation of a compliance policy and the public availability of a sufficiently detailed summary relating to such data (I).

For the provisions of this regulation to have a real effect (the CJEU would say "effective"), the content of the summary must itself take account of the compliance policy and contain information needed to help rightsholders exercise and enforce their rights (II).

This leads to guidelines (III) for designing the summary template (IV).

I. Current situation

1. Data collection and processing are of major importance, but are carried out under conditions that do not guarantee respect for EU values and law.

a. Collecting and exploiting human-generated data has become a strategic issue for AI model providers.

- *Large Language models (LLMs), for example, require large amounts of data for training.*

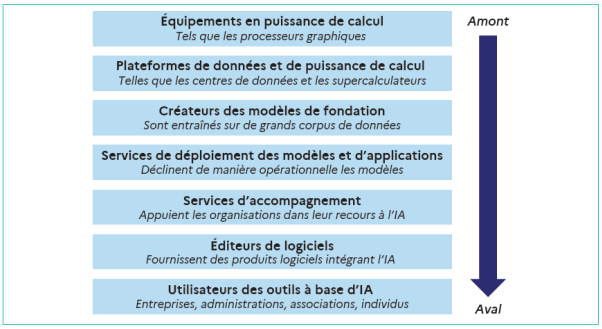
AI model suppliers are in the middle of a value chain that includes, upstream, equipment (graphics processing units—GPUs—for computing power, servers for data storage), data and computing power platforms such as Amazon Web Services, electricity, human talent, and data[4], and, downstream, companies, administrations, organisations, and individuals who derive an economic benefit (including productivity gains and quality) from what is produced by AI models reliant on upstream items.

---

[2] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper et Neil Thompson, *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence*, MIT, August 13 2024, p. 40.

[3] United Nation, AI advisory board, Governing AI for Humanity, September 2024, p. 29.
See also: Hagendorff, *Mapping the Ethics of Generative AI: A comprehensive scoping Review*, University of Stuttgart, 2024; Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Stevie Bergman, A., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., de Haas, S., Ibitoye, Y., Dafoe, A., Goldberg, B., Krier, S., Reese, A., Witherspoon, S., Hawkins, W., Rauh, M., Wallace, D., Franklin, M., Goldstein, J. A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Ringel Morris, M., King, H., Agüera y Arcas, B., Isaac, W., Manyika, J., *The Ethics of Advanced AI Assistants*, Google DeepMind, 2024.

[4] See Commission de l'intelligence artificielle, AI: notre ambition pour la France, March 2024, p. 21 and seq.



Rapport *IA : notre ambition pour la France* [*AI Report: Our Ambition for France*], March 2024

In this chain, nothing would be possible without input, which constitutes a set of key ingredients. These ingredients are acquired (database retrieval, data harvesting), cleaned (filtering and structuring), then prepared (tokenization and vectorization) to obtain an output (text, image, music, etc.).

These data, and particularly cultural data, are the only inputs into the chain whose commercial value, although obvious, is being called into question[5].

- *While the need for quality data is growing, exploitable sources are drying up.*

It has now been scientifically demonstrated that training on **synthetic data** generated by AI models degrades the template's performance and ultimately leads to its **deterioration.**[6] However, an increasing proportion of the data available on the Internet consists of synthetic data, which reduces the quality of the data collected and therefore of the models. The quality and specialization of training data not only impact the quality of the model itself, they can reduce the risk of hallucinations.

The need for data has not, however, diminished. As models are perfected and proliferate, the need increases. Some AI specialists even highlight a **performance plateau** in large language models, given that a significant portion of 'available' data on the Internet has already been utilized.

As the proportion of quality data derived from human interaction declines, the **need for data** grows.

- *The creation of high-quality datasets, especially those generated by human interaction, is becoming crucial.*

Data generated by human interactions is a rare and precious resource. This is particularly true of contemporary human creations, which are essential for creating models that are in step with the times. In its contribution to the UK Parliament's House of Lords Select Committee on Communications and Digital, OpenAI pointed out that limiting data training to **public domain** books and works created over a century ago would be an interesting experiment, but would not provide AI systems tailored to the **needs** of **today's citizens.**[7] Yet contemporary creations are, by design, the most likely to be protected by copyright and/or related rights.

---

[5] Adi Robertson, *Mark Zuckerberg: creators and publishers 'overestimate the value' of their work for training AI*, TheVerge magazine, 25 September 2024.

[6] Ilia Shumailov, Zakhar Shmaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal, *AI models collapse when trained on recursively generated data*, Nature journal, 24 July 2024.

[7] OpenAI written evidence (LLM0113), House of Lords communications and digital select committee inquiry: Large language models (p. 4).

These data may be collected via the Internet. They can be unstructured datasets (URLs or content), such as those provided by Common Crawl, an organisation that has been consistently gathering all the data available on the Internet for several years using “web crawlers” or web harvesters. These databases can then be reworked to create structured datasets. For example, training bases such as LAION or BOOKS3 are another source of supply for training AI models. However, there have been reports of illegal content being included.

Several types of data are used. In addition to data protected by copyright, it is worth mentioning the personal data of social media users, such as X or Meta (videos, sounds, text posted by users), or the instructions given to a generative AI by a user ("prompts"), which may themselves contain personal data or content protected by copyright.

b. A framework for the conditions under which such data is recovered (by harvesting or otherwise) and used is unsatisfactory.

- *No standards are currently in place upstream.*

Numerous technologies exist to enable rights holders to communicate with data harvesters whether and how content can be used.[8] They fall into two main categories, depending on the approach used to identify the content. The first approach is to use identifiers based on the location of the content (website or domain: location-based identifiers). All content present on the virtual site is concerned. This first category includes robots.txt, ai.txt, DeviantArt's noai meta-tags, the use of http headers, and domain registration in a do-not-train registry. The second approach is to allow rights holders to indicate how protected content may or may not be used (unit-based identifiers). This may involve creating an identifier to link metadata to the work (see, for example, the International Standard Content Code). It can also involve establishing reference standards for integrating metadata into digital content in order to trace its provenance. That is the approach of C2PA, the Coalition for Content Provenance and Authenticity, founded by companies like Microsoft and Adobe. Lastly, it may involve registering a work (e.g., haveibeentrained.com).

It is also worth noting the TDMRep opt-out technique, designed by rights holders, which is both location and unit-based.

However, these technologies are more or less effective depending on the type of content being identified. The content localization approach is suitable for text content. The approach based on the identification of works and other protected objects is more relevant to content in other file formats. It is therefore difficult to imagine a verification model that would be appropriate for all forms of content. And even supposing such a system could be created, its real effectiveness would remain doubtful, since rights holders are not always the source of all publications and online postings of their content.

What is more, these technologies are sometimes deliberately ignored, even when, like the "robots.txt" protocol, they are widely available.[9]

Against this backdrop, and while it would appear neither operational in view of the rapid changes taking place in the sector, nor even desirable to impose a single solution, the European Commission, noting that it is difficult to imagine a system that would work for all types of content, is exploring the possibility of creating a centralized rights register, a "unit-based

[8] Paul Keller, *Considerations for opt-out compliance policies by AI model developers*, Open_Future, 16 May 2024.

[9] Katie Paul, *Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says*, Reuters, 21 June 2024.

identifiers" solution, which would complement other technical tools, in line with what Renate Nikolay, Deputy DG of DG Connect, explained on 9 September 2024.[10]

According to the Commission, this register could serve as the foundation for the future **licensing market** for AI model training. Some rights holders, however, strongly argue that the absence of a reference in such a register cannot be considered as implying free use[11] and that, in any case, such a register would require considerable resources to implement and update, and would raise important questions in terms of liability in the event of oversight and/or error. Further information from the Commission will no doubt provide answers to the legitimate questions raised by rights holders.

- *Downstream, unlearning methods are not operational.*

The aim of machine unlearning is to remove information from the knowledge learned by an AI model.

To be exact, unlearning involves retraining a model using a same dataset, without the disputed data. The cost of such retraining makes it an unrealistic solution.

Approximate unlearning can be implemented by a multitude of techniques divided into three groups, depending on whether the unlearning is carried out by modifying the data[12], the learning protocol[13], or the trained model.[14]

However, due to the probabilistic and opaque nature of the learning process, there is currently no reliable indicator for measuring the effectiveness of approximate unlearning, which in theory should not degrade model performance. Additionally, these methods do not always provide verifiable guarantees.[15]

c. The regulation on the protection of personal data and the directive on copyright in the digital single market were adopted at a time when the massive use of content by generative AI models was still unforeseen. As such they may no longer be capable of satisfactorily guaranteeing that the rights of European citizens are being respected.

- *For personal data.*

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, known as the General Data Protection Regulation (GDPR), has a particularly wide scope. It provides a framework for the processing of personal data, understood in a broad sense, carried out by organisations established on the territory of the European Union

[10] *UE La DG Connect tient à son registre de l'« opt-out » de l'IA*, Briefing Médias, Contexte magazine, 12 September 2024.

[11] This would be tantamount to imposing formalities prior to protection—a model prohibited by the Berne Convention.

[12] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, Yisen Wang, *Unlearnable examples: Making personal data unexploitable*, conference paper ICLR 2021, 13 January 2021; Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli, *Fast yet effective machine unlearning*, 31 May 2023.

[13] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot, *Machine Unlearning*, 42e IEEE symposium of security and privacy, 15 December 2020; Yinzhi Cao et Junfeng Yang, *Towards making systems forget with machine unlearning*, IEEE symposium of security and privacy, 2015.

[14] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mohan Kankanhalli, *Can bad teaching induce forgetting? Unlearning in deep networks using an incompetent teacher*, 31 May 2023.
Aditya Golatkar, Alessandro Achille, Stefano Soatto, *External sunshine of the spotless net: selective forgetting in deep networks*, 31 March 2020.

[15] Alexis Léautier, *Comprendre le désapprentissage machine : anatomie du poisson rouge*, CNIL, 26 May 2023.

or by organisations that, regardless of their place of establishment, carry out an activity aimed at targeting or supplying goods and services to European residents.

Each member state has a data protection authority (the CNIL in France) that supports and monitors stakeholders. If the regulation is based on a **logic of compliance** and encourages data protection by design and by default, within an approach that fosters **accountability**, graduated sanctions may be imposed in cases of non-compliance with its obligations.

Recent disputes illustrate the questions raised by the use of personal data by AI model providers. After the Irish Data Protection Commission brought legal proceedings, the platform X pledged in September 2024 that it would no longer use the personal data of its European users to train its artificial intelligence program. The Meta group, targeted by complaints lodged in eleven European countries by the None of Your Business (NOYB) organisation, announced in June 2024 that it was giving up, for the time being, on using the Facebook and Instagram posts of its users in Europe to train its AI models. Last April, NOYB filed a complaint against OpenAI for failing to rectify inaccurate personal data produced by its ChaptGPT service.

In particular, many question the relevance of **legitimate interest** as a legal basis for processing.[16] Specific conditions need to be met to qualify for an exception allowing for processing, as reiterated by the European Data Protection Board in its October 2024[17] guidelines, in line with the case law of the CJEU and the opinions of the G29.

- *For copyright and related rights.*

In particular, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market defines the rules applicable to text and data mining, i.e., "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations" (Article 2, 2).

Article 3 provides an exception to the monopoly on protected content, authorizing research organisations and cultural heritage institutions to carry out text and data mining on works or other protected objects, subject to two conditions. The first is objective, and concerns lawful access to these protected objects. The second, subjective, relates to the purpose of the search: it must be carried out for scientific research purposes. The exception is regulated and specifically targeted. It cannot benefit economic players.

Article 4 provides for a second exception, allowing the use of protected content without a designated purpose. This exception from the monopoly is broader in scope, as it authorizes all players to carry out text and data mining for **any purpose, including commercial ones**. Article 4 upholds the objective condition of **lawful access** to content and introduces a clause under the control of rights holders, stipulating that the exception shall apply on condition that the use of works " has not been expressly reserved by [them] in an appropriate manner, such as machine-readable means in the case of content made publicly available online". This is the so-called ***opt-out*** clause, which implies a return to the monopoly (principle of authorization and, where applicable, remuneration).

However, the condition of lawful access does not appear to be systematically met, as press reports suggest that, for example, novels protected by copyright, or press content protected by related rights, appear in databases used to train large models.[18]

---

[16] See in particular the 2nd series of CNIL fact sheets on AI, subject to consultation until Oct. 2024.

[17] EDPB (CEPD), *Guidelines 1/2024 on processing of personal data based on Article 6(1) (f) GDPR*, 8 Oct. 2024.

[18] Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, The Atlantic, 19 August 2023.

What is more, as we have seen, the effectiveness and efficiency of the opt-out scheme are not currently credible. And one might wonder what is meant by the term "machine-readable".[19] Interpretations can differ, as illustrated by Robert Kneschke vs. LAION e.V., a case brought before the Hamburg District Court. What form must the reservation of rights clause take to be considered "readable" by data harvesting robots on the Internet?[20] Is it the user's responsibility to implement a widely used protocol, such as robots.txt, or should it be the harvester's responsibility to be capable of reading all types of instructions, including those written 'in plain text' within the HTML code? In the absence of standardization and of a common vocabulary to unambiguously define authorized uses, forcing AI providers to take into account the myriad solutions currently available could be considered unreasonably burdensome. However, recital 18 of the 2019 directive specifies that machine-readable processes include "the terms and conditions of a website or a service". This open approach suggests that it is the responsibility of harvesters to give themselves the ability to read available information, and not the responsibility of rights holders to use an imposed technology. What is more, this would be an insurmountable burden for the owners, who are not the originators of all online content.

It is therefore interesting to note that in the above-mentioned case, the Hamburg District Court, whose position will likely have to be confirmed on appeal, also held that a reservation of use written in natural language was "machine-readable" within the meaning of Directive (EU) 2019/790.[21]

The European Commission is interested in this subject, and has organised meetings on the reservation of rights clause. It is also included in the Hungarian Presidency's questionnaire to member states.[22]

2. **To put an end to a situation that is detrimental to innovation and citizens, the European Union has adopted a regulation on artificial intelligence, which includes an obligation of transparency, the scope of which the task force must clarify with a view to negotiations between member states.**

    a. **This situation is detrimental to innovation and citizens.**

- *It is a source of uncertainty for businesses and citizens alike.*

This situation of uncertainty regarding the application of the text and data mining exception, particularly concerning the enforceability of its conditions, is disastrous for small businesses, but also for larger ones, which stress the need for legal certainty to enable the market to flourish.

As is already the case in the United States[23], legal uncertainty can only lead to a proliferation of disputes and transactions that are detrimental to market development.

---

[19] See CSPLA, Report Transposition des exceptions de fouille de textes et de données, December 2020.

[20] See Paul Keller, *Machine readable or not? – notes on the hearing in LAION e.v. vs Kneschke*, 22 July 2024, Institute for Information Law.

[21] Hamburg District Court, 27 September 2024, p 15: "Die Kammer neigt allerdings dazu, als "maschinenverständlich" auch einen allein in "natürlicher Sprache" verfassten Nutzungsvorbehalt anzusehen". However, Mr. Kneschke's claim was rejected on the merits, as the court considered that the organisation benefited from a special provision of German law which authorizes text reproductions and data mining for scientific research purposes, and that Mr. Kneschke had not established that the organisation also pursued commercial ends. On this point, the compliance of the decision with European law is questionable.

[22] *Hungarian Presidency policy questionnaire on the relationship between generative Artificial Intelligence and copyright and related rights*, 27 June 2024.

[23] For copyright in the United States, see for example nine recent cases reviewed by Luiza Jarovsky (LinkedIn): UMG Recordings, Capitol Records, Sony Music Entertainment, Atlantic Recording Corporation, Atlantic Records, Rhino Entertainment, The All Blacks, Warner Music International & Warner Records vs. Suno (24/06/2024); Andre Dubus III & Susan Orlean vs. NVIDIA (02/05/2024); The Intercept Media vs. OpenAI &

Only the establishment of clear and stable conditions will enable the market to develop in a calm and sustainable manner.

- *It is detrimental to citizens, as it encourages a race to the bottom.*

During its hearings, the task force heard that neglecting to exploit data of dubious origin meant running the risk of falling behind in the race for innovation. But rights cannot be sacrificed on the altar of innovation.

Waiting only makes the situation worse, because there is no invisible hand leading to a self-regulating market, at least in this field.

Finally, this status quo will only last for so long, and a reasonable solution will have to be implemented at some point. The passage of time will only reinforce the unequal playing field, with its attendant competitive risks.

- *It hinders innovation.*

This situation strengthens the dominant companies, which are in a position to take on long and costly litigation, or to sign advantageous agreements with rights holders, who may even be forced to negotiate on an exclusive basis, when in fact their business model is to sell their rights to a multitude of players.

Clearly, the unregulated exploitation of all personal and cultural data cannot continue indefinitely, so regulation is inevitable. Delaying its emergence only raises the barrier to entry for future players in this sector, since in the meantime, dominant players can take advantage of existing loopholes to increase their lead.

This situation illustrates the **unproductive opposition between innovation and regulation**. Appropriate regulation is needed to enable innovation[24], so that innovation can be synonymous with progress.

The EU, which lags behind in the sector, has an opportunity to set itself apart by developing a **trustworthy AI model** that respects ethical criteria. This is why the framework had to be modified, which is what the AI Act did.

b. The Artificial Intelligence Act (AI Act) aims to provide a framework that is both innovation-friendly and respectful of EU values.

- *The aim of the AI Act is to improve the functioning of the internal market.*

This regulation[25] creates a framework guaranteeing the free circulation of AI-based goods and services, favourable to innovation, while respecting the values of the European Union.[26]

Microsoft (28/02/2024); The NY Times vs. Microsoft & OpenAI (27/12/2023); Mike Huckabee, Relevate Group, David Kinnaman, Tsh Oxenreider, Lysa TerKeurst & John Blase vs. Meta Platforms, Microsoft, Bloomberg, EleutherAI (17/10/2023); Author's Guild and others vs. Open AI (19/09/2023); J.L., C.B., K.S., P.M., N.G., R.F., J.D. & G.R vs. Google (11/07/2023); Kadrey, Silverman & Golden vs. Meta (07/07/2023); Paul Tremblay & Mona Awad vs. Open AI (28/06/2023). A compilation can be found on this website: https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/

[24] Anu Bradford, *The False Choice Between Digital Regulation and Innovation*, 6 October 2024, Columbia University.

[25] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) 300/2008, (EU) 167/2013, (EU) 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 13 June 2024.

[26] Recital 1:

In an open letter published on 19 September 2024, some 30 companies, including Meta and Spotify, warned that fragmentation and inconsistencies in EU regulation would affect the EU's ability to remain competitive and innovative.[27]

On the contrary, the task force considers that AI Act completes a law that was written prior to wide-scale access to generative AI. It improves coherence among standards. This is particularly true of the link with the GDPR, as the CNIL has pointed out.[28] This is also true of the rules applicable to copyright and related rights, whose effectiveness is called into question by a lack of transparency.

Far from hampering the competitiveness of European companies, the AI Act should enable them to align themselves with the highest standard. The regulation takes up the **principle of extraterritoriality** present in many European texts on digital technology and applies (in accordance with Article 2, clarified by Recital 106[29]) to all companies wishing to operate on the European market. Companies rarely wish to exclude themselves from the European market. And the high cost of training a foundation model will undoubtedly deter suppliers from training a regional model "adapted" to the European market alone.

Following the example of the GDPR in particular, we can also expect the **AI Act to spread** beyond the European Union, once again illustrating the "Brussels effect".[30] A number of bills are already taking the AI Act as a model. In the United States, a federal AI Research, Innovation, and Accountability Act bill has been submitted to the Senate. Furthermore, despite vetoing California's state bill SB 1047 on 29 September 2024, Governor Gavin Newsom has approved eight AI-related texts that will apply in the state. One of these texts, the **AB 2013 law** (appended), lays down a transparency obligation that requires the public disclosure of information relating to training data for generative AI models, including the obligation to specify whether datasets include data protected by copyright. It is also interesting to note the **proposed final judgment**, disclosed in November 2024 by the US Department of Justice (DOJ) in its case against Google, accused of being a monopoly. The DOJ has asked the platform to provide online publishers, sites, and content creators with "an easily usable mechanism" to exercise their right to **opt-out** and prevent their content from being used for training.[31]

In Canada, the Artificial Intelligence and Data Act (AIDA) takes its inspiration from European regulations, as stated in the companion document.

---

[27] *Europe needs regulatory certainty on AI*.

[28] CNIL, 12 July 2024, *Entrée en vigueur du règlement européen sur l'IA : les premières questions-réponses de la CNIL*.

[29] "Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union "

[30] Anu Bradford, *The Brussels Effect*, Columbia Law School, 2012.

[31] US District Court for the District of Columbia, Case No. 1:20-cv-03010-APM, Nov. 2024, p. 12: *"Google must provide online Publishers, websites, and content creators an easily useable mechanism to selectively opt-out of having the content of their web pages or domains used in search indexing; used to train or fine-tune AI models, or AI Products; used in retrieval-augmented generation-based tools; or displayed as AI-generated content on its SERP, and such opt-out must be applicable for Google as well as for users of the Search Index. Google must provide for an opt-out specific to itself and each index user on a user-by-user basis and must transmit all opt-outs to index users in a useable format. Google must offer content creators on Google-owned sites (all Google owned or operated properties including YouTube) the same opt-out provided to Publishers, websites, and content creators. Google must not retaliate against any Publisher, website, or content creator who opts-out pursuant to this provision "*.

- *The AI Act establishes a transparency requirement.*

As data access is crucial for model providers, the transparency obligation has been one of the most debated issues, particularly at the last Trilogue Meeting in December 2023. It applies to all models, including open models.

Guaranteeing transparency with respect to data used to train general-purpose AI models is a vital prerequisite for the emergence of this market. As expressed by the AI Act, transparency is necessary to ensure **fair competition** between suppliers of general-purpose AI models.[32]

The effectiveness of a transparency obligation will also depend on the emergence of **an ethical and competitive market**, one that respects the value chain and compensates all input. A durable economic model can never ground itself on an opaque, uncompensated use of objects belonging to third parties.

Once transparency has been achieved, the market can be established and compensation models clarified.[33] The European Union supports the emergence of a "licensing market".

This is what underlies two obligations stipulated in Article 53 of the AI Act, requiring model providers, including those published under a free and open license, to:

- Implement a policy aimed at complying with the body of EU law on copyright and related rights: point c of 1 of article 53;
- Write "a sufficiently detailed summary of the content used for training of the general-purpose AI model". This summary must be made "publicly available" and conform to a template provided by the AI Office[34], as per point d of the same paragraph.

Providers of AI models will be able to rely on codes of good practice, the drafting of which is encouraged and facilitated by the AI Office[35].

These various processes are underway at the European level.

- *Implementation is the subject of intensive discussions with the Commission and the AI Office.*

A multi-stakeholder consultation on reliable general-purpose AI models under the AI legislation was conducted by the AI Office and closed on 18 September 2024.[36]

The consultation is part of the Act's implementation timetable[37], which stipulates that codes of good practice will be ready by 2 May 2025, and that the transparency obligation will apply from 2 August 2025[38] or, for model providers placed on the market or in service before that date, 2 August 2027.[39] The first deadline applies to all new versions of a model.

Working groups have been set up by the Commission, with the task of drawing up the first Code of Practice for general-purpose AIs[40].

---

[32] Recital 106.

[33] This is the subject of the "AI Compensation" economic and legal mission currently underway at CSPLA.

[34] Office established by European Commission decision of 24 January 2024 (https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:C_202401459) published in the Official Journal of the European Union of 14 February 2024.

[35] Article 56.

[36] AI legislation: Give your opinion on Trustworthy General-Purpose AI | Building Europe's digital future; answers could be submitted via a form.

[37] Timetable for implementation of the European Artificial Intelligence Act (artificialintelligenceact.eu)

[38] Article 113, b.

[39] Article 111, 3.

[40] A first draft was released on 14/11/2024.

On 21 November, at a "Copyright" working group meeting on a Code of Best Practices, the Commission also announced a first draft of the summary template circulated by the AI Office for January 2025.

At the same time, Hungary, which holds the presidency of the Council of the EU until 31 December 2024, has sent each member state a questionnaire concerning, among other things, artificial intelligence. It also indicated that it would pay particular attention to the subject and to preparations for the implementation of the AI Act at European and national levels[41].

c. The task force set up by the Minister of Culture aims to clarify the scope of the provisions of article 53, 1, d, and to propose a summary template that can be used on behalf of France at the European level.

With a short mandate (from 15 April 2024 to 30 November 2024), the task force worked under **current law**. The scope was not to imagine what the AI Act could or should have been, nor to advocate for the introduction of new solutions requiring reforms, but to **clarify the scope of existing provisions**, within the limits of the subject matter at hand (copyright and related rights) and the task force's timetable.

This work will inform France's positions at the European level, as has been pointed out by Thomas Courbe, Director General for Enterprise and France's representative on the European AI Committee, an advisory body created by Article 65 of the AI Act and tasked with ensuring the effective implementation of AI legislation across the EU, in particular by coordinating national authorities.[42]

It is important to add, however, that in view of the final wording of the provision, the **obligation of transparency extends well beyond** content protected by copyright and related rights, which are the primary targets.[43] The wording now includes personal data within the scope of the obligation, which implies clarifying how the AI Act ties in with the GDPR—an issue that the CNIL has begun to investigate, by organising, on 11 October 2023, an initial public consultation on the constitution of learning databases for AI systems, which resulted in the publication of practical fact sheets.[44] Between 10 June and 1 October 2024, the CNIL also ran a new public consultation on new fact sheets, accompanied by a questionnaire devoted to overseeing the development of artificial intelligence systems.[45]

The task force considers this to be a fundamental issue that will undoubtedly draw the attention of the authorities in the coming months. The intersection of personal data law with literary and artistic property is highly relevant in this context, particularly with the development of generative AI that reproduces the voice or image of artists, thereby compounding infringements—copyright, related rights, personal data rights, and personality rights.

In addition, the obligation of transparency should also address the issue of **data representativeness** (training biases that could, in particular, generate and amplify

[41] Program of the Hungarian presidency of the council of the European Union in the second half of 2024, page 37.
[42] Le Monde, Régulation européenne de l'AI : la bataille se poursuit entre créateurs de contenu et entreprises de la tech, 20 June 2024: "There is work to be done to elaborate on the practical implementation of sufficiently detailed summaries. That is why the government has tasked the Conseil supérieur de la propriété littéraire et artistique [Higher Council for Literary and Artistic Property] with two copyright missions. These proposals will inform our positions at the European level."
[43] Recital 107: the summary is drawn up with a view to ensuring transparency of training data "including text and data protected by copyright law" (emphasis added). If training data were limited to texts and data protected by copyright, this clarification would be unnecessary. The set of training data covered by the transparency obligation is therefore broader than just copyrighted content.
[44] Consultation publique - fiches pratiques sur la constitution de bases de données pour la conception de systèmes d'IA - Synthèse des contributions (cnil.fr), February 2024.
[45] Artificial intelligence: new public consultation on the development of AI systems | CNIL, June 2024.

discrimination) as well as the **diversity of cultural expressions and the promotion of French and Francophone culture.**[46]

The relationship with **competition law** provisions, and the powers of the French Competition Authority (ADLC), is another area that remains to be assessed, in particular the possible sanctions that could be imposed when non-compliance with the transparency obligation undermines the functioning of the market (specifically on the grounds of abuse of dominant position) or constitutes an unfair commercial practice that competitors could act upon.[47] Generally speaking, any breach of a compliance obligation is likely to constitute a competitive infringement. Both European and national (French) authorities are working on this issue. In France, the ADLC recommended, in its opinion 24-A-05 of 28 June 2024, that the data market be built by ensuring a "balance between fair compensation for rights holders and access for model developers to the data they need to innovate, taking into account the diversity of data use cases".[48]

These aspects are beyond the scope of the task force's mission and will therefore not be assessed. The proposed summary template will deal only with cultural data, and will not include elements concerning the representativeness of the data.

II. Analysis

1. The obligation to set up a compliance policy and the obligation to make a sufficiently detailed summary available to the public share the same objective: to improve transparency.

a. The AI Act seems to consider that these are two obligations to be addressed in isolation.

Such as they are presented in the AI Act, the summary template (article 53, 1, d) and the internal policy on respecting rights (same provision, point c) are two independent obligations. The risk of treating the two together is that interpretations could fall under copyright, and therefore exceed the scope of the AI Act.

However, that should not be a concern. In fact, in the First Draft of the General-Purpose AI Code of Practice, published by the AI Office on 14 November 2024, the two subjects are dealt with under the same heading ("Rules related to copyright").

b. For the task force, the two obligations are inseparable.

In the task force's view, the AI Act creates a new and autonomous "**compliance by design**" obligation for AI providers, which is only meaningful if the issue of compliance policy and that of a sufficiently detailed summary are addressed jointly.

In fact, these two obligations share the **same objective of transparency**. This analysis is supported by the very wording of Recital 107, which states that sufficiently detailed summaries

[46] V. Francophonie Summit (Oct. 2024): Declaration of Villers-Cotterêts, art. 20.

[47] See in particular Cass. Com., 27 September 2023, no. 21 - 21.995 ruling that a company's failure to comply with its legal obligations may constitute an act of unfair competition. And CJEU (Grand Chamber), 4 October 2024, Lindenapotheke, C. 21/23, which holds that the GDPR does not preclude Member States from providing, in their national law, the possibility for competitors of the presumed violator of this regulation to invoke the violation before civil courts as an unfair commercial practice.

[48] Opinion 24-A-05 of 28 June 2024, on the competitive operation of the generative artificial intelligence sector, p. 95.

shall be made publicly available "in order to increase transparency". Moreover, recital 108 refers to these two obligations as a single obligation[49]. This link is also apparent from the material scope given by the text to the two sub-paragraphs c and d of article 53.1, since they are the only sub-paragraphs of this article that apply to **all** AI suppliers, including those producing free models.[50]

Admittedly, unlike the sufficiently detailed summary, the text of the regulation does not state that the compliance policy must be made available to the public. The regulation only stipulates that model providers are required to "implement" it.[51]

However, the relevance of the information in the sufficiently detailed summary is necessarily assessed in the light of the measures implemented by the supplier to comply with its copyright obligations. The compliance policy is in a way the inverse of the detailed summary: what the latter says explicitly, the former necessarily says implicitly.

Consequently, for the sake of consistency and good articulation between the summary template and the Code of Practice, and because the obligations are complementary and concern the same field of application here (literary and artistic property), the compliance policy should be mentioned in the summary, **at least in outline**. The Code of Practice will undoubtedly be more comprehensive in terms of the elements required.

c. The summary template must include elements relating to compliance, and in particular respect for the reservation of rights.

Without requiring that the compliance policy be detailed in the summary, the task force considers that its main elements should be included.

In particular, the summary template should invite providers to specify which **protocols** are recognized by the data harvesters they use, either directly or via third parties, e.g., whether the robots.txt protocol is respected. However, this protocol would not be the only system accepted. As mentioned, there is no reason to exclude another "machine-readable" process, especially as robots.txt protocol is considered ineffective for some content.

In the case of datasets obtained free of charge or in return for payment from a third party, it should be indicated in particular whether steps have been taken to ensure that these data have been collected in compliance with the law (guarantee of the existence of an authorization or license).

With regard to the text and data mining exception provided for in Directive 2019/790, some rights holders dispute its applicability to the training of AI models.[52] The exception, designed before the emergence of generative AI models, would be aimed exclusively at exploiting the semantic content of works, whereas the exploitation of data by AI models would not be limited

---

[49] "With regard to the obligations imposed on providers of general-purpose AI models to put in place a policy to comply with Union copyright law and make publicly available a summary of the content used for the training, the AI Office should monitor" underlined by the task force.

[50] Recital 104 justifies this as follows: " ... given that the release of general-purpose AI models under free and open-source license does not necessarily reveal substantial information on the data set used for the training or fine-tuning of the model and on how compliance of copyright law was thereby ensured, the exception provided for general-purpose AI models from compliance with the transparency-related requirements should not concern the obligation to produce a summary about the content used for model training and the obligation to put in place a policy to comply with Union copyright law... ".

[51] Art. 53, 1, c.

[52] See, for example the position of the European Writers' Council, *EWC second Statement on the AI Act Proposal*, July 2023: https://europeanwriterscouncil.eu/23ewc_on_aiact/ and more broadly: *Joint Statement to Ursula von der Leyen and the new elected European Parliament on the impact of AI on the European creative community*, July 2024: https://europeanwriterscouncil.eu/247js_aiimpact_europeancreativecommunity/.

to semantic content, but would also extend to syntactic content.[53] Similarly, one could question the compliance of the exception with the three-step test, a filter applicable to all exceptions under Article 5, 5 of Directive 2001/29 and by reference to Article 7 of Directive 2019/790.

This subtle position implies clarifying the current legal framework, no doubt after several years of negotiations (or litigation all the way to the CJEU), which will necessarily be detrimental to rights holders and the emergence of new innovative companies.

Conversely, in terms of the applicability of this exception, it may be noted that the European legislator has explicitly referred to the exception for text and data mining in the AI Act, both in the recitals and in the provisions[54], with Article 53, 1, c, mentioning that the compliance policy must "in particular" aim "to identify and comply with (...) a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790". The European Commission has also taken a logical position in favour of applying the exception, as can be seen from Thierry Breton's answer to a parliamentary question in March 2023.[55]

It is again worth underlining here that in its decision Robert Kneschke v. LAION e.v of 27 September 2024, the Hamburg District Court applies the reservation of rights clause from Article 4 of Directive (EU) 2019/790.

In any case, the AI Act letter explicitly states that the obligation of transparency extends to the question of compliance with the reservation of rights clause. Not mentioning it in the summary template would therefore, as the law stands, mean diminishing the scope of the transparency obligation.

2. Transparency does not mean letting players regulate themselves; it can go as far as requiring a list of content used.

a. The summary cannot be limited to listing the main data sources while awaiting the creation of a data market.

Some model providers consider that the transparency obligation should be limited to listing the main data sources used for training. They argue, firstly, that recital 107 specifies that the summary could include "the main collection or data sets that went into training the model, such as large private or public databases" and, secondly, that account should be taken of "the need to protect trade secrets and confidential business information".

In particular, they argue, a list of URLs from harvested sites should not be required since revealing them would infringe on trade secrecy.

Trade Secrets

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure defines a "trade secret" to be information that meets all of the following conditions: "a) it is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the

[53] Tim W. Dornis, Sebastian Stober, *Urheberrecht und Training generativer KI-templatele - technologische und juristische Grundlagen, 29* August 2024.
[54] Recitals 105 and 106.
[55] https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.html

kind of information in question; / (b) it has commercial value because it is secret; / (c) it has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret;" (article 2, 1).

These three cumulative criteria are set out in article L. 151-1 of the Commercial Code.

Following this logic, it would be sufficient to mention the names of the main datasets, to indicate whether public data have been used (without specifying which), to mention the nature of the data (image, text, etc.), and to explain the principles guiding data processing.

However, this level of information is not sufficient if the aim is to create "**useful**" legislation. It is **insufficient to archive the objective set out by the lawmaker**: "to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law".[56]

Furthermore, recital 107 mentions the main datasets or collections by way of **example**, and not in an exhaustive manner.

On the contrary, this recital even specifies that while trade secrecy may limit the degree of **technical** detail provided by the summary, the "summary should be generally **comprehensive in its scope**" (emphasis added).

Finally, it is worth recalling that the invocation of **trade secrecy** naturally has its **limits**. Under domestic law, article L. 151-7 of the French Commercial Code stipulates that trade secrecy may not be invoked against judicial and administrative authorities acting, in particular, in the exercise of their powers of investigation, control, authorization, or sanction. And we note with interest that, in Dun & Bradstreet Austria GmbH C-203/22 concerning the processing of personal data by an AI, which led to the refusal to terminate or extend a cell phone contract on the grounds that the person did not have sufficient financial solvency, Advocate General Jean Richard de la Tour considered, on 12 September 2024[57], that trade secrets cannot override an individual's right under the GDPR to understand how a decision affecting them is made. This position appears to be transposable to the rights a person derives from the copyright provisions of European texts. Trade secrecy cannot, by emptying a sufficiently detailed summary of substance, override the right that a rightsholder derives from the AI Act to have access to items that can help them "exercise and enforce their rights under Union law"[58]. Finally, the Trade Secrecy Directive even envisages the possibility of a Union rule requiring the disclosure of information to the public, including trade secrets, for reasons of public interest.[59]

Providing a list of URLs, however long it may be, does not therefore appear contrary to the provisions of the AI Act, if it is necessary to achieve the objective sought by the European lawmaker.

Trade secrecy is also difficult to uphold when using the Common Crawl.

While trade secrecy must be respected when a summary is made available to the public, that is not the case when bilateral discussions are underway (confidentiality agreements are frequently

[56]Recital 107.

[57]https://curia.europa.eu/juris/document/document.jsf?text=&docid=290022&pageIndex=0&doclang=FR&mode=req&dir=&occ=first&part=1&cid=2434261

[58]Recital 107.

[59] Article 1, 2: "This Directive shall not affect / (...) / b) the application of Union or national rules requiring trade secret holders to disclose, for reasons of public interest, information, including trade secrets, to the public (...)".

signed in other fields, so that could apply here as well), and even less so when the request comes from an administrative or judicial authority.

b. The text does not exclude the listing of protected content used for model training.

Recital 108 explicitly states that the AI Office cannot verify a provider's compliance with copyright and related rights obligations on a 'work-by-work' basis. This applies both to the implementation of a policy ensuring compliance with EU copyright law and to the publication of the training data summary for public access.[60]

Some stakeholders deduce from this clarification on the AI Office's verification work that the granularity of the summary cannot go down to the level of a single work, or rather of the protected content.

But it only follows from these recitals, which, as we have seen, refer to a summary that is "generally **comprehensive in its scope**", that if the summary includes a list of the content used, the AI Office is not required to check that this list is exhaustive, nor that the use made of this content is lawful. This is a kind of **"admissibility" check**, verifying compliance with formalities, **rather than a substantive examination** of the merits. In this initial phase of checking compliance, a "work-by-work" assessment is excluded, as stated in Recital 108. However, it is possible that **a substantive examination may be carried out at a later stage, particularly in the event of a complaint.**[61]

This is consistent with the overarching role assigned to the AI Office by the regulation, as well as the resources allocated to it, which do not allow for such exhaustive verification as a first step.

c. The normative scope of the summary must be proportionate to the objective pursued: to help interested parties assert their rights.

Article 53, 1, d, necessarily must be "**effective**". The obligation of transparency is therefore not simply a formal obligation that can be avoided by filling in a long administrative form requesting irrelevant information. Its implementation must make sense, as organisations representing creators and rights holders pointed out in a letter to MEPs on 29 October 2024.[62]

To understand the normative scope of this obligation, its objective needs to be considered. The legislator has explicitly stated that the purpose of the sufficiently detailed summary is to help copyright owners "to exercise and enforce their rights under Union law".[63]

Unless ignoring legislative intent, it would be futile—and ineffective—to focus solely on the term "summary". The intent must be understood in the context of the other requirements, in an interconnected fashion. It would not be acceptable to require a summary that fails to meet its objectives and, consequently, holds no normative value.

Thus, **a results-driven and overview reading of article 53, 1, d** gives full meaning to an expression that seems, at first glance, to be an oxymoron if we focus on the terms "summary" and "detailed" alone. **The summary is *sufficiently* detailed to meet this objective.** In other

[60] Recital 108.

[61] See point d below: "Efforts must be pursued to ensure that transparency achieves its intended outcomes, namely creating a market and enabling compensation for content."

[62] Joint letter of creators and rights holders organisations, 29 October 2024, Brussels.

[63] Recital 107.

words, the degree of detail is assessed in relation to the objective, with one limit, that of trade secrecy.

To achieve this objective, it is necessary and sufficient to enable rights holders to determine whether their protected works and objects *have been used*. It does not matter if billions of lines have to be filled in. This is not technically impossible for digital players accustomed to handling massive amounts of data, and rights holders (sometimes through their representatives) are increasingly adept at managing such volumes.

But to achieve this objective, it is not necessary to detail, in a public summary, *how* this data has been used.

In particular, requiring information on data filtering or tokenization processes to be made public would be contrary to trade secrecy.

The use of the term "summary" therefore refers to this lack of completeness due to the non-disclosure of technical information only.

In fact, the AI Act specifies that the summary concerns the "content" used for training, not the "data" used. The first term refers more broadly to sources, while "data" refers to a more structured (organised, filtered) way of representing information. The use of the term "content" is therefore in line with the proposed approach.

This is only the first step on the road to ensuring that rights are respected, but not the last. Given these conditions, what are the next steps? Because, unless we want to empty this obligation of all substance, there has to be a follow-up.

In practical terms, how can rights be exercised and enforced? Does the AI Act provide for a form of "follow-up right" to obtain additional information if necessary?

These questions are not unrelated to the task force's mission, as clarifying how the summary can be used has a reciprocal impact on its content.

d. Efforts must be pursued to ensure that transparency achieves its intended outcomes, namely creating a market and enabling compensation for content.

Clearly, the situation today is unsatisfactory. On the one hand, the task force had access to tangible evidence showing that when a rights holder suspects unauthorized use of protected works and requests additional information, the model provider requires that the rights holder specify the name of the content (or identifiers) and the sources from which this content was allegedly retrieved, which amounts to placing an insurmountable burden of proof on the rights holder. On the other hand, it can be complex for model providers to respond effectively to a constant influx of more or less precise requests from millions of individual people.

A framework for discussion must therefore be proposed.

The AI Act only outlines the rest of the process. But that is not surprising. This regulation is not a copyright text. It limits the Commission's supervisory role to verifying compliance with the obligations it has created, and the AI Office is not intended to identify and sanction potential copyright infringements.[64]

Ignorance of the transparency obligation is not equivalent to copyright infringement. And honouring the obligation of transparency does not guarantee that copyright has necessarily been respected. It is even because the obligation of transparency is respected that a rights holder is

[64] See in particular article 88 and recitals 108 and 161.

put in a position to identify a potential infringement of rights. But the procedure that ensues is more a matter of "*enforcement*" than of substantive law.

In short, the transparency obligation **creates a bridge to copyright and related rights**, but the AI Act does not offer a specific procedure for this.

However, unless the aim is to systematically judicialize the issue by invoking the right to information provided by Directive 2004/48/EC of 29 April 2004, on the enforcement of intellectual property rights[65], or, under general law, measures to gather evidence (CPC, Article 145), all parties have an interest in ensuring that a subsequent procedure allows for an **exchange of information under satisfactory conditions**, particularly with regard to trade secrets

Two approaches, not mutually exclusive, can be envisaged based on the **law as it stands** (as a reminder, the task force mission letter calls for an assessment of the scope of the transparency obligation and a list of the necessary information, without requiring any changes to the applicable law), simply through an interpretation of the AI Act.

The first approach involves a **direct exchange between rights holders (or their representatives) and AI suppliers.** Clearly, an exchange between good-faith actors could help resolve some situations. The summary template would need to designate a single point of contact to enable communication and any direct complaints. In this phase involving professionals, it should be possible to provide details of the actual use of protected content without invoking trade secrecy, preserved by the signing of confidentiality agreements (a common practice in negotiations in other fields). Access to more detailed information is necessary for a transparent assessment of any damage. The same applies to compensation.[66] This first approach does not require any change in positive law. All that would be required is a mention of a point of contact in the summary.

The second approach is generally provided for by the AI Act. Although the specific subject of copyright and related rights was probably not anticipated, the texts are nonetheless applicable. As such, the Commission has the powers to "monitor and control" compliance with the provisions of Chapter V (including relevant provisions—the chapter relating to general-purpose AI models), and the performance of these tasks is entrusted to the **AI Office** (art. 88, 1), which for this purpose has "all the powers of a market surveillance authority" (AI Act, art. 75, 1) within the meaning of Regulation (EU) 2019/1020 of 20 June 2019. In particular, in application of article 88, 2 of the AI Act, the Office is likely well-suited to requesting documentation and information, as provided for in article 91. In addition, Article 85 of the AI Act enshrines the right to lodge **a complaint with the market surveillance authority** responsible for verifying compliance.[67] As a reminder, the compliance required is twofold: article 53, 1, c covers the implementation of measures aimed at respecting copyright and related rights, while article 53, 1, d implies declaring the sources collected for model training. With broad investigative and enforcement powers (art. 14 of regulation 2019/1020), the authority would be responsible for this verification, and its status would protect it from being challenged on the grounds of trade secrecy.

---

[65] See article 8 on the right to information, transposed by law no. 2007-1544 of 29 October 2007. - CPI, art. L. 331-1-2.

[66] The need for a high degree of transparency in the assessment of uses and compensation is identical in the case of the related rights of media publishers (CPI, art. L. 218-4, al. 3), and it is the lack of transparency that is driving the current spate of legal actions.

[67] The task force acknowledges that the subject of the claim can only pertain to the measures outlined in the AI Act and not to the demonstration of a violation of copyright or related rights. The issue of the territoriality of copyright directives remains a complex topic requiring expertise and alignment with the extraterritorial scope of the AI Act.

However, if the authority is given **a legal capacity to act** by the AI Act, it is useful at this stage to question its **operational capacity** to manage claims likely to arise from rights holders in 27 member states[68], unless the new body receives a substantial increase in resources. Given the number of potential complaints and the complexity of the issues, processing times may not be satisfactory. In order to lighten the Office's workload, this task (or part of it) could be delegated to a national authority, according to a procedure yet to be defined. This proposal would have to be assessed on the basis of current law.

At the very least, whether the procedure is European (AI Office) or delegated to a national authority,[69] in the event of information deemed insufficient, inaccurate, or incomplete in both the internal compliance policy and the summary, the supplier will have to respond to the authority's injunctions, in particular by providing proof that the content subject to complaint has not been used. Indeed, at this stage (complaint before an authority), **the oversight carried out by the authority would no longer be procedural but substantial**, in line with the scope of the powers entrusted to a market surveillance authority.

This analysis is favourable to both rights holders and AI providers. For the former, it is easier to establish **proof of use of their content**. For the latter, this procedure is also a means of supplementing the disclosed information **without risking the compromise of competitive data**. Indeed, since, as seen above, the AI Act requires the summary to be comprehensive in its scope, and since the communication of technical rules, particularly filtering rules, could infringe trade secrecy, this procedure makes it possible to confirm or deny the use of protected data, by revealing the results in a secure environment.

This administrative procedure will also offer good-faith players a flexible **mediation** framework, designed to facilitate the resolution of a dispute. It would not be compulsory, however, and would therefore be **without prejudice to legal recourse**.

In the event of legal action, the judge could also **draw all necessary conclusions from the findings made by the authority (violation of the AI Act provisions)**. While revealing the overall method is likely to infringe on trade secrecy, confirming or denying the use of content is not the same thing.

Ultimately, through this claims procedure—as it is interpreted—the AI Act allows for a kind of "**follow-up right**", in the hands of both rights holders and AI providers.

At the public summary stage, the goal is to identify sources collected for training—the ingredients. However, the recipe with its preparation instructions (filtering methods, tokenization and vectorization processes, etc.) is a trade secret and does not need to be included in a public summary. However, in the event of a claim or legal action, the “recipe” must be accessible if the right is to be **effective**.

The task force has derived several guidelines from these elements for the development of the summary template.

[68] As a reminder, claims can only address a breach of the required compliance standards.
[69] Territorial jurisdiction here should depend on the claimant.

III. Guidelines for the summary template.

1. The template must be "simple and useful" to enable the AI provider to develop its summary.

As mentioned, the summary itself must be **complete in terms of content**.

The template, meanwhile, must remain **simple and effective.**[70] It must provide effective guidance for AI model providers.

These two directives are not contradictory: Completeness is not a symptom of complexity but a mark of efficacy. The idea of utility supports an **ends-oriented reading** of the provision.

2. The main elements of the compliance policy should be listed upstream, as they justify the presence or absence of certain elements downstream.

As training an AI model does not, logically, lead to the licit use of illicit data, elements relating to compliance policy will be required for data collected directly as well as from third parties, whether from Common Crawl-type databases or from information provided by users (prompts).

3. Secondly, when it comes to content information, the degree of detail required depends on the reliability of the sources.

Since access to copyright-free content, i.e., content in the public domain or content whose use is expressly authorized by its owner ("free license"), is, by design, lawful, there is no need to require a fine granularity of information. On the other hand, available information (such as identifiers) must be mentioned to enable verification by rights holders. The task force also warns of the lack of global harmonization of protection durations and the need to verify both the actual absence of protection and the scope of any potential authorization.

The same applies to content covered by contractual arrangements. Requiring an AI model supplier to specify with whom it has signed such contracts, or even to provide information about the content of these contracts, would contravene trade secrecy. It may seem paradoxical for companies to state that the list of contracts signed is a trade secret and cannot be made public, while at the same time disclosing in the press the signing of agreements or stating publicly (or in response to letters from holders) that they have no need to sign such agreements. But these statements are themselves part of the company's strategy. Although, in an ideal situation, such a list of past contracts would contribute to greater transparency, the task force considers that respect for trade secrecy prevents it from being made public. On the other hand, it does not seem out of the question to require suppliers to specify whether or not such agreements exist.

For other content, notably public content that is not free of copyright or content that has been licensed for use (e.g., databases that have been made available), more detailed information will be required, as these are the data sources most likely to contain pirated content. Of course, information on the names of the content or rights holders is not available, and it cannot be required to include what cannot be provided. However, the associated metadata and identifiers must be included.

It is also essential to provide a list of URLs specifying harvesting dates; without this, rights holders will be unable to identify the potential use of their works for training purposes, and the objective set by the legislator will not be achieved. There are no insurmountable technical obstacles; companies that have created models with billions of parameters have the capacity to list billions of URLs, and rights managers are becoming increasingly adept at handling such

[70]Recital 107.

vast quantities of data. Legally, the fact that a URL reveals information about the content accessed is not sufficient, as this involves unfiltered data, to constitute a breach of trade secrets.

If a dataset contains both open content and "other content", or if a model provider is unable to distinguish between the two, the entire dataset should be treated as "other content" subject to the highest transparency requirement. This will only encourage companies to consolidate their compliance policies upstream, so as to be able to distinguish open content data from other data.

Generally speaking, the AI provider must ensure that copyright and related rights are respected in the context of the AI Act, by implementing an internal policy. In particular, when a supplier contracts with a third party to use a dataset, it must ensure that copyright and related rights have been respected (compliance with the opt-out clause, existence of upstream licenses, etc.). In a similar sense, to benefit from the "text and data mining" exception, the content needs to have been “legally accessed”.

Therefore, the provider must include in the summary the **methodology** enabling them to address this **dual requirement of legality**.

4. **The summary template should require important contextual information upstream.**

For the reasons already given, the summary must indicate the point of contact at the issuing AI provider.

It must also mention whether the model is created ex nihilo or from another model. And if it performs a simple update.[71]

If the content of the agreements includes trade secrets, as previously mentioned, it would be useful to at least disclose the existence of such agreements.

[71]Recital 109.

IV. Summary template

Contact point for all inquiries and procedures.

Is this a summary update after a change to the model? y / n
If yes, provide a summary of the initial model.

Is the model training based in whole or in part on an existing AI model? y / n
If so, link to the summary of this model.

Is any of the content used subject to agreements? y / n
If so, in which sector?
If the partner is a public body, please specify.

Open content*					
Harvested from the Internet				**From third parties**	
Dataset	Harvesting method	Harvested domain name	Size and type of data (image, sound, multimodal, etc.)	Dataset	If applicable, link to the dataset
(list of datasets)	*(description) (1)*	*(for each dataset, list of harvested domains)*	*(for each dataset, size and data type)*	*(list of datasets)*	*(for each dataset, link if applicable)*

**only for datasets consisting exclusively of open content data, i.e., in the public domain after protection under copyright and related rights, or content whose use has been expressly authorized by the rights holder under a free license. The task force warns about the scope of a free license, which may nevertheless prohibit this type of use. Otherwise, or if there is any doubt on the license, the dataset falls under the category "Other content".*

Other content (not open)											
Harvested directly or by an authorized third party from the Internet					Datasets from third parties			Prompts		Synthetic data generated from human data	
Dataset	Harvesting method	Methodology to ensure compliance with EU law, including opt out, exclusion of pirate websites, etc.	Harvested URLs + harvest date	Size and type of data (image, sound, multimodal, etc.)	Dataset	Methodology to ensure compliance with EU law (e.g., reservation of rights clause, etc.)	Description	Description	Methodology for ensuring compliance with Union law	Dataset	Methodology for ensuring compliance with Union law
(list of datasets with dates)	*(description) (1)*	*(description) (2)*	*(for each dataset, list of URLs with collection date)*	*(for each dataset, size and type of data...) (3)*	*(list of datasets with dates)*	*(description, including legal basis for data collection) (4)*	*(5)*	*(specify whether user prompts can be saved and used for model training)*	*(description) (6)*	*(list of datasets with dates, templates used)*	*(description, including methods to reduce the risk of counterfeiting, etc.)*

A non-exhaustive list of questions that could be included in the summary template by the AI Office:

(1) Identification of harvesters used? Legal basis for harvesting? Harvest dates?
(2) Legal basis for collection? Method used to identify relevant metadata? Methods to prevent metadata deletion during training or content generation? Recognized standards and protocols? Methods used to comply with machine-readable instructions? Other methods used to ensure lawful processing? Methods implemented to ensure subsequent compliance if protected content has been used without authorization? Certification or external oversight used to ensure the legality of the operation? Methods for neutralizing rights-restricted data? Identification of recognized standards (ISBN type, etc.)?
(3) Training data size? Size of each data type?
(4) E.g., data provider guarantee?
(5) Dataset identification and size? Data type? List of unique identifier types (ISBN, DOI, ISAN, etc.) included? Link to the dataset if applicable or, failing that, list of contents (by available identifiers) of the dataset to date with sufficient detail relating to the content to make it easier for rights holders to exercise their rights? For each type of identifier, percentage of items with this identifier?
(6) How to determine whether the prompt reproduces protected data? How is such data neutralized?

Annexes

List of contributors and people interviewed

In view of the short timeframe available (and the summer holidays), the task force held an open session (including other ministries and civil society players outside the CSPLA) from the outset, and gave priority to soliciting written contributions. Hearings and oral exchanges based on contributions also took place.

The task force would like to thank all those who worked so urgently at a time when so many requests were being made.

Autorité de la concurrence [French Competition Authority] (ADLC)

Société civile pour l'administration des droits des artistes et musiciens interprètes [Civil Society for the Administration of Performing Artists' and Musicians' Rights] (ADAMI)

Aday

Alliance de la presse d'information générale [Alliance of General Information Press] (APIG)

Allonia

Autorité de régulation de la communication audiovisuelle et numérique [Audiovisual and Digital Communication Regulatory Authority] (ARCOM)

Association des services internet communautaires [Association of Community Internet Services] (ASIC)

Bibliothèque nationale de France [French National Library] (BNF)

Botscorner

Canal plus

Centre of the Picture Industry (CEPIC)

Centre français d'exploitation du droit de copie [French Centre for the Exploitation of Copyrights] (CFC)

Centre national du cinéma et de l'image animée [National Centre for Cinema and Moving Images] (CNC)

Chambre syndicale de l'édition musicale [Union of Music Publishers]

Commission nationale de l'informatique et des libertés [French Data Protection Authority] (CNIL)

CMI France

Eurocinema

European Authors' societies (GESAC)

European Magazine Media Association

European Newspaper Publishers' Association

European Publishers Council

Federation of European publishers (FEP)

Fireflies.ai

Gaumont

Google

GESTE

GFII

Institut national de l'audiovisuel [French National Audiovisual Institute] (INA)

International Federation of the Phonographic Industry (IFPI)

La Sofia

Les Voix (organisation)

Ligue des auteurs professionnels [League of Professional Authors]

Linkup

Microsoft

Mistral AI

News Media Europe

Numeum

Open Future Foundation

Panodyssey

French Ministry of Culture / French Ministry of Economy, Finance, and Industry - Pôle d'Expertise de la Régulation Numérique [Digital Regulation Expertise Centre]

Ministry of the Economy, Finance, and Industry - General Directorate for Enterprises

Photoroom

PopScreen Games

Procirep

RELX

Société civile des Auteurs Réalisateurs Producteurs [Society of Authors, Directors, and Producers] (ARP)

Société des auteurs et compositeurs dramatiques [Society of Dramatic Authors and Composers] (SACD)

Syndicat des éditeurs de la presse magazine [Trade Union for Magazine Press Publishers] (SEPM)

Trust My content

Société des auteurs, compositeurs et éditeurs de musique [Society of Authors, Composers and Publishers of Music] (SACEM)

SAIF

Société civile des auteurs multimédia [Society of Multimedia Authors] (SCAM)

Société civile des producteurs phonographiques [Society of phonographic producers] (SCPP)

Société des auteurs dans les arts graphiques et plastiques [Society of Authors in the Graphic and Plastic Arts] (ADAGP)

Société des Gens de Lettres [Society of Literary Authors] (SGDL)

Syndicat des catalogues de films de patrimoine [Heritage Film Catalogues Union] (SCFP)

Syndicat des éditeurs de la presse magazine [Trade Union for Magazine Press Publishers] (SEPM)

Syndicat national des auteurs et compositeurs [French National Union of Authors and Composers] (SNAC)

Syndicat national de l'édition [French National Publishing Union] (SNE)

Syndicat national de l'édition phonographique [French National Phonographic Publishing Union] (SNEP)

Syndicat des producteurs indépendants [Union of Independent Producers] (SPI)

SPEDIDAM

STM

Union des producteurs phonographiques français indépendants [Union of Independent French Phonographic Producers] (UPFI)

Union Nationale des Syndicats d'Artistes Musiciens [French National Union of Musician Artists]

Vivendi

Task Force Mission Letter

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Conseil supérieur de la propriété littéraire et artistique
[Higher Council for Literary and Artistic Property]

Ms. Alexandra Bensamoun
University Professor

Paris, 12 April 2024

SUBJECT: Task force on the implementation of the European regulation establishing harmonized rules on artificial intelligence

Madam,

Article 53 of the draft European regulation establishing harmonized rules on artificial intelligence (AI) includes an obligation for providers of general-purpose AI models to take measures to respect copyright and, in particular, the framework laid down by the 17 April 2019 directive on copyright and related rights in the Digital Single Market (DAMUN). Among these measures, suppliers must develop and make publicly available a "sufficiently detailed summary" of the data used to train their model.

This transparency on the sources that have enabled AI systems to be trained upstream is essential to enable copyright and related rights holders to check that the conditions for lawful access to and use of their works and services—and in particular their possible opposition to any data mining ("opt out")—have been respected.

However, the scope of this transparency obligation is subject to a number of limitations, set out in the draft regulation, the implementation of which needs to be clarified. This applies in particular to the scope of suppliers concerned by this obligation, the level of detail of the information to be provided, the impact of industrial and trade secrets on the disclosure of information, and the form of the disclosure thus imposed.

To facilitate the implementation of this transparency obligation, the draft regulation has tasked the European Artificial Intelligence Office, created by a European Commission decision of 24 January 2024, with developing a simple and effective summary template for the training data used by AIs. In carrying out this task, the office will consult stakeholders, including experts from the scientific and educational communities, citizens, civil society organisations, and social partners.

The recent report by the Artificial Intelligence Commission, set up by the Government in 2023 (*IA: notre ambition pour la France*), meanwhile, recommends "implementing and evaluating the transparency obligations set out in the European AI Regulation by encouraging the development of standards and a suitable infrastructure".

Following on from your previous work (report on the legal and economic challenges of artificial intelligence in the cultural creation sectors, January 2020, and report on text and data mining exceptions, December 2020), the Minister of Culture would like the CSPLA to launch a new task force, firstly, to assess the scope of the transparency obligation set out in the European regulation, taking into account the questions mentioned above, and, secondly, to draw up a list of the information that you feel must be communicated, according to the cultural sectors concerned, to enable authors and holders of related rights to exercise their rights.

I am entrusting you with this mission, for which you will be assisted by a rapporteur. You will also be able to rely on departments at the Ministry of Culture, in particular the General Secretariat (legal and international affairs department). You will hold hearings with CSPLA members, as well as with entities and personalities whose contributions you consider useful, in particular the departments of the Ministry of the Economy, Finance, and Industrial and Digital Sovereignty. Professor Frédéric Pascal, a member of the CSPLA, may also assist in your work.

It would be desirable to be able to present the progress of your work at the next CSPLA plenary session in early summer, and to be able to present your report in December, after discussions with interested CSPLA members.

Thank you for accepting this mission.

Yours faithfully,

Olivier Japiot

Chairman of CSPLA

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (extracts)

Recitals

(1) The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation. This Regulation ensures the free movement, cross-border, of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation.

(…)

(104) The providers of general-purpose AI models that are released under a free and open-source licence, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available should be subject to exceptions as regards the transparency-related requirements imposed on general-purpose AI models, unless they can be considered to present a systemic risk, in which case the circumstance that the model is transparent and accompanied by an open-source license should not be considered to be a sufficient reason to exclude compliance with the obligations under this Regulation. In any case, given that the release of general-purpose AI models under free and open-source licence does not necessarily reveal substantial information on the data set used for the training or fine-tuning of the model and on how compliance of copyright law was thereby ensured, the exception provided for general-purpose AI models from compliance with the transparency-related requirements should not concern the obligation to produce a summary about the content used for model training and the obligation to put in place a policy to comply with Union copyright law, in particular to identify and comply with the reservation of rights pursuant to Article 4(3) of Directive (EU) 2019/790 of the European Parliament and of the Council ([40]).

(105) General-purpose AI models, in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other

subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works.

(106) Providers that place general-purpose AI models on the Union market should ensure compliance with the relevant obligations in this Regulation. To that end, providers of general-purpose AI models should put in place a policy to comply with Union law on copyright and related rights, in particular to identify and comply with the reservation of rights expressed by rightsholders pursuant to Article 4(3) of Directive (EU) 2019/790. Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union.

(107) In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form

(108) With regard to the obligations imposed on providers of general-purpose AI models to put in place a policy to comply with Union copyright law and make publicly available a summary of the content used for the training, the AI Office should monitor whether the provider has fulfilled those obligations without verifying or proceeding to a work-by-work assessment of the training data in terms of copyright compliance. This Regulation does not affect the enforcement of copyright rules as provided for under Union law.

(…)

(161) It is necessary to clarify the responsibilities and competences at Union and national level as regards AI systems that are built on general-purpose AI models. To avoid overlapping competences, where an AI system is based on a general-purpose AI model and the model and system are provided by the same provider, the supervision should take place at Union level through the AI Office, which should have the powers of a market surveillance authority within the meaning of Regulation (EU) 2019/1020 for this purpose. In all other cases, national market surveillance authorities remain responsible for the supervision of AI systems. However, for general-purpose AI systems that can be used directly by deployers for at least one purpose that is classified as high-risk, market surveillance authorities should cooperate with the AI Office to carry out evaluations of compliance and

inform the Board and other market surveillance authorities accordingly. Furthermore, market surveillance authorities should be able to request assistance from the AI Office where the market surveillance authority is unable to conclude an investigation on a high-risk AI system because of its inability to access certain information related to the general-purpose AI model on which the high-risk AI system is built. In such cases, the procedure regarding mutual assistance in cross-border cases in Chapter VI of Regulation (EU) 2019/1020 should apply *mutatis mutandis*.

(…)

(156) In order to ensure an appropriate and effective enforcement of the requirements and obligations set out by this Regulation, which is Union harmonisation legislation, the system of market surveillance and compliance of products established by Regulation (EU) 2019/1020 should apply in its entirety. Market surveillance authorities designated pursuant to this Regulation should have all enforcement powers laid down in this Regulation and in Regulation (EU) 2019/1020 and should exercise their powers and carry out their duties independently, impartially and without bias. Although the majority of AI systems are not subject to specific requirements and obligations under this Regulation, market surveillance authorities may take measures in relation to all AI systems when they present a risk in accordance with this Regulation. Due to the specific nature of Union institutions, agencies and bodies falling within the scope of this Regulation, it is appropriate to designate the European Data Protection Supervisor as a competent market surveillance authority for them. This should be without prejudice to the designation of national competent authorities by the Member States. Market surveillance activities should not affect the ability of the supervised entities to carry out their tasks independently, when such independence is required by Union law.

(157) This Regulation is without prejudice to the competences, tasks, powers and independence of relevant national public authorities or bodies which supervise the application of Union law protecting fundamental rights, including equality bodies and data protection authorities. Where necessary for their mandate, those national public authorities or bodies should also have access to any documentation created under this Regulation. A specific safeguard procedure should be set for ensuring adequate and timely enforcement against AI systems presenting a risk to health, safety and fundamental rights. The procedure for such AI systems presenting a risk should be applied to high-risk AI systems presenting a risk, prohibited systems which have been placed on the market, put into service or used in violation of the prohibited practices laid down in this Regulation and AI systems which have been made available in violation of the transparency requirements laid down in this Regulation and present a risk.

(…)

Articles

(…)

SECTION 2

Requirements for high-risk AI systems

Article 53

Obligations of providers of general-purpose AI models

1. Providers of general-purpose AI models shall:

(a) draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation, which shall contain, at a minimum, the information set out in Annex XI for the purpose of providing it, upon request, to the AI Office and the national competent authorities;
(b) draw up, keep up-to-date and make available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets in accordance with Union and national law, the information and documentation shall:

   (i) enable providers of AI systems to have a good understanding of the capabilities and limitations of the general-purpose AI model and to comply with their obligations pursuant to this Regulation; and
   (ii) contain, at a minimum, the elements set out in Annex XII;
(c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;
(d) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

2. The obligations set out in paragraph 1, points (a) and (b), shall not apply to providers of AI models that are released under a free and open-source licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception shall not apply to general-purpose AI models with systemic risks.

3. Providers of general-purpose AI models shall cooperate as necessary with the Commission and the national competent authorities in the exercise of their competences and powers pursuant to this Regulation.

4. Providers of general-purpose AI models may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.

5. For the purpose of facilitating compliance with Annex XI, in particular points 2 (d) and (e) thereof, the Commission is empowered to adopt delegated acts in accordance with Article 97 to detail measurement and calculation methodologies with a view to allowing for comparable and verifiable documentation.

6. The Commission is empowered to adopt delegated acts in accordance with Article 97(2) to amend Annexes XI and XII in light of evolving technological developments.

7. Any information or documentation obtained pursuant to this Article, including trade secrets, shall be treated in accordance with the confidentiality obligations set out in Article 78.

(…)

CHAPTER IX

POST-MARKET MONITORING, INFORMATION SHARING AND MARKET SURVEILLANCE

(…)

SECTION 4

Remedies

Article 85

Right to lodge a complaint with a market surveillance authority

Without prejudice to other administrative or judicial remedies, any natural or legal person having grounds to consider that there has been an infringement of the provisions of this Regulation may submit complaints to the relevant market surveillance authority.

In accordance with Regulation (EU) 2019/1020, such complaints shall be taken into account for the purpose of conducting market surveillance activities, and shall be handled in line with the dedicated procedures established therefor by the market surveillance authorities.

SECTION 5

Supervision, investigation, enforcement and monitoring in respect of providers of general-purpose AI models

Article 88

Enforcement of the obligations of providers of general-purpose AI models

1. The Commission shall have exclusive powers to supervise and enforce Chapter V, taking into account the procedural guarantees under Article 94. The Commission shall entrust the implementation of these tasks to the AI Office, without prejudice to the powers of organisation of the Commission and the division of competences between Member States and the Union based on the Treaties.

2. Without prejudice to Article 75(3), market surveillance authorities may request the Commission to exercise the powers laid down in this Section, where that is necessary and proportionate to assist with the fulfilment of their tasks under this Regulation.

(…)

Article 91

Power to request documentation and information

1. The Commission may request the provider of the general-purpose AI model concerned to provide the documentation drawn up by the provider in accordance with Articles 53 and 55, or any additional information that is necessary for the purpose of assessing compliance of the provider with this Regulation.

2. Before sending the request for information, the AI Office may initiate a structured dialogue with the provider of the general-purpose AI model.

3. Upon a duly substantiated request from the scientific panel, the Commission may issue a request for information to a provider of a general-purpose AI model, where the access to information is necessary and proportionate for the fulfilment of the tasks of the scientific panel under Article 68(2).

4. The request for information shall state the legal basis and the purpose of the request, specify what information is required, set a period within which the information is to be provided, and indicate the fines provided for in Article 101 for supplying incorrect, incomplete or misleading information.

5. The provider of the general-purpose AI model concerned, or its representative shall supply the information requested. In the case of legal persons, companies or firms, or where the provider has no legal personality, the persons authorised to represent them by law or by their statutes, shall supply the information requested on behalf of the provider of the general-purpose AI model concerned. Lawyers duly authorised to act may supply information on behalf of their clients. The clients shall nevertheless remain fully responsible if the information supplied is incomplete, incorrect or misleading.

AB Act 2013 (California): Generative artificial intelligence: training data transparency

THE PEOPLE OF THE STATE OF CALIFORNIA DO ENACT AS FOLLOWS:

SECTION 1.

Title 15.2 (commencing with Section 3110) is added to Part 4 of Division 3 of the Civil Code, to read:

TITLE 15.2. Artificial Intelligence Training Data Transparency

3110.

For purposes of this title, the following definitions shall apply:

(a) "Artificial intelligence" means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

(b) "Developer" means a person, partnership, state or local government agency, or corporation that designs, codes, produces, or substantially modifies an artificial intelligence system or service for use by members of the public. For purposes of this subdivision, "members of the public" does not include an affiliate as defined in subparagraph (A) of paragraph (1) of subdivision (c) of Section 1799.1a, or a hospital's medical staff member.

(c) "Generative artificial intelligence" means artificial intelligence that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence's training data.

(d) "Substantially modifies" or "substantial modification" means a new version, new release, or other update to a generative artificial intelligence system or service that materially changes its functionality or performance, including the results of retraining or fine tuning.

(e) "Synthetic data generation" means a process in which seed data are used to create artificial data that have some of the statistical characteristics of the seed data.

(f) "Train a generative artificial intelligence system or service" includes testing, validating, or fine tuning by the developer of the artificial intelligence system or service.

3111.

On or before January 1, 2026, and before each time thereafter that a generative artificial intelligence system or service, or a substantial modification to a generative artificial intelligence system or service, released on or after January 1, 2022, is made publicly available to Californians for use, regardless of whether the terms of that use include compensation, the developer of the system or service shall post on the developer's internet website documentation regarding the data used by the developer to train the generative artificial intelligence system or service, including, but not be limited to, all of the following:

(a) A high-level summary of the datasets used in the development of the generative artificial intelligence system or service, including, but not limited to:

(1) The sources or owners of the datasets.

(2) A description of how the datasets further the intended purpose of the artificial intelligence system or service.

(3) The number of data points included in the datasets, which may be in general ranges, and with estimated figures for dynamic datasets.

(4) A description of the types of data points within the datasets. For purposes of this paragraph, the following definitions apply:

(A) As applied to datasets that include labels, "types of data points" means the types of labels used.

(B) As applied to datasets without labeling, "types of data points" refers to the general characteristics.

(5) Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.

(6) Whether the datasets were purchased or licensed by the developer.

(7) Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140.
(8) Whether the datasets include aggregate consumer information, as defined in subdivision (b) of Section 1798.140.
(9) Whether there was any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the artificial intelligence system or service.
(10) The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing.
(11) The dates the datasets were first used during the development of the artificial intelligence system or service.
(12) Whether the generative artificial intelligence system or service used or continuously uses synthetic data generation in its development. A developer may include a description of the functional need or desired purpose of the synthetic data in relation to the intended purpose of the system or service.
(b) A developer shall not be required to post documentation regarding the data used to train a generative artificial intelligence system or service for any of the following:
(1) A generative artificial intelligence system or service whose sole purpose is to help ensure security and integrity. For purposes of this paragraph, "security and integrity" has the same meaning as defined in subdivision (ac) of Section 1798.140, except as applied to any developer or user and not limited to businesses, as defined in subdivision (d) of that section.
(2) A generative artificial intelligence system or service whose sole purpose is the operation of aircraft in the national airspace.
(3) A generative artificial intelligence system or service developed for national security, military, or defense purposes that is made available only to a federal entity.