



Text and Data Mining exceptions in the Generative AI Era: preserving European innovation and competitiveness

Marco Scialdone*

Abstract: the rapid ascension of generative Artificial Intelligence (AI) has sparked intense debate surrounding copyright frameworks and upstream data ingestion. This paper examines the critical role of Text and Data Mining (TDM) exceptions, specifically under Articles 3 and 4 of the EU Copyright in the Digital Single Market (CDSM) Directive, as the foundational infrastructure for European AI development. By analyzing historical technological shifts, recent European legislative implementations (including the Italian Copyright Law updates and the EU AI Act), and comparative international models like Japan's "non-enjoyment" framework, this brief argues against protectionist rollbacks. Restricting TDM exceptions introduces severe market consolidation risks, drives innovation out of the European Union, and exposes consumers to localized algorithmic biases from foreign-trained models. Instead, this paper advocates for a balanced approach that pairs robust TDM exceptions with standardized machine-readable opt-outs and statutory monetization mechanisms at the commercialization phase.

Contents: 1. The historical continuum of technology and creative panics - 2. The regulatory landscape of text and data mining in Europe – 2.1. National implementations: the Italian precedent – 2.2. The AI Act and the Code of Practice – 2.3. Judicial clarification – 3. Comparative models: the Japanese "non-enjoyment" framework - 4. Market and economic implications of restrictive copyright regimes - 5. Strategic recommendations for a balanced AI data infrastructure.

1. The Historical Continuum of Technology and Creative Panics

The contemporary regulatory anxiety surrounding generative AI and copyright is not an unprecedented phenomenon¹. Legal frameworks protecting creativity have historically evolved in response to technological disruptions, beginning with the invention of the movable type printing press in 1450. The subsequent proliferation of the publishing industry

¹ M. Scialdone, *Intelligenza artificiale e diritto d'autore: la tutela giuridica della creatività non umana*, in (Edited by), G. Taddei Elmi e Alfonso Contaldo, *Intelligenza Artificiale. Algoritmi giuridici, ius condendum o "fantadiritto"?*, p. 201-216, Pacini Giuridica, Pisa, 2020.

necessitated the creation of "book privileges" (the direct historical precursors to the modern right of reproduction) predating comprehensive intellectual property regimes.

More recently, two distinct technological revolutions have forced adjustments to copyright law:

- **the advent of the home computer (1980s):** the large-scale diffusion of personal computers disrupted the traditional original/copy dichotomy. Digital works could suddenly be duplicated into an infinite number of perfect copies at zero marginal cost, challenging the traditional prerequisite of individual author authorization.
- **the proliferation of the Internet (late 1990s):** the digitization and global distribution of intellectual works enabled a paradigm shift from a *"Read-Only Culture"* to a *"Read-Write Culture"*, where existing content serves as the structural foundation for new creations, fostering a widespread *"Remix Culture"*².

While these historical phases positioned machines as functional tools for human artists, generative AI introduces a radical inversion: the human now serves as an input provider functional to the production of outputs by artificial intelligence systems. History demonstrates that systemic panics within cultural industries ultimately expand, rather than destroy, the creative ecosystem:

- **1839:** French painter Paul Delaroche declared, *"From today, painting is dead!"* upon witnessing the daguerreotype. Instead, photography freed painting to explore impressionism and abstraction.
- **1906:** musicians protested the phonograph, claiming "talking machines" would destroy vocal cords and eliminate artistic development³.
- **1982:** the Motion Picture Association of America famously equated the VCR to the film industry as *"the Boston Strangler is to a woman alone at home"*⁴. VCR, later, became one of Hollywood's most lucrative revenue streams.

Copyright law has consistently adapted to facilitate technological transitions rather than block them. Text and Data Mining represents the modern extension of this continuum: it functions not to plagiarize or copy expressions for consumer enjoyment, but as an automated analytical technique to identify patterns, trends, and correlations to generate entirely new knowledge.

2. The regulatory landscape of Text and Data Mining in Europe

² L. Lessig, *Remix: making art and commerce thrive in the hybrid economy*, New York, Penguin Press, 2009.

³ L. Lessig, *ibidem*, "When I was a boy ... in front of every house in the summer evenings, you would find young people together singing the songs of the day or old songs. Today, you hear these infernal machines going night and day. We will not have a vocal cord left. The vocal cord will be eliminated by a process of evolution, as was the tail of man when he came from the ape" (quoting John Philip Sousa's speech in 1906 before U.S. Capitol to warn legislators about the phonograph)

⁴ In 1982, MPAA president Jack Valenti famously testified before Congress, stating: *"I say to you that the VCR is to the American film producer and the American public as the Boston Strangler is to the woman home alone"*, <https://time.com/archive/6908240/what-jack-valenti-did-for-hollywood/>.

Under the current CDSM Directive framework, Article 3 establishes a mandatory exception for research and cultural institutions, while Article 4 creates a broader exception for any entity, subject to a rightsholder "opt-out". This structure serves as a resilient pillar for balanced technological growth, as demonstrated by its integration into national legal systems to address AI training.

2.1. National implementations: the Italian precedent

In Italy, Article 25, paragraph 1, letter b) of Law No. 132 of 23 September 2025 introduced Article 70-*septies* into the Italian Copyright Law (*Legge sul Diritto d'Autore* – LDA). This provision clarifies the application of existing exceptions to advanced computational models:

- it explicitly confirms that reproductions and extractions for the purposes of text and data mining through AI models and systems, including generative AI, are permitted, provided access to the underlying databases was lawfully obtained.
- rather than enacting a novel exception, the provision mandates compliance with existing structures. Article 70-*ter* permits extractions for educational or non-commercial scientific research (requiring author acknowledgment), while Article 70-*quater* applies to commercial uses, rendering the machine-readable opt-out mechanism critical.

2.2. The AI Act and the Code of Practice

Concurrently, Regulation (EU) 2024/1689 (AI Act) reinforces copyright compliance for General Purpose Artificial Intelligence (GPAI) model providers during the training phase⁵, independent of the jurisdiction where the training occurs. Providers must fulfill two primary obligations⁶:

⁵ See Recital 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 for 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 authorization 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 purposes of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve the rights on their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where the right to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorization from rightsholders if they want to carry out text and data mining over such works”.

⁶ See Recital 106: “...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 less stringent copyright standards than those provided in the Union”

1. implement a policy to identify and respect the reservations of rights pursuant to the CDSM Directive.
2. draw up and make publicly available a sufficiently detailed summary of the content utilized for training.

The *Code of Practice for General-Purpose AI Models* (published 10 July 2025), a voluntary instrument drafted by independent experts and recognized by the European Commission, is predicated on the assumption that TDM exceptions fully encompass generative AI training, thereby offering legal certainty and reducing administrative burdens for signatories.

2.3. Judicial clarification

This regulatory view is supported by current judicial precedent. The Regional Court of Hamburg (judgment of 27 September 2024⁷, subsequently upheld on appeal⁸) established that TDM exceptions are applicable to generative AI training datasets. In a suit brought by a photographer against a non-profit organization creating image-text datasets, the court ruled that the reproduction fell within the scientific research exception, even if the upstream training could eventually facilitate the creation of competing works. In an *obiter dictum*, the court reaffirmed that under commercial exceptions (Art. 4), rightsholder opt-outs can be effectively executed via machine-readable natural language, such as website terms and conditions.

3. Comparative models: the Japanese "non-enjoyment" framework

In evaluating future optimizations of the CDSM Directive, the European Union can draw insights from Japan's legislative framework. Through its 2018 amendments, Japan introduced Article 30-4 of the Japan Copyright Act (JCA), establishing a highly favorable environment for machine learning.

By drawing the boundary of protection around the *endogenous enjoyment* of a work, the Japanese framework separates the commercial market for creative consumption from technical data processing.

Significantly, Article 30-4 overrides private contractual restrictions. If a developer acquires lawful access to a work, subsequent copying for non-enjoyment data analysis remains non-infringing, even if website terms and conditions attempt to prohibit TDM.

⁷ Germany - Hamburg District Court, 310 O.22723, LAION v Robert Kneschke, [27 September 24] <https://www.euipo.europa.eu/en/law/recent-case-law/germany-hamburg-district-court-310-o-22723-laion-v-robert-kneschke>

⁸ Higher Regional Court Hamburg Confirms AI Training was Permitted (Kneschke v. LAION) [https://www.twobirds.com/en/insights/2025/germany/higher-regional-court-hamburg-confirms-ai-training-was-permitted-\(kneschke-v,-d,-laion\)](https://www.twobirds.com/en/insights/2025/germany/higher-regional-court-hamburg-confirms-ai-training-was-permitted-(kneschke-v,-d,-laion))

To clarify the boundaries of "unreasonable prejudice" to rightsholders, the Japanese Agency for Cultural Affairs published its *General Understanding on AI and Copyright in Japan* in March 2024⁹. The document defines standard "enjoyment" markets by medium:

- **Literary works:** reading text for its narrative, intellectual, or informational value.
- **Musical works:** listening to and appreciating the composition.
- **Artistic/Cinematographic works:** viewing and experiencing the visual expression.

When an AI system tokenizes or processes these works for mathematical calculations, the expressive elements are not perceived by human senses (constituting "non-enjoyment"). However, the *General Understanding* outlines two critical exceptions where the statutory allowance is deactivated:

1. the unauthorized reproduction of fee-based databases compiled specifically for computational analysis directly competes with the rightsholder's core market.
2. if a dataset is curated intentionally to mimic a specific creator's expressive style to generate competing works, the purpose of enjoyment coexists with training, constituting infringement absent explicit consent.

Importantly, the Japanese system separates the *upstream learning phase* (protected under Article 30-4) from the *downstream output phase*. Standard infringement rules apply to AI-generated outputs: if an output exhibits substantial similarity and reliance relative to a copyrighted work, it remains actionable.

4. Market and economic implications of restrictive copyright regimes

A robust and open TDM framework is critical to preserving market competition and economic viability within the digital single market. Imposing restrictive licensing requirements or narrowing current exceptions introduces three structural risks:

- **Market consolidation and monopolization:** mandating granular, hyper-fragmented licensing agreements for upstream training data requires vast compliance capital. Only a limited number of massive, legacy technology conglomerates possess the resources to secure sweeping global licensing deals. European startups, SMEs, and open-source innovators would be systematically locked out, resulting in a highly concentrated market characterized by monopolistic pricing and diminished choice.
- **Innovation deficit:** if the European regulatory environment becomes a hostile deadlock, developers will relocate operations or source pre-trained models from jurisdictions with flexible data ingestion frameworks (such as Japan or the United States). Consequently, Europe risks becoming a passive importer of foreign AI technologies rather than a global leader.
- **Localized algorithmic bias:** sourcing pre-trained AI models from foreign jurisdictions poses severe risks. For example, financial risk-scoring or loan-approval models trained exclusively on foreign datasets will inherently reflect foreign socio-economic realities,

⁹ See, https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf

legal standards, and cultural biases. To ensure safe, legally compliant, and culturally aligned AI systems, European developers must be permitted to train models on broad, diverse, and legally compliant European data pools.

5. Strategic recommendations for a balanced AI data infrastructure

To foster innovation while maintaining equitable balances across the digital ecosystem, the following strategic directions are recommended:

- **preserve the integrity of Articles 3 and 4:** safeguard existing TDM exceptions against protectionist revisions that would artificially choke algorithmic development within the single market.
- **Harmonize opt-out frameworks:** address operational ambiguities under Article 4 by establishing standardized, transparent, and globally interoperable machine-readable formats for opt-outs (e.g., unified technical protocols) to lower compliance friction for both rightsholders and developers.
- **Bridge the value gap via monetization:** rather than blocking upstream data ingestion for opted-out content, fair compensation should be shifted to the commercial monetization phase. A statutory remuneration scheme, akin to compulsory licensing models, should be explored, whereby commercial AI developers contribute a statutory levy based on revenues, distributed to creators via collective management organizations.

** Expert Fellow at Techno Polis, Adj. Professor of AI Law at European University of Rome*