European Journal of Risk Regulation (2026), 1–18
doi:10.1017/err.2026.10118

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Algorithmic Crawling vs. Copyright Protection: Defining the Boundaries of Fair Use in Text and Data Mining for AI Training

Siyang Chen and Xinxuan Long

Faculty of Law, Macau University of Science and Technology, Macau
Corresponding author: Xinxuan Long; Email: 524871117@qq.com

Abstract

The advancement of AI relies on text and data mining (TDM) to acquire training data, yet its large-scale algorithmic crawling activities fundamentally conflict with copyright regimes. Through comparative legal research, this article reveals the divergent governance approaches and efficacy deficiencies within the three major jurisdictions of the European Union, the United States and China. We argue that the current regulatory model adopted by the EU may adversely affect the dynamism of Europe's AI industry. At the policy level, the EU should construct a synergistic "Cost-Benefit-Governance" framework, reducing compliance costs through differentiated regulation, collective licensing, and fiscal support measures. It should also leverage unified internal rules and international multilateral platforms to foster governance consensus. In legal practice, the EU should advance the clarification and tiered application of copyright exceptions. This can be achieved by refining rules, implementing data classification governance and innovating safe harbour liability mechanisms to enhance legal predictability, thereby balancing technological innovation and copyright protection. This framework aims to mitigate the inhibitory effect of current systems on AI innovation and address jurisdictional barriers through international cooperation.

Keywords: AI training; algorithmic crawling; copyright law; fair use; TDM

I. Introduction

The development of large-scale AI language models is fundamentally dependent on the computational analysis of vast quantities of textual and digital resources.[1] TDM serves as a pivotal technique in transforming unstructured raw information into structured data suitable for AI model training.[2] This process is giving rise to an acute and deepening tension with traditional copyright frameworks designed to protect creative works, inevitably increasing the risk of copyright infringement.[3]

[1] UMF Dimlo, V Rupesh, and Y Raju, "The Dynamics of Natural Language Processing and Text Mining under Emerging Artificial Intelligence Techniques" (2024) 15 International Journal of Systems Assurance Engineering and Management 4512.

[2] I Novalija and others, "From Code to Concept: A Semantic Approach to AI Innovation Discovery in Open Source Software Repositories" (2025) 13 IEEE Access 129995.

[3] E Bonadio, P Dinev and L McDonagh, "Can Artificial Intelligence Infringe Copyright? Some Reflections" in R Abbott (ed), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar Publishing 2022).

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

The core tension between TDM and copyright protection currently lies in the insufficiently defined boundaries of permissible use.[4] This legal ambiguity not only constrains innovation in AI development but also underscores the urgency of establishing a clearer and more adaptive legal framework. Furthermore, the territorial nature of copyright law exacerbates the legal risks associated with TDM.[5] While data resources on the internet are inherently global, standards of copyright protection and exceptions for permissible use vary significantly across different jurisdictions.[6] A TDM activity considered lawful in one jurisdiction may constitute infringement in another, immensely complicating compliant data collection and usage for AI developers operating internationally.

This paper, through a systematic comparison of the governance approaches in the three major jurisdictions of the European Union, the United States and China, reveals a core paradox: contemporary copyright systems, when confronted with the structural challenges posed by TDM, either stifle innovation by excessively restricting data access (as seen in the EU) or exacerbate legal risks due to ambiguous rules (as observed in the US and China), failing to effectively balance creators' rights with the public interest in AI technological development. The deeper root of this paradox lies in a fundamental misalignment between the traditional copyright law logic of infringement determination centred on "expressive use" and the technologically "non-expressive" nature of TDM; furthermore, the conflict between the territoriality of copyright and the globalised flow of data amplifies this systemic inadequacy. Based on this, the present study aims to address the following questions and propose solutions to the risks TDM poses to the copyright framework:

- What specific challenges does TDM present to the EU copyright system?
- What insights can the EU draw from the divergent stances adopted by various nations regarding TDM, and how should the EU respond to such risks?

Thesis and Contribution: This article posits that the EU's relatively stringent regulatory framework for TDM may, to a certain extent, constrain the innovative vitality and developmental pace of its domestic AI industry. The EU should transition towards a diversified governance model that integrates legal, technological, market and policy instruments, and implement more lenient regulatory measures for small and medium-sized enterprises. Concurrently, leveraging regulatory architectures such as the Court of Justice of the European Union (CJEU) and the AI Office, the EU can facilitate the establishment of unified behavioural guidance on TDM across its member states. The EU's regulatory approach to TDM must transcend the adversarial framework of rights-holders versus users, and instead construct a multi-stakeholder governance ecosystem that encompasses legislators, regulators, rights-holders, technology platforms and end-users.

Roadmap: The article is structured as follows. Section II clarifies the function and role of TDM in AI training. Section III discusses how the EU regulates the risks posed by TDM and the problems with this approach. Section IV compares how the EU, the US and China address the copyright challenges raised by TDM. Section V puts forward policy recommendations and solutions for building collaborative governance and improving the exceptions under EU copyright law.

[4] K Tyagi, "Copyright, Text & Data Mining and the Innovation Dimension of Generative AI" (2024) 19 Journal of Intellectual Property Law & Practice 557.

[5] F Eesteban de la Rosa and JC Fernandez, "Text and Data Mining in Cross-Border Environments: Regime in European Law, Comparative Law and in the European System of Private International Law" (2024) Revista de Derecho Comunitario Europeo <https://digibug.ugr.es/handle/10481/90940> (last accessed 12 June 2026).

[6] W Lu and W Yuzhi, "Comparative Study of TDM Technology Systems from an International Perspective" (2024 9th International Conference on Computer and Communication Systems (ICCCS), Xi'an, China, 2024).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

II. The functional positioning of TDM in AI training

TDM is a fundamental analytical method for contemporary AI systems. It refers to the process of extracting rich semantic information from texts to understand their content and meaning.[7] TDM is widely used to uncover hidden patterns and significant information from large volumes of unstructured written materials.[8] The technical workflow of TDM in AI training unfolds through several interconnected stages, which can be categorised into information extraction, information retrieval, summarisation, clustering and classification.[9] Specifically, the information extraction phase primarily involves automated algorithmic crawling, where raw digital texts are systematically acquired from online sources, databases and other repositories. This is followed by a preparatory information retrieval step, wherein the raw data – often containing advertisements, errors and irrelevant formatting code – undergoes intensive preprocessing and cleaning. This includes removing irrelevant content, correcting mistakes, handling missing values and normalising the text through techniques such as lowercasing and stemming to enhance its retrievability.[10]

The core mining process, known as feature extraction and transformation, serves subsequent summarisation, clustering and classification tasks. During this stage, the cleaned data is tokenised into words or subword units. Entity recognition techniques are employed to identify key objects (e.g., personal names, location names), and vectorisation converts these linguistic features into numerical representations. Ultimately, unstructured text is refined into a structured feature vector matrix – a process that inherently involves summarisation.[11]

The resulting structured matrix is directly input into AI models – particularly neural networks – for training. This is the phase where the model learns complex statistical relationships within the data, typically aiming to achieve effective clustering or classification.[12] Once trained, the model can be deployed for practical application and inference. Finally, an essential optimisation mechanism involves establishing a feedback loop: performance gaps identified during deployment are used to evaluate and adjust subsequent stages of the TDM process, forming a continuous cycle of improvement.[13]

III. The legislative evolution and judicial practice of EU copyright exception rules for TDM

1. The systematic construction of the EU's TDM exception clauses

The current EU legal framework governing TDM and copyright is characterised by a progression from a principled foundation towards increasingly detailed and specific

[7] M Jiang and others, "A Variety of Text Mining Technology and Tools Research", *Proceedings of the 2014 International Conference on Mechatronics, Electronic, Industrial and Control Engineering* (Published by Atlantis Press 2014).

[8] A Kaushik and S Naithani, "A Comprehensive Study of Text Mining Approach" (2016) 16 International Journal of Computer Science and Network Security 69.

[9] SS Tandel, A Jamadar and S Dudugu, "A Survey on Text Mining Techniques", (2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019).

[10] J Han, C Wang and A El-Kishky, "Bringing Structure to Text: Mining Phrases, Entities, Topics, and Hierarchies" (Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14), New York, USA, 2014).

[11] D Kavitha and A Mala, "A Brief Survey of Text Mining: Domains, Implemented Algorithms and Evaluation Metrics" (2024) 23 Journal of Information & Knowledge Management 2450078.

[12] X Zhang and D Tian, "Study on Text Classification Methods" (International Conference of China Communication, NanNing, China, 2010).

[13] E Kariri and others, "Exploring the Advancements and Future Research Directions of Artificial Neural Networks: A Text Mining Approach" (2023) 13 Applied Sciences 3186 <https://www.mdpi.com/2076-3417/13/5/3186> (last accessed 12 June 2026).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

obligations. The EU's system of copyright exceptions and limitations, founded upon Article 5 of the InfoSoc Directive, constitutes a closed and exhaustive statutory list. This reflects the civil law tradition of authors' rights, which emphasizes legal certainty. The DSMD introduced a significant update to this system by establishing a dual-track exception regime: research organisations benefit from a mandatory exception, while commercial applications are subject to a rights reservation mechanism.[14] It should be noted that the DSMD does not provide a clear definition of TDM,[15] leading some to question whether AI training activities fall within the scope of the statutory concept of TDM.[16] Subsequently, the enacted AI Act further established a risk-based regulatory framework, imposing transparency and copyright management requirements on general-purpose AI models.[17] Furthermore, the AI Act further explicitly stipulates that all providers of general-purpose AI models must comply with EU copyright law. This allows AI model training activities that meet the exception criteria to be exempt from copyright infringement liability, thereby providing essential legal certainty for AI research and development while respecting copyright protections. Moreover, the AI Act extends the regulatory scope, imposing clear requirements particularly on the training and use of general-purpose AI models. It shifts copyright compliance responsibilities earlier to the model development phase,[18] reflecting a regulatory shift from *ex post* infringement determinations to *ex ante* compliance guidance. Building on this foundation, the EU recently introduced the Code of Practice for General-Purpose AI (GPAI CoP), which provides granular interpretation of the statutory duties under the AI Act. Although this Code falls within the category of "soft law," it complements the aforementioned "hard law" instruments, collectively forming a comprehensive regulatory chain covering data usage, model development and output dissemination.

Of particular note are the relevant provisions of the AI Act, which took effect in August 2025 and are poised to substantially impact copyright disputes related to generative AI. The Act strengthens transparency obligations; while this does not alter the legal nature of potential infringements, it makes them considerably easier to identify. Model providers are now required to publish detailed summaries of their training data and implement opt-out mechanisms for rightsholders. The heightened compliance burden significantly increases operational costs for businesses.[19] All providers of general-purpose AI models must conduct rigorous due diligence on the copyright status of training data and establish mechanisms to respect rightsholders' reservations. For non-profit organisations like LAION[20] or SMEs such as HowardsHome,[21] the cost of building and maintaining a compliant system can be prohibitively high. Consequently, even though the statutory exception for TDM remains applicable in theory, the practical transparency and data provenance requirements may stifle non-profit, distributed, community-sourced data projects.

[14] 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 (2019) OJ L130/92 (Digital Single Market Copyright Directive), Arts. 3–4.

[15] T Margoni and M Kretschmer, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology" (2022) 71 GRUR International 685.

[16] TW Dornis, "The Training of Generative AI Is Not Text and Data Mining" (2025) 47 European Intellectual Property Review 65.

[17] M Maryna, "In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective" (2023) 48 European Law Review 443 <https://urn.fi/URN:NBN:fi-fe2025082791910> (last accessed 12 June 2026).

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 1689/1, Arts. 51–6.

[19] L Pinkney and others, "Voices from the Frontline: Social Work Practitioners' Perceptions of Multi-Agency Working in Adult Protection in England and Wales" (2008) 10 The Journal of Adult Protection 12.

[20] *Getty Images (Stuttgart) LLC v LAION e.V.*, Landgericht Hamburg (LG Hamburg), Beschluss vom 8. Februar 2024, Az. 310 O 115/23.

[21] *DPG Media v Howards Home Rechtbank Amsterdam*, C/13/737 170/HAZA23-690 (30 Oct 2024).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

The EU's legal provisions concerning TDM are grounded in the InfoSoc Directive. Through the interplay of the DSMD, the AI Act and the GPAI Cop, a progressively stringent and detailed regulatory framework has been formed. The InfoSoc Directive provides the foundational norms for copyright exceptions and limitations; however, it did not explicitly include TDM within the scope of such exceptions, effectively leaving room for the judicial application of copyright exceptions to TDM. Building upon this foundation, the DSMD excluded TDM conducted for non-scientific research purposes from the scope of copyright exceptions unless rightsholders had not expressly reserved their rights. This shift has created a significant dilemma for AI training for commercial purposes, introducing substantial legal risk. The AI Act does not, in essence, alter the legal position of TDM as established by the DSMD; it reiterates that providers of AI models must comply with the "opt-out" reservation of rights expressed in Article 4(3) of the Directive. Furthermore, the AI Act requires AI model providers to draft and implement a policy aimed at respecting Union copyright law, notably through the use of state-of-the-art technical means. This subjects commercial TDM to more rigorous legal oversight. For TDM, the GPAI Cop can effectively be viewed as a detailed elaboration of this technical requirement stipulated by the AI Act. Observing the trajectory of this regulatory development reveals the EU's intention to mitigate the risks TDM poses to the copyright system through increasingly stricter regulation.

Under the new framework, the central controversy is shifting. It is moving away from substantive debates over whether an act constitutes infringement, and towards uncertainties regarding procedural compliance. For instance, while the AI Act mandates that companies publish a "sufficiently detailed" summary of training data, what constitutes "sufficiently detailed" remains undefined. The GPAI CoP offers best-practice guidance rather than hard standards. Although nominally voluntary, the AI Act stipulates that providers not adhering to the GPAI CoP must demonstrate compliance through alternative means approved by the European Commission.[22] In practice, this makes adherence to the Code the lowest-risk option for companies, potentially transforming voluntary norms into *de facto* mandatory standards. Ultimately, the legal risk for businesses is not eliminated but transmuted: from the substantive risk of infringement to a procedural and technical compliance risk, centred on the adequacy of processes and disclosures, judged against ambiguous and evolving standards. This shift in risk profile confirms the academic view that relying on copyright exceptions to justify AI training faces fundamental challenges. The EU's regulatory approach effectively steers the industry toward a compliance model centered on licensing.[23]

Despite the EU's intention to establish a comprehensive regulatory regime for TDM, its practical implementation has diverged from the legislative intent.[24] The EU continues to face multiple challenges in reconciling TDM with copyright protection.[25] On one hand, ambiguities within the legal provisions, insufficient coordination among Member States and a lack of uniform standards for machine-readable formats have hindered the effective enforcement of rights reservations. On the other hand, high compliance costs pose a barrier for small and medium-sized enterprises, whilst regulatory fragmentation increases

[22] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L, Art. 56.

[23] R Eleonora, "Copyright Exceptions and Fair Use Defences for AI Training Done for 'Research' and 'Learning', or the Inescapable Licensing Horizon" (2025) 16 European Journal of Risk Regulation 961.

[24] V Novak, "The EU Copyright Exceptions for Text and Data Mining: (Un)Suitable Model Solution for the Development of Artificial Intelligence?" (2022) 33 Australian Intellectual Property Journal 92.

[25] S Mark and S Irina, "Development of the Exceptions and Limitations Regime in the EU Copyright Law on Text and Data Mining" (2020) 461 Tomsk State University Journal 249 <https://cyberleninka.ru/article/n/razvitie-rezhima-isklyucheniy-i-ogranicheniy-v-avtorskom-prave-es-v-sfere-intellektualnogo-analiza-teksta-i-dannyh> (last accessed 12 June 2026).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

the operational burden for multinational corporations. At the Member State level, divergent stances have emerged due to varying industrial interests, with a perceptible tension between the preferences of digital economy powerhouses and culturally prominent nations, further complicating policy alignment. Additionally, external pressures, including influence from the United States and industry lobbying, may lead to the delay or dilution of certain rules, exacerbating the risk of internal market fragmentation. Analysing the impact on different stakeholders, whilst rightsholders have gained new theoretical avenues for remuneration, they still encounter practical difficulties in exercising their rights. Research institutions, although permitted to use data under exception clauses, may suffer indirect losses due to constraints within the commercial ecosystem. Collectively, these factors test the EU's capacity to balance technological innovation with copyright protection. In the short term, regulatory uncertainty is likely to persist. In the long term, a failure to achieve effective coordination could adversely affect the EU's position in the global AI competition.

2. *Case studies on TDM exception disputes in the EU*

Despite maintaining a unified framework at the legislative level, the practical implementation of EU rules exhibits significant divergence among Member States, reflecting underlying coordination deficits and legal uncertainty at the Union level.[26] Although some studies suggest that TDM complies with the "three-step test,"[27] its practical application remains contested in academic discourse. German courts have demonstrated a strict, dichotomous approach, heavily emphasising the purpose of use when applying exceptions. In LAION v Robert Kneschke, the Hamburg Regional Court ruled that the creation of a dataset for non-profit research purposes fell within the TDM exception for scientific research under German copyright law.[28] Conversely, in the case brought by GEMA against OpenAI, the Munich court, citing the model's commercial nature, declined to apply the more lenient research exception. Instead, it invoked the general exception provision and found infringement due to the rightsholder's prior reservation of rights.[29]

When determining infringement, German courts consider not only the nature of the use but, more critically, its impact on the potential market for the original work. In GEMA v OpenAI, ChatGPT's output of song lyrics was deemed a direct reproduction of the work's core content, conflicting with the normal exploitation of the work and thus constituting substantial use. This judicial inclination is rooted in Germany's long-standing author's rights tradition, demonstrating a value orientation that tends to favour the protection of rightsholders' interests when balancing technological development and copyright.

An examination of rulings from German courts reveals that, based on the legal position of TDM established by the DSMD, the focus of litigation concerning TDM has shifted from the traditional question of whether an action constitutes a copyright exception to whether it serves a scientific research purpose. While this shift circumvents interpretive disputes over copyright exceptions found in traditional judicial decisions, it has led to a new institutional trap. Determining whether TDM activities in AI model training serve a scientific research purpose involves a significant degree of subjective judgement. This assessment of purpose effectively depends on whether the involved entity is non-profit

[26] M Smith, "Challenges in the Implementation of EU Law at National Level" (Briefing Requested by the JURI committee, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, November 2018) PE 608.841 <https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/608841/IPOL_BRI(2018)608841_EN.pdf> (last accessed 12 June 2026).

[27] M Senftleben, "TDM, GenAI and the Copyright Three-Step Test" (2025) 57 International Review of Intellectual Property and Competition Law 67<https://ssrn.com/abstract=5373903> (last accessed 12 June 2026).

[28] *Supra*, note 20.

[29] *GEMA v OpenAI*, 42 O 14139/24 (LG München I, 11 November 2025).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

and whether the AI in question is ultimately intended for commercial use.[30] In reality, however, the commercial nature of an AI model's training process is not strongly correlated with the commercial status of the entity undertaking it or the model's subsequent deployment for commercial purposes.

By contrast, the Dutch court in the DPG Media v. HowardsHome case exhibited a different reasoning. The Amsterdam District Court did not automatically preclude the application of the TDM exception solely based on commercial character, but instead focused on the specific characteristics of the use. The court considered the provision of short text snippets accompanied by links to the original content an act of aggregation that neither substituted nor directly competed with the source material.[31] Furthermore, the court adopted a stricter view on technical specifics, ruling that a rights reservation must precisely correspond to the specific technical means employed to be legally effective. This reflects a Dutch judicial insistence on the principle of technological neutrality, rejecting broad-brush restrictions on automated processing.

The Dutch ruling displays a distinct pragmatism, employing quantitative standards to classify short excerpts as permissible and emphasising their referential, rather than substitutive, function. This seeks a balance between investor rights and the free flow of information. Unlike the German emphasis on protecting creators, the Dutch approach appears more inclined to preserve space for data flow and technological innovation. Collectively, these cases illustrate that when reconciling AI development with copyright protection, national courts within the EU often engage in nuanced, case-specific balancing of interests, resulting in a pluralistic landscape of judicial practice.

However, this landscape of fragmented jurisprudence is at a potential turning point. The Hungarian court has referred the case of Like Company v. Google Ireland to the CJEU, a move anticipated to foster a more harmonised legal interpretation at the EU level. The core legal issues revolve around whether text processing during model training constitutes a regulated act of reproduction, and if so, whether such training can fall within the TDM exception.[32] A ruling by the CJEU on these matters will transcend the individual case, providing an authoritative definition for pivotal concepts such as "reproduction" and the "TDM exception."

The recently enacted provisions of the AI Act and the associated GPAI CoP, while not altering the legal standard for determining copyright infringement itself, are poised to exert considerable influence on judicial decisions through their transparency obligations and code of conduct mechanisms. The Act's requirement for providers of general-purpose AI models to publish detailed summaries of their training data brings greater clarity to the factual matrix of disputes.[33] Rightsholders may find it easier to demonstrate whether their works were included in a training dataset, while providers can leverage these public records to evidence their compliance measures. Should a provider have voluntarily adhered to the relevant code of practice, courts may consider this a factor demonstrating diligent effort, potentially favourably influencing liability assessments. Nevertheless, the broad definitions within certain provisions of the Act, coupled with potential variances in implementation across Member States, risk generating new divergences and further exacerbating legal uncertainty.

Tracing the evolution of relevant EU judicial practice reveals a discernible pivot from an initial encouragement of technological exploration towards a phase of intensified

[30] *Supra*, note 23.

[31] *Supra*, note 21.

[32] Case C-250/25 *Like Company v Google Ireland* [2025] OJ C 250/25 (pending)<https://eur-lex.europa.eu/eli/C/2025/3039/oj/eng> (last accessed 12 June 2026).

[33] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 1689, 12 July 2024, Art. 53(1)(d).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

regulation. Early rulings adopted a more permissive stance towards TDM, aiming to nurture an emerging technological field. As the technology matured and its market impact expanded, more recent judgements have displayed a stronger inclination to reinforce copyright protection, signalling a corrective recalibration towards upholding creators' rights. This trajectory underscores the enduring challenge for the EU in constructing its TDM framework: achieving an effective and sustainable balance between legal harmonisation, technological innovation and copyright protection.

3. Legislative motivations and policy directions behind the EU's TDM rules

The EU's policy trajectory in the realms of AI and copyright represents the outcome of protracted negotiations and compromise among multiple stakeholders.[34] In fact, EU copyright law-making has not been driven by copyright-specific rationales, but rather by the objective of establishing an internal market.[35] In the early days, online platforms extensively utilised copyrighted content under safe harbour provisions without providing fair remuneration to creators, sparking debates over an inequitable value distribution. Industry bodies, notably the Federation of European Publishers, actively lobbied for a legislative response.[36] This prompted the EU to introduce copyright filtering obligations for online service providers within the DSMD, aiming to rebalance the interests through reinforced platform accountability.

Subsequently, the drafting process of the AI Act brought these tensions to a head. Initially, the legislation demonstrated the EU's strategic intent to assert its digital sovereignty and uphold fundamental rights via high-standard regulation. By establishing what it envisioned as the world's most rigorous regulatory framework, the EU sought to leverage its rule-making power as an asymmetric competitive advantage, thereby steering the global development of AI to align with its values and market rules. However, the legislative process encountered substantial resistance from industry. US tech firms, led by OpenAI, openly questioned the feasibility of certain provisions,[37] at one point threatening to exit the European market,[38] while engaging in intensive lobbying efforts. Simultaneously, divisions emerged within the EU itself, with Member States such as France and Germany expressing concerns that excessive regulation could stifle the growth of domestic AI enterprises like Mistral AI.[39]

Under this dual pressure, legislators ultimately made significant concessions.[40] The final version of the Act abandoned the initial concept of automatically categorising all General-Purpose AI systems as high-risk, opting instead for a new regulatory tier that

[34] JP Quintais, "The New Copyright Directive: A tour d'horizon – Part I" (Kluwer Copyright Blog, 17 May 2019) <http://copyrightblog.kluweriplaw.com/2019/05/17/the-new-copyright-directive-a-tour-dhorizon-part-i/> (last accessed 12 June 2026).

[35] A Trapova, "Copyright for AI-Generated Works: A Task for the Internal Market?" (2023) 48 European Law Review 187.

[36] A Bergman-Tahon, "EU Copyright Modernised: Was it Easy?" (International Publishers Association, 17 April 2019) <https://internationalpublishers.org/eu-copyright-modernised-was-it-easy/> (last accessed 12 June 2026).

[37] TOI Tech Desk, "Germany's Two Biggest Technology Companies Are Unhappy with EU's AI Regulations; Call It "Toxic" For ..." (The Times of India15 July 2025) <https://timesofindia.indiatimes.com/technology/tech-news/germanys-two-biggest-technology-companies-are-unhappy-with-eus-ai-regulations-call-it-toxic-for-/articleshow/122504598.cms> (last accessed 12 June 2026).

[38] Reuters, "OpenAI may leave the EU if regulations bite - CEO" (Reuters, 25 May 2023) <https://www.reuters.com/technology/openai-may-leave-eu-if-regulations-bite-ceo-2023-05-24/> (last accessed 12 June 2026).

[39] J Rone, "The Sovereign Cloud" in Europe: Diverging Nation State Preferences and Disputed Institutional Competences in the Context of Limited Technological Capabilities" (2024) 31 Journal of European Public Policy 2343.

[40] F Palmiotto, "The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation" (2025) 16 European Journal of Risk Regulation 770.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

imposes stricter obligations solely on the most powerful models presenting systemic risks. The evolving stance of OpenAI exemplifies a broader pragmatic shift in corporate attitude. Its approach transitioned from highly publicised threats of withdrawal in 2023 to stating, following an unsuccessful legal challenge in 2025, that it was considering an appeal.[41] This indicates a growing focus on navigating operational space within the compliance framework, rather than outright confrontation. This evolution signifies that the relationship between regulation and innovation has entered a new stage, where the emphasis pivots from adversarial debates on principle prior to legislation to the practicalities of compliance adaptation and dispute settlement thereafter.

Transparency requirements, championed by rights holders, have also been incorporated into the legislation, mandating that generative AI systems disclose copyright information related to their training data. This directly addresses concerns raised by the creative industries. The final version of the Act emerges as a hybrid outcome of competing interests, where various demands have been partially accommodated. While the technology sector faces compliance pressures, it has also averted more extreme regulatory measures. However, it is worth noting that stringent regulations centred on building data barriers could inadvertently increase compliance costs for domestic companies, reducing the diversity and quality of AI models[42] and ultimately constraining the long-term competitiveness of the EU's AI industry.

Moving forward, businesses will need to navigate dual compliance requirements under both the DSMD and the AI Act. Such complexity may gradually lead the industry to develop best practices. As more companies adopt codes of conduct and national regulatory bodies within Member States become more established, the EU's approach to AI governance could influence the formation of global standards through the Brussels Effect. Nevertheless, the divergence between the EU's strict regulatory path and the US principle of fair use remains a significant obstacle to international regulatory harmonisation.

Recently, the European Commission has considered suspending certain provisions, including granting grace periods to companies and postponing transparency-related penalties.[43] This reflects an ongoing trend of dynamic policy recalibration. Such moves respond not only to lobbying by US tech giants but also to internal calls from European industries seeking to avoid stifling innovation. These developments illustrate regulators' efforts to strike a balance between external pressures and domestic demands, while also signalling a reassessment of the potential adverse effects of the initially stringent regulatory framework.

IV. A comparative legal examination of TDM exceptions in major non-EU jurisdictions

1. The "transformative use" standard in US law and its application

In addressing the copyright challenges posed by TDM, the judicial practice in the United States exhibits a distinct pragmatic inclination, deeply rooted in its fair use doctrine. Originating from the common law tradition of judicial precedent, this doctrine was codified in Section 107 of the Copyright Act of 1976.[44] It is not a specific catalogue of rights

[41] M Klos, "Open AI must pay GEMA licence fee for ChatGPT" (JUVE Patent, 11 November 2025) <https://www.juve-patent.com/cases/open-ai-must-pay-gema-licence-fee-for-chatgpt/> (last accessed 12 June 2026).

[42] M Kretschmer, T Margoni and P Oruç, "Copyright Law and the Lifecycle of Machine Learning Models" (2024) 55 IIC – International Review of Intellectual Property and Competition Law 110.

[43] B Moens and T Bradshaw, "European CEOs urge Brussels to halt landmark AI Act" (Financial Times, 3 July 2025) <https://www.ft.com/content/a825759e-aec8-4184-bc73-f604f169204c> (last accessed 12 June 2026).

[44] BA Greenberg, "Copyright Trolls and Presumptively Fair Uses" (2014) 85 University of Colorado Law Review 54, 91.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

but an open-ended principle of judicial equity. At its core, judges are required to conduct a comprehensive balancing of four flexible statutory factors on a case-by-case basis: the purpose and character of the use, the nature of the copyrighted work, the amount and substantiality of the portion used and the effect of the use upon the potential market for or value of the work. This allows for certain unlicensed, yet socially beneficial uses, such as criticism, comment, news reporting, teaching, scholarship or research.

Compared to the EU's exhaustive list of copyright exceptions, although both share the legislative objective of balancing the interests of rightsholders with public interests like freedom of expression, education and innovation – and both must comply with the internationally accepted three-step test – their legal logics and implementation paths are fundamentally different. The US fair use doctrine is a flexible mechanism combining rules and standards. It is inherently common-law in nature, serving as a safety valve grounded in utilitarian philosophy and the First Amendment. It is capable of proactively adapting to technological shifts – from videocassette recorders and search engines to AI training – embodying both forward-looking potential and inherent uncertainty. In contrast, the EU's list of exceptions is a product of the civil law *droit d'auteur* tradition, manifested as a closed, exhaustive enumeration within statutory law. User conduct must strictly fit within a specific provision to qualify for an exemption. While offering greater legal certainty, it lacks the inherent flexibility to respond to novel developments.

This fundamental paradigmatic divide between common law judicial dynamism and civil law legislative rigidity is vividly illustrated in US judicial practice. Specifically, courts, by flexibly interpreting the transformative use standard within the fair use doctrine, have regarded AI training as a transformative use of the original works.[45] The purpose is not merely to reproduce the content of the originals, but to learn their underlying patterns through algorithms to generate entirely new functionalities and value. This approach has, in effect, fostered a relatively permissive innovation environment for the AI industry. This judicial stance not only reflects the law's adaptability to technological developments but also, on a deeper level, serves the strategic objective of enhancing the nation's technological competitiveness.

Several landmark cases clearly illustrate the conditional support adopted by US courts regarding copyright issues in AI training data. In The New York Times Company v OpenAI Inc, the dispute extended beyond the training phase to the model's output behaviour, indicating that the assessment of transformativeness requires a holistic consideration of both the technical process and the consequences of use.[46] The ruling in Richard Kadrey v Meta Platforms Inc was more groundbreaking, as the court explicitly recognised that the training process itself constitutes highly transformative use, given its aim is to create tools capable of text comprehension and generation, rather than to disseminate the original expression.[47] This judgment provided significant judicial support for technology companies using publicly available data for model training.

This judicial stance centered on transformative use becomes particularly salient when contrasted with the UK's approach to similar disputes. In the cross-border parallel litigation of Getty Images v Stability AI, the legal paths of the two nations diverge markedly: on the US front, the core of the litigation revolves closely around the principle of transformative use; whereas in the UK, due to the strict technical definition of infringing copies under the law and jurisdictional hurdles, the focus of the lawsuit has been shifted towards trademark infringement, with the court having ruled that the AI model itself does not constitute a copyright-infringing copy.[48] This contrast highlights the unique defensive

[45] Q Jenny, "Does Training AI Violate Copyright Law?" (2021) 36 Berkeley Technology Law Journal 1407.
[46] *The New York Times Company v OpenAI Inc*, No 1:23-cv-11195 (SDNY 27 December 2023).
[47] *Richard Kadrey v Meta Platforms Inc*, No 23-cv-03417, 2025 WL 1752484 (ND Cal 25 June 2025).
[48] *Getty Images (US) Inc v Stability AI Ltd* [2025] EWHC 2863 (Ch).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

space that the flexibility of the US fair use doctrine provides to AI companies, while the UK's more traditional, text-based interpretation may impose more stringent evidentiary challenges on rights holders.

Of course, the US support for transformative use is not unconditional. The ruling in Kristen Bartz v Anthropic PBC introduced a crucial limitation, with the judge unequivocally stating that the legality of the data source is a prerequisite for claiming transformative use. Even if the training purpose is innovative, obtaining data through pirated channels inherently constitutes infringement.[49] This position has prompted the industry to shift towards compliant data sources and spurred the emergence of a new data licensing market. Notably, companies of different sizes face disparate situations in this transition: well-resourced large firms can mitigate risks through bulk licensing agreements, whereas startups may struggle due to compliance barriers.

This judicial direction effectively constitutes an implicit industrial policy. Through a relatively permissive interpretation of transformativeness, US courts have created a regulatory environment for domestic AI enterprises that is more favourable than that faced by their EU counterparts. This competitive strategy is particularly evident in cases like the lawsuit by Westlaw's parent company against the startup Ross Intelligence, where the court's stringent ruling protected the interests of established content giants while inadvertently raising barriers to market entry.[50] Furthermore, the U.S. Copyright Office has established the disclosure requirements for AI-generated content in copyright registration at the administrative level by issuing the Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence. At the same time, through the release of the Copyright and Artificial Intelligence series of reports, it has guided and stimulated deeper discussions on issues such as the legality of training data. The legality of training data sources is currently being addressed by Congress through legislative proposals like the Generative AI Copyright Disclosure Act and has not yet been fully or definitively clarified through administrative guidance.

The United States has currently developed a unique governance ecosystem: statutory legislation remains silent, the judiciary exerts influence through the accumulation of case law and de facto standards emerge commercially through market negotiations. This pluralistic dialogue mechanism maintains the legal system's flexibility but also introduces significant uncertainty. Licensing negotiations between large technology firms and content providers are shaping new industry norms, with judicial rulings either endorsing or correcting these developments. This dynamic equilibrium continuously moulds the industrial landscape, fostering technological progress while potentially raising future antitrust concerns or triggering legislative intervention.

In summary, to consolidate and maintain its global leadership in AI, the US has adopted a strategy of flexibly interpreting the transformative use doctrine in its judicial practice. This approach establishes a resilient balance between copyright protection and technological innovation,[51] effectively creating a favourable innovation climate for domestic companies. This mechanism provides the necessary legal latitude for AI development while steering the industry towards compliance through market forces. Its fundamental rationale is to proactively serve the sustained development needs of strategic national industries through the adaptive interpretation of legal principles.

[49] *Kristen Bartz v Anthropic PBC*, No 23-cv-04637, 2025 WL 1752483 (ND Cal 25 June 2025).

[50] *Thomson Reuters Enterprise Centre GmbH v Ross Intelligence Inc*, No 1:20-cv-00613, 2025 WL 1752482 (D Del, 11 February 2025).

[51] M Sag, "The New Legal Landscape for Text Mining and Machine Learning" (2018) 66 Journal of the Copyright Society of the USA 291.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

2. An examination of the status quo of the fair use system in China's copyright law

China's legal regulations and judicial practice in the field of TDM reflect profound policy considerations. As competition intensifies between Chinese AI enterprises and their international rivals, particularly leading US companies, legislators must carefully balance the relationship between copyright protection and technological innovation to avoid falling behind in the global race. China's approach to applying its fair use rules to copyright issues arising from TDM demonstrates distinct pragmatism.[52] Presently, unlike the EU, China has not established a specific copyright exception for TDM. Legal determinations primarily rely on the general provisions for fair use within the Copyright Law,[53] combined with a comprehensive assessment utilising legal instruments such as the Anti-Unfair Competition Law. This model reflects a pragmatic strategy that prioritises regulating the output end while adopting a relatively tolerant stance towards the input end. It aims to maintain market order and content security while preserving a degree of exploratory space for the development of the AI industry.

Several representative cases clearly illustrate the reasoning and standards applied by Chinese judicial practice in handling related controversies. China's first copyright infringement case involving the training phase of a large-scale AI image generation model directly addresses the core issue of data acquisition for AI training. An illustrator accused the developer and operator of an AI painting software of using their publicly posted works to train the AI model without authorisation, pushing the judicial review focus forward to the input stage of training data. The core dispute lies in whether this act of reproduction for training purposes constitutes infringement.[54] The plaintiff's choice to sue directly over the training act itself, rather than acting only after infringing content is generated, indicates that content creators are beginning to actively assert their rights in the AI era. The Chinese court's characterisation of the nature of this training activity in this case will directly influence the entire industry's development model.

In the case concerning the Shuabao app crawling short videos from Tiktok, the Chinese court demonstrated its focus on maintaining fair competition order during the data processing and utilisation phases. From the perspective of corporate strategy, Tiktok's decision to base its lawsuit on unfair competition rather than copyright infringement highlights the platform's intent to protect the aggregate value of its data collection. The court ruled that the crawling constituted unfair competition, providing an alternative protective path for data-related interests through competition law.[55] This approach sidesteps the debate over whether a data collection constitutes a copyrightable work while effectively curbing free-riding behaviour, reflecting a pragmatic judicial effort to uphold market fairness.

The world's first ruling on the tort liability of a generative AI platform, issued by the Guangzhou Internet Court, exemplifies strict regulation of output-side conduct. In this case, the court held the AI platform operator liable for users generating images

[52] J Hua, "Copyright Exceptions for Text and Data Mining in China: Inspiration from Transformative Use" (2022) 69 Journal of the Copyright Society of the USA 123.

[53] Copyright Law of the People's Republic of China (2020 Amendment), s 24(1), (2), (6), (13) (China) Article 24 of the law stipulates over ten specific scenarios in an illustrative manner, such as limited use for personal study or classroom teaching. However, AI training activities are typically initiated by commercial entities, and the scale and purpose of their data copying often fall outside the scope of these conventional exceptions. Consequently, judicial practice frequently requires a creative interpretation and application of the catch-all provision in Item 13 of the same article – "other circumstances prescribed by laws or administrative regulations" – which affords courts significant discretionary space in delineating the boundaries of such activities.

[54] Case of *Four Illustrators v Rednote & Trik AI* (Beijing Internet Court, 20 June 2024) (case number not yet publicly available) <https://www.thepaper.cn/newsDetail_forward_30355990> (last accessed 12 June 2026).

[55] *Beijing Weibo Shijue Technology Co Ltd v Beijing Chuangrui Wenhua Media Co Ltd* (2021) Jing 73 Min Zhong No 1011 (Beijing Intellectual Property Court, 2021).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

substantially similar to the "Ultraman" character using its service. The liability was based on the operator's failure to implement necessary proactive screening measures and to fulfil a reasonable duty of care, constituting contributory infringement.[56] Notably, the court supported the plaintiff's claim to cease the specific infringing acts in question but rejected the request to delete all related training data from the platform.[57] This indirectly indicates that the court distinguishes between data use in the training phase and infringing outcomes at the output stage, leaving interpretive space for training acts – not aimed at reproducing the expressive content of a work and not unreasonably prejudicing the legitimate interests of the right holder – to potentially be recognised as fair use.

The role of China's Copyright Law in the field of TDM can also be observed in disputes that did not proceed to litigation. For example, China National Knowledge Infrastructure (CNKI) once sent a letter to Shanghai Mita Network Technology Co., Ltd., asserting that its copyright in compilations had been infringed, alleging that Mita AI's search engine had crawled and displayed titles and abstracts of a large volume of academic literature without authorisation.[58] The dispute was ultimately not taken to court but was resolved swiftly within 24 hours of Mita receiving the letter, through a public statement, removal of the relevant data and disconnection of links. The root cause of this outcome lies in the significant ambiguities within China's Copyright Law regarding such novel data utilisation disputes, making it difficult for either party to ensure victory in court. For CNKI, the basis of its claim rests on treating the collection of numerous literature titles and abstracts as a protected compilation. However, this is challenging in judicial practice. Chinese courts have increasingly stringent requirements regarding the originality of selection and arrangement in databases. CNKI's database emphasises comprehensiveness of coverage over creative arrangement, creating substantial uncertainty about whether it would be recognised as possessing sufficient originality. For Mita, the potential defence of fair use also carries high risks. Although its actions had a public-interest nature in facilitating academic research, under China's Copyright Law, the key criteria for determining fair use include that the act "does not unreasonably prejudice the legitimate interests of the copyright owner" and is consistent with the "three-step test." Faced with commercial, large-scale and systematic data crawling and substitutive display, these conditions are difficult to satisfy, and a court would likely find that such actions substantially harm the market value of CNKI's database. Therefore, the ambiguity of the law creates a grey area of comparable risk for both parties. Litigation is not only time-consuming and costly but could also produce precedents unfavourable to one's own side or the entire industry. Through a commercial warning and swift settlement, CNKI maintained its data barriers at a relatively low cost, while Mita averted existential threats and shifted to alternative data sources.

A review of these cases and regulatory practices reveals that China has adopted a multi-level, dynamically coordinated governance strategy in the field of TDM. At the top-level design level, regulations such as the Interim Measures for the Management of Generative Artificial Intelligence Services[59] emphasise the responsibilities of service providers regarding content security, labelling obligations and copyright compliance, demonstrating

[56] *Shanghai Xinchuanghua Cultural Development Co Ltd v Guangzhou Tab Network Technology Co Ltd* (2024) Yue 0192 Min Chu No 113 (Guangzhou Internet Court, 8 February 2024).

[57] Y Wan and H Lu, "Copyright Protection for AI-Generated Outputs: The Experience from China" (2021) 42 Computer Law & Security Review 105581.

[58] "CNKI Sends Infringement Notice to Meta AI Search Website" (The Paper, 16 August 2024) <https://m.thepaper.cn/newsDetail_forward_28421385> (last accessed 12 June 2026).

[59] Interim Measures for the Management of Generative Artificial Intelligence Services (Joint Order of Seven Departments Including the Cyberspace Administration of China No. 15) (2023) <https://www.gov.cn/gongbao/2023/issue_10666/202308/content_6900864.html> (last accessed 12 June 2026) – Regulates security assessments, data governance, and content controls for generative AI services.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

a focus on regulating the output and application ends. The Guidance on Deepening the Implementation of the "AI Plus" Initiative further proposes improving data property rights and copyright regimes to align with AI development, and encourages the exploration of value-contribution-based data cost compensation and revenue sharing mechanisms.[60] At the judicial level, courts exhibit caution, exploring boundaries on a case-by-case basis. The core consideration often centres on whether the use affects the normal exploitation of the work and unreasonably prejudices the legitimate interests of the right holder. At the enforcement level, initiatives such as the "Clear and Bright Governance of AI Technology Misuse" campaign led by the Cyberspace Administration of China exemplify a proactive regulatory approach. By reinforcing primary responsibility on enterprises and intervening at the source of training data management, authorities have urged companies to clean existing datasets and establish strict web-crawling standards, thereby translating principle-based requirements into concrete compliance guidance. Such administrative oversight also demonstrates flexibility and local adaptability: cyberspace authorities in regions such as Beijing, Shanghai and Zhejiang have adopted measures tailored to local industrial contexts – including multi-stakeholder governance, awareness-raising campaigns and technical blocking mechanisms – to jointly promote lawful and compliant TDM. This "output-focused, input-lenient" governance model, under the current circumstance of unclear copyright rules, objectively affords the AI industry a degree of flexibility and grey area regarding data acquisition and model training. It reflects a characteristically pragmatic wisdom of "crossing the river by feeling the stones."

3. Comparison and analysis of the EU, US and Chinese models

In addressing the copyright challenges posed by TDM, the United States, the European Union and China have developed distinct governance pathways shaped by their respective policy priorities. These approaches not only reflect differences in regulatory design across the three major legal systems but also reveal deeper tensions among competing objectives such as technological innovation, cultural protection, and industrial security.

The EU's legal and policy framework for TDM, centred on the DSMD and the AI Act, aims to balance technological innovation with copyright protection.[61] However, its underlying logic is marked by significant internal challenges. The "opt-out" mechanism in the DSMD, which styles itself as an exception yet is effectively regarded as a right that rightsholders must actively reserve to exclude material from training datasets,[62] and which allows rightsholders to exercise their rights at any time, results in a dynamic copyright status for high-value data. Moreover, the opt-out mechanism, reliant on non-standardised signals, exacerbates uncertainty and fails to ensure transparency or fair compensation.[63] This uncertainty makes it difficult for businesses to ensure long-term compliance, thereby dampening incentives for innovation[64] and potentially placing European AI developers at a

[60] State Council, "Opinions on Deepening the Implementation of the 'AI+' Action" (Guofa [2025] No. 11, 21 August 2025) <https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html> (last accessed 12 June 2026).

[61] A Guadamuz, "The EU's Artificial Intelligence Act and Copyright" (2024) 28 The Journal of World Intellectual Property 213.

[62] B Sobel, "A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning", in J-A Lee, R M Hilty and K-C Liu (eds), *Artificial Intelligence and Intellectual Property* (Oxford 2021).

[63] E Alonso and N Lucchi, "AI and Copyright "Hallucinations": Does the Text and Data Mining Exception Really Support Generative AI Training?" (2025) 47 European Intellectual Property Review 515.

[64] A Pereira, "Copyright in Text and Data Mining and the Development of Artificial Intelligence in Europe" (2023) 66 Boletim De Ciencias Economicas 2535.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

competitive disadvantage globally.[65] Concurrently, the AI Act's requirement to publish summaries of training data to enhance transparency may conflict with the protection of corporate trade secrets. Technically, the accurate, real-time identification of copyright status within vast datasets is both costly and formidable.[66] Supplementary voluntary codes of practice, due to their non-binding nature and reliance on widespread industry adoption, offer limited practical coordination. The layering of these rules also risks fragmenting the internal market, as divergent transposition by Member States increases legal complexity.

In the international context, the EU's path contrasts sharply with the US and Chinese models. The US leverages its fair use doctrine, allowing courts to address TDM flexibly through case law. This approach, focused on promoting innovation and maintaining global technological leadership, fosters a relatively permissive environment for the AI industry.[67] However, its reliance on *ex-post* judicial rulings creates legal uncertainty, and high litigation costs can stifle startups. The absence of a prior exemption mechanism and the practical impossibility of licensing vast datasets on a per-work basis leave businesses in a persistent state of legal ambiguity – a significant drawback the EU should avoid emulating. Nevertheless, the adaptive nature of the US fair use model is worthy of consideration, as it allows courts to adjust standards dynamically with technological progress, avoiding legislative rigidity and providing flexibility for innovation.

China adopts a pragmatic regulatory model, foregoing a specific TDM exception in favour of a comprehensive assessment utilising the general fair use clause in its Copyright Law alongside instruments like the Anti-Unfair Competition Law. This strategy focuses regulatory effort on the output end, allowing flexibility for copying activities at the input stage.[68] While this provides the industry with operational leeway during its developmental phase, the ambiguity surrounding input activities means businesses may operate based on regulatory tolerance rather than legal certainty, potentially affecting long-term investment confidence. This inherent vagueness is a disadvantage the EU should not replicate, as it can lead to protracted instability. However, China's emphasis on coordinating development with controllability offers a valuable insight for the EU. For instance, its approach of managing risks by reinforcing service provider obligations, rather than excessively restricting data input, can help foster technological iteration while safeguarding rights.

The future trajectory of the law remains uncertain. The EU may undertake de facto flexible adjustments through soft-law instruments like guidelines and promoted codes of practice, providing the industry with breathing room. The definitive interpretation of key legal concepts will likely fall to the CJEU, though this process is protracted and its outcomes unpredictable. Ultimately, EU policy will likely continue to oscillate between upholding rights protection standards and striving for AI leadership, suggesting that instability in its legal environment may persist in the short to medium term.[69]

[65] GM Ziaja, "The Text and Data Mining Opt-out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European Artificial Intelligence Innovations?" (2024) 19 Journal of Intellectual Property Law & Practice 453.

[66] A Buick, "Copyright and AI Training Data – Transparency to the Rescue?" (2024) 20 Journal of Intellectual Property Law & Practice 182 <https://doi.org/10.1093/jiplp/jpae102> (last accessed 12 June 2026).

[67] MA Lemley and B Casey, "Fair Learning" (2020) 99 Texas Law Review 743.

[68] J Hua, "Copyright Exceptions for Text and Data Mining in China: A Flexible Approach" (2022) 69 Journal of the Copyright Society of the USA 123.

[69] C Handke, L Guibault and J-J Vallbé, "Is Europe Falling behind in Data Mining? Copyright's Impact on Data Mining in Academic Research", in B Schmidt and M Dobreva (eds), *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust* (Amsterdam, IOS Press 2015).

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

V. Reshaping the boundaries of TDM exceptions: exploring pathways from a collaborative governance perspective

1. Policy leadership: fostering a multi-stakeholder copyright governance ecosystem

To address the predicament faced by TDM in the application of copyright law, the EU must construct a multi-level, multi-institution collaborative governance framework. The core objective is to create a stable and affordable compliance environment for businesses, especially small and medium-sized enterprises (SMEs).[70] Specifically, at the regulatory level, the European Commission should clarify exemptions based on turnover in the implementing provisions of the DSMD, allowing small businesses to temporarily avoid some of the strictest filtering obligations. To reduce transaction costs, the European Union Intellectual Property Office (EUIPO) could collaborate with collective management organisations in Member States to develop and promote standardised, simplified licensing agreement templates. Regarding financial support, programmes such as the Digital Europe Programme or dedicated AI innovation funds could be leveraged to provide direct subsidies for SMEs' compliance expenditures. Additionally, exploring the inclusion of certain compliance costs within tax deduction schemes would substantially reduce the financial burden on enterprises from a fiscal perspective. While reducing costs for businesses, balancing the interests between rights holders and users is equally crucial. The European Commission could promote the extended collective licensing model through legislative initiatives to safeguard the actual income of rights holders. Moreover, utilising the EU Blockchain Services Infrastructure to develop a prototype micro-payment system based on distributed ledger technology could enable real-time rights verification and remuneration payment. Furthermore, the European Committee for Standardisation (CEN) should take the lead in advancing the standardisation of copyright information identification technologies.

Coordination at the governance level is also indispensable. Within the EU, binding implementation guidelines could be issued through regulatory structures established under the AI Act, such as the AI Office, to limit Member States' discretion when transposing TDM exceptions into national law. The EUIPO could also establish and maintain a unified EU TDM rules database, clearly presenting regulatory differences among Member States and enhancing the transparency of legal application. At the international level, the EU should proactively utilise multilateral platforms such as the EU–US Trade and Technology Council (TTC) to explain the rationale behind its high-standard TDM rules and actively promote their external adoption. Simultaneously, the EU should advocate its balanced governance model in international forums like the World Intellectual Property Organization (WIPO) to foster global consensus. Additionally, active efforts should be made through European Committee for Standardization (CEN) and the International Organization for Standardization (ISO) to advance the global standardisation of copyright information identification technologies, thereby reducing cross-border data flow barriers arising from regulatory differences at the infrastructure level.

In summary, by comprehensively employing policy instruments such as differentiated regulation, economic incentives, governance coordination, and international dialogue, the EU has the potential to systematically resolve the copyright exception dilemma surrounding TDM, ultimately achieving a harmonious interaction between copyright protection and technological innovation.

[70] N Jondet, "The Text and Data Mining Exception in the Proposal for a Directive on Copyright: Why the European Union Needs to Go Further than the Laws of Member States" (2018) 67 Propriétés Intellectuelles 25.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

2. Legal practice: recommendations for legislative refinement and judicial guidelines on TDM exceptions

To resolve the dilemma of applying copyright law to TDM, the construction of a theoretical framework must be grounded in transforming abstract legal principles into clear, actionable and specific rules. This systematic endeavour can be advanced along three closely interrelated dimensions:

First, efforts should focus on clarifying legal provisions to eliminate ambiguities. Specific measures could include the European Commission promoting standardised licensing request templates or endorsing industry-recognised copyright filtering tools to enhance efficiency. On a technical level, it is essential to unify technical standards for machine-readable formats, requiring that opt-out declarations adhere to specified data structures and are included in relevant EU standards catalogues. Existing AI model documentation forms could also be extended to serve as universal copyright declaration templates. Concurrently, a centralised copyright declarations portal should be established, guiding rights holders to use standardised language for submissions and ensuring interoperability with major technology platforms. Furthermore, support should be given to developing user-friendly data provenance tools to lower the technical barriers for transparency disclosures. To assist SMEs, a three-year transition period could be instituted, during which only general disclosures would be required, utilising funding from programmes such as "Digital Europe" to help them procure compliance tools. Judicially, the CJEU should be encouraged to clarify key concepts through preliminary rulings, drawing on Member State jurisprudence regarding the use of work excerpts to set explicit quantitative or proportional thresholds, thereby enhancing the practicality of judicial guidance.

Once the foundational rules are clarified, the governance approach should shift from a one-size-fits-all model to a tiered system based on the value-density of data. While the current legal framework primarily classifies and manages risks associated with AI applications, its key provisions do not genuinely follow a risk-based approach,[71] overlooking distinctions in the attributes of training data itself. For public domain and open-licence content, free circulation should be guaranteed. For vast quantities of low-originality, general data, statutory exceptions or implied licensing mechanisms could be introduced to reduce compliance costs. Regarding core data with high commercial value, carefully designed licensing and compensation mechanisms are necessary to protect rights holders' interests, accompanied by stringent antitrust scrutiny to prevent data monopolisation. Such tiered governance can alleviate difficulties in accessing core data while achieving a more resilient balance between innovation efficiency and rights protection.

Regarding liability allocation, there is a need for innovative application of safe harbour principles. AI service providers should bear a reasonable duty of care concerning model outputs, including establishing content filtering mechanisms and infringement complaint channels. If a platform has fulfilled these obligations and responds promptly, liability limitations could apply to its use of training data and user-generated content. As a counterbalance, platforms should proactively disclose the basic types and source characteristics of their training data. This not only fulfills transparency obligations and builds industry trust, but also provides practical leads for rightsholders to safeguard their rights. At the same time, the implementation of this obligation must be tailored to specific contexts, maintaining a dynamic balance between technical costs and practical utility, ensuring that transparency requirements are both principled and reasonable as well as

[71] M Ebers, "Truly Risk-Based Regulation of Artificial Intelligence How to Implement the EU's AI Act" (2025) 16 European Journal of Risk Regulation 684.

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press

actionable.[72] Through these concrete pathways, a sustainable equilibrium incentivising technological innovation while protecting legitimate rights can be constructed.

VI. Conclusion

With the implementation of the EU's AI Act and its accompanying measures, the regulation of TDM may enter a new phase. However, the inhibitory effect of the EU's prolonged stringent regulation on the AI industry may not be reversible in the short term; indeed, the enhanced regulatory force of the AI Act could lead to further suppression.

The core argument of this paper is that the EU's stringent regulation of TDM may stifle innovation, necessitating a shift towards a multi-stakeholder governance model. It also calls for the establishment of clear criteria for copyright exceptions pertaining to TDM – particularly standards for assessing non-commercial purposes and technological means that duly respect copyright.

Looking ahead, there are preliminary signs of a shift in the EU's approach to AI regulation. In November 2025, the European Commission adopted the proposal for a Digital Omnibus Regulation, representing a systemic endeavour to foster innovation and growth in the digital sector through the simplification of rules governing cybersecurity, data privacy, data access and platform governance.[73] This suggests that the core mission of future research may lie in constructing a regulatory framework within the irreversible march of technology that both protects creativity and unleashes AI's potential. The stratified governance model proposed in this article constitutes a systematic endeavour towards that objective.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of use of artificial intelligence. During the preparation of this work the authors used ChatGPT5 for checking correctness and grammar, as well as for identifying additional sources. The authors reviewed and edited the content and they take full responsibility for the content of the publication.

[72] MC Buiten, "Towards Intelligent Regulation of Artificial Intelligence" (2019) 10 European Journal of Risk Regulation 41.

[73] Commission, "Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus)" COM (2025) 837 final.

Cite this article: S Chen and X Long, "Algorithmic Crawling vs. Copyright Protection: Defining the Boundaries of Fair Use in Text and Data Mining for AI Training". *European Journal of Risk Regulation*. https://doi.org/10.1017/err.2026.10118

https://doi.org/10.1017/err.2026.10118 Published online by Cambridge University Press