COMPARATIVE ANALYSIS OF TEXT AND DATA-MINING EXCEPTION CLAUSES IN SOUTH KOREA, SINGAPORE, AND JAPAN

This article explores whether generative AI image models, which employ latent-space diffusion techniques, infringe upon the copyright of works used as training data. A doctrinal analysis based on US principles of transformative use and substantial similarity observed that the core technical process of latent-space generation does not typically replicate copyrighted expression. A comparative legislative analysis of text and data-mining ("TDM") exceptions in Japan, Singapore, and the European Union highlights their growing importance in facilitating data-driven innovation. The findings support adopting a dedicated TDM exception in South Korea to provide legal certainty and promote AI innovation.

WooJung **JON**
DPhil in Law (University of Oxford);
Associate Professor, Korea Advanced Institute of Science and Technology (KAIST), Graduate School of Future Strategy.

I. Introduction

A. Background and purpose

1 Generative artificial intelligence ("AI") technologies, including diffusion-based text-to-image systems and transformer-driven architectures, have achieved a level of creative output that challenges traditional copyright doctrines.[1] In simple terms, these models "learn" from vast datasets, internalising abstract patterns that can then be recombined to produce new content. While the "Italian Plumber"[2] anecdote highlights a scenario in which an AI system might inadvertently regurgitate original data, the normative question remains:

1 Robin Rombach *et al*, "High-Resolution Image Synthesis with Latent Diffusion Models" (2022) *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674.

2 Timothy B Lee & James Grimmelmann, "Why the New York Times Might Win Its Copyright Lawsuit Against OpenAI", *Ars Technica* (20 February 2024) <https://arstechnica.com/tech-policy/2024/02/why-the-new-york-times-might-win-its-copyright-lawsuit-against-openai/#page-3> (accessed 16 July 2025).

Does typical AI-driven image generation truly infringe upon copyrighted works used in data training?

2 Several high-profile lawsuits underscore the significance of this issue. *Getty Images (US) Inc v Stability AI Inc*[3] ("*Getty Images*") raises the question of whether ingesting protected photographs to train a commercial image generator constitutes reproduction or derivative use. *Andersen v Stability AI Ltd*[4] is a proposed class action by artists concerned about losing control over their creative labour. While the court dismissed some claims in October 2023, the plaintiffs filed amended complaints, and the case has survived subsequent dismissal motions on its core claims. As of late 2025, the case is actively in discovery and is scheduled for a jury trial beginning in September 2026. These cases illustrate the intersection of cutting-edge AI with centuries-old copyright precepts – particularly reproduction rights, derivative works, and fair use defences. Given the speed of AI innovation and the relative slowness of legislative processes, courts and commentators must grapple with whether existing legal tools sufficiently address these novel controversies.

3 Generative models generally fail to produce outputs that meet the legal threshold for copyright infringement, particularly under the crucial doctrines of transformative use and substantial similarity, as their outputs tend to recombine abstract features from multiple sources rather than replicate protectable expression *verbatim*. Nonetheless, this article acknowledges the ongoing debate and proposes policy-oriented solutions where uncertainties remain. By situating AI generation within an established legal framework, it aims to dispel the presumption that "scraping + generation = infringement" while also highlighting legitimate concerns about memorisation or direct duplication in special instances.

4 The rapid development of AI and big data analytics has revolutionised many industries, emphasising the need for adaptable legal frameworks that facilitate text and data mining ("TDM").[5] Text and

3 The original Delaware case, No 1:23-cv-00135, was voluntarily dismissed. The lawsuit was refiled in the Northern District of California on 14 August 2025 (No 3:25-cv-06891).

4 No 3:23-cv-00201 (30 October 2023, ND Cal) (US).

5 Text and data mining ("TDM") is crucial for artificial intelligence models to create content, as they rely on vast datasets; however, TDM often finds itself in a challenging relationship with copyright law due to its nature of reproducing large amounts of data and copyrighted works. See Kyungsuk Kim, "Korean Copyright Issues in Text Data Mining for Generative AI" (2024) 1 *Journal of AI Law and Regulation* 64. Recognising the need to balance these interests, 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 L 130/92, which includes two

(cont'd on the next page)

data mining extracts insights from large volumes of information, often involving copyrighted materials, and although it drives innovations in sectors like healthcare, finance, and technology, it poses significant challenges to copyright laws. These laws aim to strike a balance between protecting rights and fostering technological progress.

5 Japan and Singapore are at the forefront of providing TDM-specific legal frameworks. Japan's Copyright Act amendments (passed in 2018, effective 1 January 2019) reflect the internal limit theory, which presumes that non-expressive uses fall outside the scope of copyright protection.[6] This framework permits broad TDM use while ensuring that copyrighted works are not consumed in their expressive capacity. In contrast, Singapore's Copyright Act 2021,[7] offers a "computational data analysis" ("CDA") exception, supporting both commercial and non-commercial TDM activities. Robust safeguarding policies prevent contractual terms from overriding the statutory exemption.[8]

6 Existing literature has extensively explored TDM frameworks in selected jurisdictions, including those of the US[9] and European Union[10] ("EU"), demonstrating the growing consensus that such exceptions foster data-driven innovation across key industries. The underlying theoretical justifications for TDM – grounded in notions such as the internal limit theory – reveal why non-expressive uses of works need not infringe

specific exceptions for TDM in Arts 3 and 4, aims to improve access to protected works and boost research and innovation. See also Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, "Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU" in *Propiedad Intelectual Y Mercado Único Digital Europeo* (C Saiz García & R Evangelio Llorca gen eds) (Tirant lo Blanch, 2019) at p 27.

6 Copyright Act (Act No 48 of 1970) (as amended by Act No 33 of 2023) (Japan) Art 30-4.

7 (2020 Rev Ed) (S'pore).

8 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 244. This section permits copying for computational data analysis ("CDA"), defined in s 243 as including: "(*a*) using a computer programme to identify, extract and analyse information or data from the work or recording; and (*b*) using the work or recording as an example of a type of information or data to improve the functioning of a computer programme in relation to that type of information or data". An illustration provided states that "the use of images to train a computer programme to recognise images" is an example of CDA under para (*b*).

9 See Copyright Act 17 USC (US) § 107. See *Authors Guild v HathiTrust* 755 F 3d 87 (2nd Cir, 2014), where the Second Circuit held that creating a searchable database of digitised books constituted fair use, recognising that such use was highly transformative. Similarly, Directive (EU) 2019/790 introduced exceptions to copyright for TDM.

10 See Arts 3–4 of the Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92. Article 3 provides an exception for scientific research by research organisations and cultural heritage institutions. Article 4 allows TDM for any purpose, subject to rights-holder opt-out.

copyright[11] (similarly, EU law confirms that unprotected elements, such as software functionality, are not subject to copyright[12]). Despite the proliferation of comparative studies, South Korea's incipient efforts have received comparatively limited scholarly scrutiny.

7 Text and data mining – the automated analytical processing of large collections of text or data – has become pivotal for research and innovation in the age of AI. However, because TDM often involves making copies or extractions of copyright-protected works as input, it can conflict with traditional copyright rules.[13] In response, lawmakers in various jurisdictions have crafted exemptions or exceptions to ensure that copyright law does not unduly hinder data-driven innovation.[14] This comparative analysis examines how three jurisdictions – South Korea, Singapore, and Japan – have approached TDM exceptions in their copyright laws, and highlights how these approaches differ from those in the EU and US.

B. Literature review

8 Legal scholarship on TDM exceptions situates these policies at the intersection of innovation and intellectual property rights. For instance, Gervais argues that copyright law's scope need not extend to purely computational readings of a work, as such activities do not encroach upon the work's expressive dimension.[15]

9 Samuelson further suggests that flexible legal doctrines – like fair use in the US – can accommodate new technologies without requiring extensive statutory overhaul.[16] However, she acknowledges that unpredictability remains an ongoing concern. Beyond these jurisdiction-specific analyses, international policy bodies, such as the

11 *ILOG Inc v Bell Logic LLC* 181 F Supp 2d 3 at 8 (D Mass, 2002). See also Copyrights 17 USC (US) § 102(b) (excluding ideas, procedures, and systems from copyright protection).

12 *SAS Institute Inc v World Programming Ltd* Case C-406/10, EU:C:2012:259, [2012] 3 CMLR 4 (holding that functionality, programming languages, and data formats are unprotected ideas).

13 Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92, Recital 8 (noting that TDM can "involve acts protected by copyright" requiring authorisation absent an exception).

14 Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92, Recital 11.

15 Daniel Gervais, "The Machine as Author" (2020) 105 *Iowa Law Review* 2053 at 2067 (arguing that TDM as "machine reading" typically qualifies as fair use). See also at 2089 (on Berne Convention originality requirements).

16 Pamela Samuelson, "Possible Futures of Fair Use" (2019) 90 *Washington Law Review* 815 at 864.

Organisation for Economic Co-operation and Development ("OECD"), emphasise the economic promise of data-driven innovation, encouraging member states to implement forward-looking frameworks.[17] This *corpus* of scholarship underscores the need to balance technological advancement with the legitimate protection of copyright holders, providing a valuable context for assessing South Korea's nascent TDM proposals.

C. Research methodology

10 The purpose of this article is twofold. Firstly, it examines whether producing images through latent-space diffusion models constitutes copyright infringement under current legal doctrines, particularly when the training data consist of copyrighted works. This entails a doctrinal analysis of key concepts – such as reproduction, substantial similarity, and derivative works – in the context of AI-generated outputs, as well as an evaluation of potential defences such as fair use. Secondly, the article investigates whether South Korea requires a dedicated TDM exception to address emergent AI applications, given comparative developments and policy debates. Drawing on cross-jurisdictional insights from the US, Japan, the EU, Singapore, and other relevant legal frameworks, the study assesses whether reliance on South Korea's broad fair use provision[18] is sufficient or whether a more explicit legislative framework is necessary to promote AI-driven innovation while safeguarding copyright holders' rights. Ultimately, the analysis informs how latent-space AI generation may be reconciled with copyright principles and how South Korea might benefit from statutory reforms.

11 To guide this inquiry, a central research question is posed: How does latent-space generation differ from direct copying or "rehydration" of a copyrighted work, and why is this distinction legally significant – especially under US copyright law? In addressing this question, a structured, comparative legal methodology was employed. South Korea, Japan, and Singapore were selected as the primary jurisdictions for detailed analysis, given South Korea's ongoing legislative considerations alongside Japan's and Singapore's established TDM frameworks. The EU and US contexts serve as auxiliary reference points, highlighting broader global trends and potential policy strategies. This comparative

17 Organisation for Economic Co-operation and Development, *Data-Driven Innovation: Big Data for Growth and Well-Being* (OECD Publishing, 2015) <https://www.oecd.org/sti/data-driven-innovation-9789264229358-en.htm> (accessed 16 July 2025).

18 Copyright Act (amended up to Act No 20841 of 25 March 2025) (South Korea) Art 35-5.

approach considers multiple factors, including the permissible scope of TDM activities (commercial *versus* non-commercial), legal prerequisites for lawful access and non-expressive use, and the role of contractual overrides. By surveying how these jurisdictions manage AI-driven data processing, the article elucidates the theoretical underpinnings – such as the internal limit theory – and the practical mechanisms necessary to align copyright law with the realities of latent-space AI generation.

12 The remainder of this manuscript proceeds as follows. Part II delves into the technical underpinnings of generative AI, explaining how latent-space representations work and why they typically do not replicate copyrighted material *verbatim*. Part III provides a comparative analysis of TDM exceptions, drawing on examples from Japan, Singapore, the EU, and the US. In doing so, it highlights key legislative approaches – such as Japan's broad "non-enjoyment" principle, Singapore's robust safeguards for lawful access, and the EU's tiered TDM exceptions – situating South Korea's nascent proposals within this global landscape. Part IV concludes, synthesising these findings and addressing both doctrinal and policy considerations. Normative recommendations are offered on how South Korea could introduce a dedicated TDM exception while preserving core copyright interests, ultimately contending that carefully calibrated legislative reform would benefit both rights holders and AI-driven innovation.

II. Technical underpinnings: latent spaces and the reality of "copying"

A. Fundamentals of generative artificial intelligence

13 Generative adversarial networks ("GANs") – which train a generator-discriminator pair to synthesise realistic image samples and thereby highlight the potential of the adversarial framework – mark a paradigm shift in AI image creation capabilities.[19] Unlike traditional systems that primarily focused on classification, these models generate original content emulating patterns observed in training data; however, they typically do not replicate specific works precisely.[20] The key lies in AI's ability to grasp the overall patterns and relationships within complex,

19 Ian Goodfellow *et al*, "Generative Adversarial Networks" (2020) 63(11) *Communications of the ACM* 139 <https://doi.org/10.1145/3422622> (accessed 7 October 2025).

20 Diederik P Kingma & Max Welling, "Auto-Encoding Variational Bayes" (2014) *International Conference on Learning Representations Conference Proceedings* <https://doi.org/10.48550/arXiv.1312.6114> (accessed 7 October 2025).

high-dimensional data, allowing it to generate new content that resembles the style and structure of the original without copying it exactly.[21]

B. Latent-space representation

14 Central to the image creation process in generative AI is the concept of a latent space – an abstract mathematical representation wherein the model encodes statistical patterns and relationships learned from training data, rather than storing the data itself.[22] Each point within this space corresponds to a potential image, and navigating through this space allows the model to interpolate between different visual concepts.[23] As this representation is both compressed and abstract, the model does not store or recall entire copyrighted images. Instead, it generates new content that reflects the general patterns and structures it has learned, without precisely replicating specific works.

C. Image generation process

15 During training, the model ingests millions of images, breaking them down into constituent features and encoding them into lower-dimensional latent representations.[24] Through repeated training cycles, the model learns to recognise visual patterns by comparing its predictions to actual results and adjusting its internal settings to improve accuracy. In the generation phase, an input prompt or parameter set guides the model to locate a point in the latent space that matches the desired attributes.[25]

16 Once the latent point is identified, the model decodes these abstract features into a final image. Importantly, this decoding does not simply "reverse" the earlier encoding by reproducing identical inputs; instead, just as the hidden units in a five-layer network capture the underlying structure of the task domain and thereby generalise it to

21 Carl Doersch, "Tutorial on Variational Autoencoders" (2016) <https://doi.org/10.48550/arXiv.1606.05908> (accessed 7 October 2025).

22 Carl Doersch, "Tutorial on Variational Autoencoders" (2016) <https://doi.org/10.48550/arXiv.1606.05908> (accessed 7 October 2025).

23 Tom White, "Sampling Generative Networks" (2016) *Proceedings of the 30th Conference on Neural Information Processing Systems* <https://doi.org/10.48550/arXiv.1609.04468> (accessed 7 October 2025).

24 Ian Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks" (2016) *Proceedings of the 30th Conference on Neural Information Processing Systems* <https:// doi.org/10.48550/arXiv.1701.00160> (accessed 7 October 2025).

25 Yoshua Bengio, Aaron Courville & Pascal Vincent, "Representation Learning: A Review and New Perspectives" (2013) 35 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1798 at 1813–1817.

novel input-output pairs, it leverages those learned statistical regularities to synthesise a fresh arrangement of pixels.[26] Though near-duplicates can theoretically occur (*eg*, the so-called "Italian Plumber" problem[27]), such cases are relatively rare.[28] For most generative models trained on large datasets, the outputs reflect new combinations of learned features, akin to a human artist drawing on their training without mechanically reproducing earlier works.

D. Text-to-image models

17 While previous discussions have acknowledged the role of prompts in generative AI, they often do not fully explore how these inputs function as precise controls in image generation. Many contemporary generative AI systems, particularly text-to-image models, accept textual inputs to guide image creation.[29] These models employ sophisticated text encoders to translate linguistic descriptions into visual concepts.[30] For example, a prompt such as "a futuristic cityscape at sunset" does not retrieve an existing image but instead synthesises a novel image that reflects the described concept.

(1) Translation from text to visual output

18 Text prompts are first tokenised and embedded into numerical vectors, which transformer-based architectures then process to capture the context.[31] The system employs attention mechanisms to correlate textual references with visual attributes.[32] Finally, the processed textual information is mapped to the latent space of images, creating a visual

26 David E Rumelhart, Geoffrey E Hinton & Ronald J Williams, "Learning Representations By Back-Propagating Errors" (1986) 323 *Nature* 533 at 535.

27 Timothy B Lee & James Grimmelmann, "Why the New York Times Might Win Its Copyright Lawsuit Against OpenAI", *Ars Technica* (20 February 2024) <https://arstechnica.com/tech-policy/2024/02/why-the-new-york-times-might-win-its-copyright-lawsuit-against-openai/#page-3> (accessed 16 July 2025).

28 Martín Abadi *et al*, "TensorFlow: A System for Large-Scale Machine Learning" (2016) *OSDI'16 Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* 265.

29 Aditya Ramesh *et al*, "Zero-Shot Text-to-Image Generation" (2021) 139 *Proceedings of Machine Learning Research* 8821.

30 Tomás Mikolov *et al*, "Efficient Estimation of Word Representations in Vector Space" (2013) *International Conference on Learning Representations Conference Proceedings* <https://doi.org/10.48550/arXiv.1301.3781> (accessed 7 October 2025).

31 Ashish Vaswani *et al*, "Attention Is All You Need" (2017) *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000 at 6004 <https://dl.acm.org/doi/10.5555/3295222.3295349> (accessed 7 October 2025).

32 Kelvin Xu *et al*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" (2015) *Proceedings of Machine Learning Research* 2048 at 2049–2051.

output.[33] Translating textual concepts into imagery can yield novel and creative results.[34] However, if a textual prompt precisely describes a unique copyrighted image, the risk of near-replication may increase.

(2) Limits of transformation

19 While it is tempting to say that text-to-image models always create "new expressions", certain prompts might produce outputs that closely match existing copyrighted works.[35] Thus, although these models can demonstrate transformation, there is a possibility of reproducing protected content, especially in edge cases. As with human creators, the outcome depends on how faithfully the prompt mirrors specific originals.

E. Core principles: from general adversarial networks to transformers

20 Early generative AI research often referenced GANs, introduced by Goodfellow *et al*.[36] While GANs demonstrated the capability to synthesise images from random noise, the field has since evolved to include diffusion models,[37] stable diffusion,[38] and large-scale transformer architectures for text-to-image tasks.[39] A critical factor for copyright analysis is how these models handle training data: rather than storing entire images, they learn statistical patterns – such as edges, colours, and shapes – and encode these features into high-dimensional weight matrices.

---

33 Scott Reed *et al*, "Generative Adversarial Text to Image Synthesis" (2016) 48 *Proceedings of Machine Learning Research* 1060.

34 Yuxiang Wei *et al*, "Elite: Encoding Visual Concepts Into Textual Embeddings for Customized Text-To-Image Generation" (2023) *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 15897 at 15943.

35 Han Zhang *et al*, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks" (2017) *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* 5908 <https://doi.org/10.1109/ICCV.2017.629> (accessed 7 October 2025).

36 Ian Goodfellow *et al*, "Generative Adversarial Networks" (2020) 63(11) *Communications of the ACM* 139 <https://doi.org/10.1145/3422622> (accessed 7 October 2025).

37 Prafulla Dhariwal & Alexander Nichol, "Diffusion Models Beat GANs On Image Synthesis" (2021) *Proceedings of the 35th International Conference on Neural Information Processing Systems* 8780.

38 Diederik P Kingma & Max Welling, "Auto-Encoding Variational Bayes" (2014) *International Conference on Learning Representations Conference Proceedings* <https://doi.org/10.48550/arXiv.1312.6114> (accessed 7 October 2025).

39 Ashish Vaswani *et al*, "Attention is All You Need" (2017) *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000 <https://dl.acm.org/doi/10.5555/3295222.3295349> (accessed 7 October 2025).

This approach means that the model's "memory" functions more like a learned mapping of patterns than a repository of specific images.[40]

21 Modern transformer models, such as GPT-3, which stacks layers with *alternating dense and locally-banded sparse* attention, rely on attention mechanisms that dynamically weight different parts of the input sequence (or embedded visual features) when predicting the next token.[41] In text-to-image contexts, the model effectively learns associations between textual descriptors (*eg*, "purple sky", "floating island", "impressionist style") and underlying visual representations. Crucially, these representations encode statistical patterns and features rather than storing the exact pixel arrangement of any original training image. While rare cases of "overfitting" or "memorisation" can occur (typically affecting a fraction of the training data, particularly duplicated data), this differs fundamentally from data storage as the model encodes learned patterns in billions of parameters rather than retaining raw training data.[42] Thus, while the model's training set is broad, the final output emerges from a learned manifold – a structured, internal representation that captures the essential patterns and relationships within the training data. This manifold enables the model to generate new images that are coherent and stylistically consistent with the learned concepts, without replicating any single source image line-for-line.

F. Decoding and the myth of "reverse copying"

22 Some critics have expressed that the final decoding step in generative AI models might directly reconstruct original training images.[43] However, this characterisation oversimplifies the process. Decoding is better understood as sampling from a learned latent distribution – a statistical representation of patterns and features extracted from the training data. Rather than retrieving exact copies, the model synthesises new content by combining and interpolating these learned features. While certain outputs may resemble training data, especially in cases of

40 Matthew Sag, "Copyright Safety for Generative AI" (2023) 61 *Houston Law Review* 295 at 305.

41 Tom B Brown *et al*, "Language Models Are Few-Shot Learners" (2020) *Proceedings of the 34th International Conference on Neural Information Processing Systems* 1877 at 1880.

42 Nicholas Carlini *et al*, "Extracting Training Data from Large Language Models" (2021) *Proceedings of the 30th USENIX Security Symposium* 2633.

43 Gowthami Somepalli *et al*, "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models" (2023) *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 6048.

overfitting or memorisation, the model encodes statistical patterns in its parameters rather than storing or retrieving specific images.[44]

23 The "Italian Plumber" problem[45] underscores a genuine concern that memorisation might sometimes produce near-identical replicas. However, empirical studies indicate that such instances are unusual and often associated with overtraining on small or niche datasets.[46] In large-scale models, such as Stable Diffusion or DALL E 2, developer teams typically adopt regularisation or deduplication strategies to curb memorisation.[47] If such outliers do arise, the relevant legal question is whether these sporadic duplications define the entire system's operation (they generally do not) or merely represent edge cases that courts can address on a fact-specific basis.

G. Substantial similarity in outputs and the model itself

24 *Substantial similarity* requires that the defendant's work incorporate protectable elements of the plaintiff's copyrighted expression.[48] In many jurisdictions, courts (*eg*, in the US Second and Ninth Circuits) use a combination of "extrinsic" and "intrinsic" tests or an "abstraction-filtration-comparison" approach.[49] Artificial intelligence outputs generally combine features from multiple sources, creating new compositions that deviate from any specific reference. Consequently, these outputs rarely exhibit the direct look and feel that would rise to the level of unlawful appropriation.[50]

25 Meanwhile, plaintiffs might allege that the *model itself* constitutes an infringing copy by storing or encoding copyrighted images. Yet, as explained, the model's parameters reflect aggregated patterns rather

44 Jonathan Ho & Tim Salimans, "Classifier-Free Diffusion Guidance" (2022) *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* <https://doi.org/10.48550/arXiv.2207.12598> (accessed 7 October 2025).

45 Timothy B Lee & James Grimmelmann, "Why the New York Times Might Win Its Copyright Lawsuit Against OpenAI", *Ars Technica* (20 February 2024) <https://arstechnica.com/tech-policy/2024/02/why-the-new-york-times-might-win-its-copyright-lawsuit-against-openai/#page-3> (accessed 16 July 2025).

46 Matthew Sag "Copyright Safety for Generative AI" (2023) 61 *Houston Law Review* 295 at 327–337.

47 Amirata Ghorbani & James Zou, "Data Shapley: Equitable Valuation of Data for Machine Learning" (2019) 97 *Proceedings of Machine Learning Research* 2242.

48 *Nichols v Universal Pictures Corp* 45 F2d 119 (2nd Cir, 1930).

49 *Computer Associates International Inc v Altai Inc* 982 F2d 693 (2nd Cir, 1992).

50 Matthew Sag, "Copyright Safety for Generative AI" (2023) 61 *Houston Law Review* 295 at 334–336.

than pixel-by-pixel or code-level duplication.[51] Courts have recognised that "intermediate" copies in the course of text/data processing may be protected under fair use or TDM exceptions as long as they do not supplant the original's value in the market.[52] The *Getty Images* complaint[53] specifically points to partial watermark recognitions, contending that repeated artifacts demonstrate direct copying. From a strictly doctrinal standpoint, however, a *de minimis* appearance of partial watermarks or style elements does not automatically equate to substantial similarity across the entire output.

H. Policy rationale for intervention even if non-infringing

26 The conclusion that latent-space generation is lawful implies that no further legislation is required. However, legal uncertainty and widespread misconceptions fuel expensive, protracted litigation that can stifle innovation. Many AI developers remain unsure whether their reliance on fair use would stand if tested in appellate courts, especially as new controversies such as watermark duplication arise.

27 Targeted policy interventions might provide greater clarity. For example, Congress (or administrative bodies) could create explicit TDM carve-outs, affirming that ephemeral training-stage copies do not infringe when the final outputs do not replicate entire works. Alternatively, guidelines might require or encourage AI developers to incorporate "anti-memorisation filters" to reduce the risk of near-duplicate images. Industry stakeholders might voluntarily adopt labelling or licensing frameworks, thereby mitigating disputes and giving rights holders a mechanism to opt in or out. While AI models might be largely non-infringing*, this does not preclude legislative or administrative action* to reduce friction, ensure fairness, and protect against edge-case abuses.

III. Comparative analysis of text and data mining exceptions

A. Japan

28 Japan was an early adopter in introducing a TDM exception. In 2018, Japan amended its Copyright Act (effective 2019) to include a broad exemption for uses of works in machine learning and data

51 J Hartman, "Neural Net Weights As Data Storage: Myth or Reality?' (2022) 62 *Jurimetrics* 485 at 490–493.

52 *Authors Guild v Google Inc* 804 F3d 202 (2nd, Cir 2015).

53 *Getty Images (US) Inc v Stability AI Ltd* No 3:25-cv-06891-TLT (14 August 2025, ND Cal) (US).

analysis. Japan's TDM exception, introduced comprehensively in the 2018 amendments to the Copyright Act (though partial recognition dates back to 2009), permits both commercial and non-commercial TDM activities without requiring additional permissions. The approach includes a non-enjoyment clause under Art 30-4 of the Copyright Act of Japan, ensuring that copying for data analysis rather than for expressive consumption does not infringe on copyrights.[54] This exception is based on the theory that exploitation not for enjoyment purposes falls outside the inherent scope of copyright protection, a position congruent with the internal limit theory, whereby non-expressive uses are deemed outside the scope of copyright protection.

29 Japan's experience suggests that permitting TDM across broad contexts can foster research and market growth, but ongoing empirical evaluation is essential as AI-driven technologies advance rapidly. The exception in Art 30-4 explicitly encompasses data analysis, defined as the extraction of information (*eg*, linguistic or image data) from a large number of works.[55] Furthermore, it includes a catch-all for other non-expressive uses, such as machine processing of works where no human perceives the protected expression, beyond the examples listed. A built-in condition requires that the use "not unreasonably prejudice the interests of the copyright owner" – essentially echoing the three-step test.[56]

30 Unlike Singapore law and the EU approach, Japan's statute does not explicitly require that the works be lawfully obtained or exclude infringing sources. Thus, on its face, Art 30-4 could even apply to data

54 Copyright Act (Act No 48 of 1970) (as amended by Act No 33 of 2023) (Japan) Art 30-4 (permitting exploitation of works for "non-enjoyment" purposes including data analysis).

55 Copyright Act (Act No 48 of 1970) (as amended by Act No 33 of 2023) (Japan) Arts 30-4(i)–30-4 (iii).

56 The three-step test is a fundamental standard in international copyright law that defines the permissible scope of exceptions and limitations to exclusive rights. It originated in the 1967 Stockholm Revision of the Berne Convention for the Protection of Literary and Artistic Works (9 September 1886), 828 UNTS 221 (entered into force 29 January 1970), which introduced the test for exceptions to the reproduction right (Art 9(2) Berne Convention). This formulation was later incorporated and generalised in Art 13 of the Agreement on Trade-Related Aspects of International Property Rights (15 April 1994), 1869 UNTS 299 (entered into force 1 January 1995) ("TRIPS") which applies the test to all exclusive rights. Article 13 of TRIPS stipulates: "Members shall confine limitations or exceptions to exclusive rights to certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder." Hence, any copyright exception must satisfy the three steps or criteria. It must: (a) be confined to certain special cases; (b) not conflict with a normal exploitation of the work, and (c) not unreasonably prejudice the legitimate interests of the rights holder.

mined from unauthorised copies of works. However, Japanese authorities have clarified that the "no enjoyment" requirement and the general prejudice clause limit the scope: the TDM exception cannot be used as a pretext to distribute or consume pirated content, and it remains subject to the overarching fairness principle of not unduly harming rights holders.[57] In practice, Japan's broad exception has provided a flexible legal environment for AI development, with minimal restrictions on TDM, which contrasts with the more conditional frameworks elsewhere.

B. Singapore

31 Singapore's CDA exception, enshrined in s 244 of the Copyright Act 2021, offers a carefully calibrated framework designed to promote AI and data-driven innovation. It permits both commercial and non-commercial TDM, mandates lawful access, and precludes copyright holders from opting out of statutory exemptions – effectively prohibiting contractual overrides. By comparing this model to the EU's more research-oriented approach and the US's flexible but less TDM-specific fair use framework, it becomes evident that Singapore's measures provide a high degree of legal certainty and user empowerment.

32 Singapore overhauled its copyright law with the Copyright Act 2021, which took effect on 21 November 2021 and introduced an explicit exception for what it refers to as CDA. Found in s 244 of the Act, this exception allows anyone to make copies of works or recordings for the purpose of TDM or other data analysis. Notably, Singapore's CDA exception applies to both non-commercial and commercial activities – there is no limitation to scientific research or other specific purposes. Section 243 of the Copyright Act 2021 defines CDA broadly, covering the use of computer programs to identify, extract, and analyse information from a work, including using it as example data to improve an algorithm.[58] Several important safeguards are built into the Singapore framework: the person relying on the exception must have lawful access to the content (*eg*, *via* purchase or subscription, not *via* hacking or infringement),[59] and any copies made are not to be shared with others except in limited

57 Hugh Stephens, "Japan's Text and Data Mining (TDM) Copyright Exception for AI Training", *Hugh Stephens Blog* (10 March 2024) <https://hughstephensblog.net/2024/03/10/japans-text-and-data-mining-tdm-copyright-exception-for-ai-training-a-needed-and-welcome-clarification-from-the-responsible-agency> (accessed 15 April 2025).

58 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 243 stipulates that CDA includes using a program to "identify, extract and analyse information or data from the work" and using the work as data to improve a program's functioning.

59 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 244(2)(*d*).

cases such as with collaborating researchers or for results verification.[60] Crucially, Singapore explicitly disallows any contractual override of this exception – any contract term that purports to prohibit or restrict TDM uses allowed by the law is unenforceable.[61] This means, for example, that even if a website's terms of service attempt to forbid scraping or data mining, a user with lawful access could still lawfully mine the content under Singapore law. Because Singapore's TDM exception applies to all types of works and users and cannot be overridden by private contracts, its framework is especially permissive and innovation-friendly compared to many other jurisdictions.[62]

C. Comparative analysis

33 Several themes emerge from our comparative analysis. South Korea's draft TDM provision, as well as Japan's and Singapore's laws, include a requirement for lawful access to materials (with Japan's lawful access condition implied rather than explicitly stated). Japan and Singapore both permit TDM for commercial purposes, whereas in South Korea the inclusion of TDM for commercial purposes in the proposed TDM exception clause is still under debate. Singapore's Copyright Act 2021 expressly voids any contract term that purports to exclude or restrict the CDA exception.[63] This provides additional safeguards to users, allowing both commercial and non-commercial TDM use under the exception. In contrast, Japan does not provide statutory prohibitions

60 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 244(2)(*c*) stipulates that copies made under the CDA exception cannot be provided to any other person, except to the extent necessary for "verifying the results of the computational data analysis" for "collaborative research" or a related study.

61 A key feature of Singapore's CDA exception is the prohibition of contractual overrides. Under s 187(1)(*c*) of Copyright Act 2021 (2020 Rev Ed) (S'pore), "Any contract term is void to the extent that it purports, directly or indirectly, to exclude or restrict any permitted use under any provision in ... Division 8 (computational data analysis)." This provision imposes a blanket prohibition on contractual overrides to TDM exceptions, rendering unenforceable any contractual terms that are inconsistent with these exceptions. It ensures that users can rely on the statutory exception for CDA without being hindered by restrictive contractual terms. This legislative approach ensures that statutory rights are preserved despite any private agreements to the contrary.

62 Pin-Ping Oh, "Coming Up in Singapore: New Copyright Exception for Text and Data Mining", *Bird & Bird Insights* (19 September 2021) <https://www.twobirds.com/en/insights/2021/singapore/coming-up-in-singapore-new-copyright-exception-for-text-and-data-mining> (accessed 16 July 2025). Oh notes that extending the exception to all users (commercial and non-commercial) and disallowing contractual override in all cases makes the Singaporean approach "much more permissive" and of greater utility to businesses, compared to jurisdictions that allow rights holders to opt out.

63 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 187(1)(*c*).

of contractual overrides for TDM exceptions in its copyright law. In Japan, this principle is upheld through legal interpretation rather than explicit statutory language.[64] In South Korea, whether to bar contractual overrides remains under legislative consideration.

34 In practice, Japan's TDM exception has been used by pharmaceutical researchers to analyse large collections of medical literature, a process that could potentially accelerate aspects of drug discovery. Likewise, Singapore's clear and stable legal framework is thought to encourage financial analytics firms to utilise large datasets in their work. Meanwhile, some efforts to introduce broad TDM exceptions in Europe have encountered industry push-back, highlighting the risks of moving ahead without sufficient stakeholder buy-in. Overall, these examples underscore how critical precise legislative design is to reap the full benefits of TDM while mitigating risks. Table 1 compares the scope and key characteristics of the TDM exceptions enacted in Japan and Singapore, revealing their shared tolerance for commercial uses and refusal to allow rights holder opt-outs.[65]

Table 1. Comparison of TDM exception provisions in Japan and Singapore

Item	Japan	Singapore
Introduction year	2018	2021
Scope	Commercial/ Non-commercial allowed	Commercial/ Non-commercial allowed
Contract to restrict TDM	Void (interpretation)	Void (s 187(1)(*c*))
Lawful access required	Implicit (legitimate copy presumed)	Yes
Opt-out by rights holder	Not possible	Not possible
Key characteristics	Use permitted for "non-enjoyment" purposes	Use permitted for "computational data analysis"

64 Tatsuhiro Ueno, "The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication" (2021) 70(2) *GRUR International* 145 at 149; Artha Dermawan, "Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn From the Japanese "Nonenjoyment" Purposes?" (2024) 27 *The Journal of World Intellectual Property* 44 at 54.

65 Based on Copyright Act (Act No 48 of 1970) (as amended by Act No 33 of 2023) (Japan) Art 30-4 and Copyright Act 2021 (2020 Rev Ed) (S'pore) ss 187(1)(*c*) and 243–244.

35 Notable differences between the TDM frameworks of Japan and Singapore reflect their distinct legislative history and policy orientations. Japan initially recognised TDM-related uses in 2009 and significantly expanded the scope of its exemption in 2018. In contrast, Singapore's TDM exception took shape in 2021, embodying a more contemporary vision of technological innovation and market responsiveness. Although Japan does not explicitly list "lawful access" as a statutory requirement, the internal limit theory presumes that the use of copyrighted materials must be legitimate – meaning infringing copies are not exempt. In comparison, Singapore's framework *expressly* mandates lawful access. Conceptually, Japan's framework emphasises "non-enjoyment" or consumption of copyrighted expressions, whereas Singapore's framework foregrounds the notion of computational data analysis, signalling a proactive endorsement of AI-driven research and development.

36 Despite these distinctions, both jurisdictions share several foundational similarities that distinguish them as leaders in crafting supportive legal environments for AI and related technologies. Both Japan and Singapore allow TDM activities for commercial and non-commercial purposes, thus fostering a wide range of innovative applications across the private and public sectors. Neither jurisdiction permits contractual provisions that override statutory exemptions, ensuring that legislative intent cannot be undermined by private agreements.[66] Furthermore, rights holders cannot use any opt-out mechanism, ensuring that access to materials for TDM is not arbitrarily restricted. Thus, Japan and Singapore serve as leading examples of comprehensive, flexible TDM exception frameworks intended to stimulate research, development, and innovation in fast-evolving digital economies.

D. European Union

37 The approaches in South Korea, Singapore, and Japan illustrate a spectrum of TDM exceptions, which can be contrasted with developments in the EU and US. The EU only recently addressed TDM in its copyright law *via* the Directive on Copyright and Related Rights in the Digital Single Market.[67] This Directive requires EU member states to implement two *mandatory* TDM exceptions: one for TDM for the

66 Copyright Act 2021 (2020 Rev Ed) (S'pore) s 187(1)(*c*) explicitly renders contrary contract terms unenforceable. In Japan, this is an interpretation, not an explicit statutory provision. The mandatory nature of Art 30-4 of the Copyright Act of Japan is widely understood to override contracts, but the law itself is silent on the matter.

67 Directive (EU) 2019/790, [2019] OJ L 130/92.

purposes of scientific research benefiting research organisations and cultural heritage institutions,[68] and another for TDM for any purpose, which is available to all users including commercial entities.[69] Both EU exceptions apply only to content that has been lawfully accessed by the user.[70] However, the scope and flexibility of the two differ: the research-focused exception under Art 3 cannot be overridden by contract and rights holders are not allowed to opt out of it, reflecting a strong policy in favour of research uses. By contrast, the general TDM exception under Art 4 allows rights holders to opt out of permitting mining of their works (for example, by machine-readable means such as a meta-tag or robots.txt file, indicating reservation of rights).[71]

38 Even before these new statutory provisions, European case law had begun carving out room for non-expressive uses. Notably, the Court of Justice of the EU in *SAS Institute v World Programming Ltd*[72] held that neither the functionality of a computer program nor the act of observing or testing a program's functioning in order to create an interoperable program is protected by copyright (only the expression of the program

68 See Art 3 of Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92. Article 3 mandates an exception for reproductions and extractions by research organisations and cultural heritage institutions for purposes of scientific research TDM.

69 See Art 4 of Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92. A TDM exception that clearly permits both commercial and non-commercial uses reduces ambiguity for businesses and researchers by clarifying the legal framework for lawful TDM activities. In the EU, Art 4 of Directive (EU) 2019/790 introduces such a general exception for reproductions and extractions of lawfully accessible works for TDM by others for any purpose. It allows entities to develop algorithms without the burden of authorisation and remuneration, provided the right has not been expressly reserved by the rights holder. According to Vrakas, "These exceptions were welcomed to a certain extent, for their potential to harmonise a collective EU-wide approach towards 'lawful TDM', ultimately fostering competition." See Giorgos Vrakas, "A Literature Review of 'Lawful' Text and Data Mining" (2024) 4 *Open Research Europe* 153 at 160.

70 Lawful access clauses in the TDM exceptions of Directive (EU) 2019/790 (Arts 3 and 4) ensure that only individuals with legitimate access to content may engage in TDM, thus respecting the rights of content owners and prohibiting unauthorised access. Recital 14 of the Directive underscores this principle, stating that "research organisations and cultural heritage institutions, including the persons attached thereto, should be covered by the text and data mining exception with regard to content to which they have lawful access". Such provisions aim to balance the interests of rights holders and users, facilitating innovation through TDM while safeguarding against infringement.

71 See Art 4(3) of Directive on Copyright and Related Rights in the Digital Single Market, Directive (EU) 2019/790, [2019] OJ L 130/92. The Art 4 TDM exception "shall not apply" if rights holders have "expressly reserved" the use of their works in TDM, that is *via* machine-readable means in the case of online content.

72 Case C-406/10, EU:C:2012:259, [2012] 3 CMLR 4.

is protected).[73] This effectively allowed a form of data analysis in the software context under existing law, foreshadowing the more explicit TDM exceptions to come.

E. United States

39 In the US, there is no specific statutory exception for TDM; instead, fair use under 17 USC § 107 has been the primary mechanism authorising TDM activities. US courts have generally been receptive to characterising TDM and other non-consumptive, transformative uses as fair use. A landmark example is the case of *Authors Guild v Google Inc*[74] ("*Google Books*"), where a court held that Google's mass scanning of books to create a searchable database (and display snippet results) was a transformative use and therefore a fair use, despite the commercial nature of the project.[75] Similarly, copying images to analyse them for an image recognition AI, or ingesting large datasets to train an AI model, would likely be viewed as transformative and not substituting for the original works, thus favouring a fair use finding. The US Supreme Court's decision in *Campbell v Acuff-Rose Music Inc*[76] ("*Campbell*") underscored that even commercially motivated uses can qualify as fair use if they are transformative in purpose and character.[77] As the court famously noted, "parody has an obvious claim to transformative value" and "the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works".[78] This philosophy has paved the way for the flexible application of fair use to new contexts, especially data mining. However, because fair use is a case-by-case, open-ended doctrine, it can leave uncertainty; there is no guarantee that a particular TDM project will be deemed fair use, especially if the specific use arguably affects the market for the original. The bright-line statutory exceptions in the EU, Singapore, and Japan provide more certainty in those jurisdictions, though often with more conditions.

---

73 *SAS Institute Inc v World Programming Ltd* Case C-406/10, EU:C:2012:259, [2012] 3 CMLR 4 at [39], [46], [56] and [61]–[62].

74 804 F 3d 202 (2nd Cir, 2015).

75 *Authors Guild v Google Inc* 804 F 3d 202 at 207–208 (2nd Cir, 2015). The mass digitisation of millions of books to create a searchable database was held to be a transformative fair use since it adds value to the original and does not substitute for it.

76 510 US 569 (1994).

77 The US Supreme Court in *Campbell v Acuff-Rose Music Inc* 510 US 569 at 579 (1994) emphasised that a parody's commercial character does not preclude a finding of fair use. The court also observed at 575 that "some opportunity for fair use of copyrighted materials has been thought necessary" (quoting US Constitution Art I § 8 cl 8).

78 *Campbell v Acuff-Rose Music Inc* 510 US 569 at 579 (1994).

F. South Korea

40 South Korea's legislative efforts to establish a TDM exception coincide with the growing importance of AI and data analytics in global innovation ecosystems. Japan's experience demonstrates that permissive and theoretically grounded exemptions can foster research and market growth. Singapore's CDA framework illustrates how user safeguards and clear legal boundaries create a stable environment for technological advancement. By examining lessons from the EU and the US, South Korea should consider strategies to encourage cross-border research collaboration and respond flexibly to new technological developments.

41 In drafting its TDM exception, South Korea would benefit from adopting broadly permissive provisions that unambiguously permit both commercial and non-commercial uses. Clarifying that these activities are aimed at analysis rather than the expressive consumption of copyrighted works is essential. However, the four current proposals in South Korea illustrate an ongoing debate, with some bills (introduced by Do Jong-hwan,[79] Lee Yong-ho,[80] and Hwangbo Seung-hee[81]) implicitly or explicitly accommodating commercial use, and Lee In-young's proposal[82] restricting TDM to non-commercial, research-oriented contexts. Such internal divergence indicates that legislators seek to reconcile robust copyright protections with the desire to spur AI-driven innovation. The noted bills ultimately failed to pass before the 21st National Assembly's term expired on 29 May 2024, and thus were effectively discarded. On 11 February 2025, Representative Kim Tae-seon introduced a new TDM exception bill in the 22nd National Assembly, proposing that non-commercial data analysis be permitted under certain security conditions.[83] The bill builds on lessons from the four failed proposals of the 21st Assembly.

42 Transitionary measures, such as a phased implementation period accompanied by governmental guidance, can facilitate stakeholder adjustment to new legal environments. Furthermore, a periodic review committee could assess the exemption's efficacy and economic impact,

79 Copyright Act Amendment Bill No 2107440 (South Korea) (proposed by Representative Do Jong-hwan and 12 others, 15 January 2021).

80 Copyright Act Amendment Bill No 2117990 (South Korea) (proposed by Representative Lee Yong-ho and others, 31 October 2022).

81 Copyright Act Amendment Bill No 2122537 (South Korea) (proposed by Representative Hwangbo Seung-hee and others, 8 June 2023).

82 Copyright Act Amendment Bill No 2124685 (South Korea) (proposed by Representative Lee In-young and others, 25 September 2023).

83 Copyright Act Amendment Bill No 2208071 (South Korea) (proposed by Representative Kim Tae-seon and others, 11 February 2025).

enabling evidence-based refinements over time. Engaging a broader slate of stakeholders, including legal experts, industry representatives, and academics in the assessment of copyright policy could enable continuous evaluation of the impact of copyright exceptions. For instance, creators' organisations could participate in shaping copyright policy. Such engagement facilitates the monitoring of usage, economic benefits, and unintended consequences, allowing for data-driven recommendations to refine legislation and maintain its effectiveness and relevance.

43 As TDM technologies continue to evolve and become more closely integrated with machine learning models and cross-border data networks, South Korea's statutory regime must remain adaptable. Future research should track longitudinal data on innovation metrics, monitor industry feedback, and engage with external jurisdictions. By drawing on the experiences of Japan, Singapore, the EU, and the US – and internalising these insights – South Korea can design a TDM exception that not only advances its own AI and data analytics sectors but also contributes to the global discourse on harmonising copyright law with the demands of the digital age.

44 In sum, bridging the four South Korean TDM bills with well-established international frameworks could yield a balanced TDM exception. Policymakers should consider permitting commercial TDM under rigorous security standards, requiring lawful access, and clarifying whether derivative works are permissible for purely analytical transformations. Do Jong-hwan's and Hwangbo Seung-hee's broad approaches, Lee Yong-ho's emphasis on protective measures, and Lee In-young's focus on non-commercial education and research collectively reflect the spectrum of choices South Korea faces. Whatever the final synthesis, ensuring robust legislative language, consistent enforcement, and meaningful dialogue among stakeholders will bolster public trust and encourage innovation across the AI ecosystem – ultimately aligning South Korea's TDM framework with global best practices.

(1) Ministry of Culture and Korea Copyright Commission Guidelines

45 With the emergence of generative AI, represented by ChatGPT, expectations for AI as a content creation tool have increased. Simultaneously, concerns have arisen from a copyright protection perspective regarding unauthorised use of training data and copyright infringement by AI outputs. To address these new copyright issues in the AI era, the Ministry of Culture, Sports and Tourism ("MCST") first established a public-private AI-Copyright System Improvement Working Group in 2023. Subsequently, the MCST and the Korea Copyright Commission ("KCC") jointly published a *Guide on Generative AI and Copyright*, encouraging AI developers to obtain licenses or

pay compensation for works used in training whenever possible, and suggesting that rights holders should be allowed to opt out of having their content used.[84] The guide advises that AI operators must secure legitimate usage rights through appropriate compensation before using copyrighted works for AI training. Furthermore, it encourages copyright holders opposed to such use to clearly state their opposition or employ technical countermeasures, aiming to balance the interests of creators and technological advancement.

46 In 2024, the working group was divided into learning and output subcommittees to discuss a wide range of copyright-related issues, including the use of copyrighted works such as training data, recognition of copyright for AI outputs, and responses to copyright infringement. Thus, they identified key issues requiring consensus on legal and institutional improvement measures due to their widespread impact on the industry.

47 In June 2025, the MCST and the KCC jointly distributed guidelines on copyright registration standards for creative works using AI outputs,[85] and a copyright infringement assessment of AI outputs.[86] The establishment of this working group is considered an important attempt to balance AI technology and copyright law as well as improve related legal systems. This working group is expected to provide helpful input and guidance to the legislative process and may thus be drawn upon for future bills and legislative action that may become law.

(2) Copyright infringement lawsuit by South Korean Broadcasters against Naver over use of data for artificial intelligence training

48 On 13 January 2025, the three major South Korean terrestrial broadcasters (Korean Broadcasting System ("KBS"), Munhwa Broadcasting Corporation ("MBC") and Seoul Broadcasting System ("SBS")) initiated a copyright infringement lawsuit against Naver, South

84 Ministry of Culture, Sports and Tourism & Korea Copyright Commission, *A Guide on Generative AI and Copyright* (December 2023) < https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=52591> (accessed 15 April 2025).

85 Ministry of Culture, Sports and Tourism & Korea Copyright Commission, *Guidelines for Copyright Registration of Works Utilizing Generative Artificial Intelligence* (2025-02) (June 2025) <https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=54253> (accessed 16 July 2025).

86 Ministry of Culture, Sports and Tourism & Korea Copyright Commission, *Guidelines for Preventing Copyright Disputes Arising From Generative Artificial Intelligence Outputs* (2025-03) (June 2025) <https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=54252> (accessed 16 July 2025).

Korea's leading internet company, in Seoul Central District Court. As of late 2025, the case remains actively pending. The first court hearing was held 18 September 2025 and the second hearing on 6 November 2025. This legal action represents a significant development in the ongoing global debate about the use of copyrighted content for training AI systems. The lawsuit alleges that Naver unlawfully utilised news articles from these broadcasting companies to train its generative AI models, "HyperCLOVA" and "HyperCLOVA X," without proper authorisation or compensation.[87]

49 On 13 January 2025, the Korea Broadcasting Association ("KBA"), led by its President, Bang Moon-shin, announced that the three major broadcasters had filed a lawsuit against Naver in the Seoul Central District Court. The legal action seeks damages for copyright infringement and violations of the Unfair Competition Prevention and Trade Secret Protection Act,[88] along with an injunction to prohibit further use of their content for AI training purposes.[89] This case marks the first lawsuit filed by domestic media organisations against a major technology company in South Korea regarding the unauthorised use of news data for AI training.

50 Prior to initiating legal proceedings, the KBA had taken preliminary steps to address their concerns. In December 2023, the association sent formal notices to several domestic and international technology companies, including Naver, Kakao, Google Korea and Microsoft. These notices explicitly stated that separate compensation agreements would be necessary for using the broadcasters' news content, audio and video materials for AI training, prohibiting such use without permission. This legal strategy is complicated by market-based solutions emerging concurrently. In September 2025, one of the plaintiffs, SBS, announced its own formal fee framework for licensing its news data for AI training.[90] This development suggests the market is not simply waiting for a legislative TDM exception but is actively using the threat of litigation to create a licensing framework.

87 Kim Sang-hyeop, "Three Major Broadcasters File Lawsuit Against Naver for Unauthorised Use of News Content", *KBS News* (13 January 2025) <https://news.kbs.co.kr/news/pc/view/view.do?ncd=8150969> (accessed 15 April 2025).

88 (South Korea) Art 2, item 1, subitems (ka) and (pa).

89 *KBS, MBC, SBS v Naver Corporation* Case No 2025GaHap5105 (13 January 2025, Seoul Central District Court).

90 Lee Yoon-seo, "SBS Becomes First Broadcaster in Korea to Set Fees for AI Training Use of News Data", *The Korea Herald* (10 September 2025) <https://www.koreaherald.com/article/10572608> (accessed 10 November 2025).

IV. Conclusion

51 Generative AI image models rely on *latent-space* abstractions of their training data rather than storing or reproducing any protected imagery precisely. The technical evidence surveyed in this article illustrates that these systems learn conceptual patterns, artistic styles or common visual elements, by encoding statistical relationships in their parameters, without retaining the precise expressive details of any single input work. As a result, the output image is a novel combination of learned features, not a facsimile of a prior creation. From a legal standpoint, this means the AI's operations typically fail to meet the threshold of infringement: there is no *unauthorised reproduction* of protected expression that an ordinary observer would recognise as substantially similar to a specific original work. In essence, copying in the AI training context is non-literal and non-expressive, falling outside the scope of what copyright law forbids.

52 Under US copyright doctrine, this conclusion is reinforced by the principles of transformative use and the substantial similarity test. The US Supreme Court in *Campbell* emphasised that a use is "transformative" if it alters the original with new expression, meaning, or message, serving a different purpose than the original. Training a generative model is highly transformative: the purpose is to extract data patterns and enable new image creation, which is fundamentally distinct from the purpose of the original images (to be viewed for their expressive content). Indeed, the Second Circuit's decision in *Google Books* affirmed that making digital copies of entire books for a search index was a "highly transformative" use because it served a new informational purpose and did not substitute for the original market.[91] By analogy, an AI's ingestion of millions of images to generate a latent image-text space is similarly transformative – it is a process aimed at knowledge extraction and creativity, not at disseminating the original images. Moreover, US courts require that an accused's work share substantial protected expression with the plaintiff's work for infringement to exist. Given generative AI's training mechanics, any output is an algorithmic synthesis of countless inputs; it is exceptionally unlikely to replicate any single source image in a way that would satisfy the Second or Ninth Circuit standards for substantial similarity (which filter out unprotectable elements and assess the total concept and feel of the works). In an ordinary case, the latent-space output does not appropriate the original artist's creative choices in a manner recognisable to the lay observer, meaning no infringement occurs. Put simply, the generative process extracts ideas, styles, and features – those

91 *Authors Guild v Google Inc* 804 F 3d 202 at 216–218 (2nd Cir, 2015) <https://www.law.berkeley.edu/wp-content/uploads/2016/05/Authors-Guild-v-Google-804_F.3d_202.pdf> (accessed 16 July 2025).

elements often deemed unprotectable – without encroaching on the protected expression of any one artwork.

53 This understanding aligns with fair use jurisprudence and emerging policy trends. Courts have consistently favoured uses that do not usurp the market for the original work and that offer new social value. For instance, in *Google Books,* the court found no infringement in Google's unlicensed text scanning precisely because the output (searchable snippets) did not expose a substantial portion of the books' expressive content or serve as a market substitute. Likewise, an AI model's training phase does not reveal or compete with the original images – it yields only *abstract learned parameters*. The fair use doctrine's goal of promoting innovation strongly supports treating such intermediate copying as lawful, much as the Ninth Circuit did when it permitted intermediate software copies for reverse-engineering in *Sega Enterprises Ltd v Accolade Inc*.[92] The *transformative use* paradigm from *Campbell* and its progeny suggests that AI training, as a tool for creating new works and knowledge, should be protected in order to further "the Progress of Science and useful Arts".[93] Even after the Supreme Court's recent clarification in *Andy Warhol Foundation for the Visual Arts Inc v Goldsmith*[94] that transformation alone is not dispositive;[95] the key point remains that AI training does not target the expressive value of the originals and thus does not encroach on the core market rights of authors. Absent deliberate prompting to mimic a particular work, the generated images typically will not be substantially similar to any one source. In short, both the purpose of generative AI training and the *outcome* of that process place it outside the realm of copyright's infringement tests.

54 Comparative legal developments underscore this conclusion and point the way toward an appropriate legislative response. Facing these issues early on, Japan introduced Art 30-4 of its Copyright Act in 2018 to expressly permit uses of copyrighted works for *information analysis*. This provision allows copying for purposes such as machine learning and data mining so long as the use is not intended to enjoy the work's expressive content and does not *unreasonably prejudice* the copyright owner's interests. In practice, Japan's law gives AI developers a safe harbour to ingest large volumes of data in training generative models, recognising that such utilisation is socially beneficial and fundamentally different

92 977 F 2d 1510 (9th Cir, 1992) (finding that copying for a non-expressive, functional purpose was fair use).

93 *Campbell v Acuff-Rose Music Inc* 510 US 569 at 575 (1994).

94 598 US 508 (2023).

95 *Andy Warhol Foundation for the Visual Arts Inc v Goldsmith* 598 US 508 at pp 28–29 (2023) <https://www.supremecourt.gov/opinions/22pdf/21-869_87ad.pdf> (accessed 16 July 2025).

from the normal consumption of the works. The Singapore Copyright Act 2021 similarly introduced s 244, a forward-looking exception for "computational data analysis". This exception explicitly authorises making copies of works for data analysis purposes – including AI model training – in both commercial and non-commercial contexts. Notably, Singapore built in robust safeguards: any copies made may only be used for analysis and cannot be distributed except for limited verification or collaborative research, users must have *lawful access* to the works, and critically, no contract can override this exception. By preventing contractual terms from prohibiting TDM, the law ensures that the TDM right remains effective even against private restrictions. The EU has also embraced TDM exceptions through Directive (EU) 2019/790. Article 3 of the Directive requires Member States to permit TDM for scientific research by research institutions (an exception that cannot be contracted away), and Art 4 mandates an exception for *TDM by anyone for any purpose* on lawfully accessible content, subject only to the right of a rights holder to opt out of this use (*eg*, *via* machine-readable reservations). These advances reflect a growing international consensus that copyright laws must adapt to facilitate AI and big data innovation. Jurisdictions are carving out space for computational uses of works – uses that generate knowledge and new creativity without supplanting the market for originals.

55 In contrast, South Korea has yet to codify a similar exception, and this absence now puts its innovators and rights holders on uncertain legal ground. The South Korean Copyright Act does contain a general fair use provision (Art 35-5, analogous to 17 USC § 107), but it has not been tested in the courts for AI training data. Unlike Japan and Singapore, South Korea lacks a specific TDM exception, leaving a grey area as to whether unlicensed dataset compilation is permitted. A government-issued *Guide on Copyrights for Generative AI* ("Guidelines") in 2023 acknowledged the ongoing debate and lack of clear precedent on this issue. Tellingly, the Guidelines stopped short of declaring AI training categorically lawful under fair use; instead, it *cautioned AI developers to obtain licenses* for training data or otherwise ensure permission, while also advising rights holders on measures (like robots.txt exclusions and contract terms) to control AI use. This cautious approach signals that, under current law, companies in South Korea run a legal risk when using copyrighted content to train AI without consent as evidenced by the high-profile lawsuit brought against Naver by South Korea's three major broadcasters. However, that same dispute is also spurring market-based solutions, such as plaintiff SBS's September 2025 introduction of an AI news licensing framework, a solution that a broad TDM exception might pre-empt. The broadcasters argue that Naver's use of their content – absent any license or compensation – usurps their rights and business interests, effectively *free-riding* on proprietary media archives. This dispute, the first of its kind in South Korea, starkly

illustrates the uncertainty and conflict that arise from the vacuum of a clear TDM exception. It has become evident that relying on case-by-case fair use analysis or broad theories of transformation is untenable as a long-term solution; both innovators and creators need *ex ante* clarity on what is permissible.

56 South Korea must therefore act decisively to modernise its copyright framework. In light of the technical findings and comparative legal trends discussed, the South Korean Legislature should adopt a dedicated TDM exception that authorises the training of AI models on copyrighted works, provided certain conditions are met. This statutory exception should permit *reproducing and processing lawfully accessed works for the purpose of data analysis* (encompassing both non-commercial research uses and commercial AI development), thereby removing the ambiguity that currently plagues AI innovators. Simultaneously, the South Korean TDM exception must incorporate sensible safeguards to protect rights holders. These can be drawn from other models abroad discussed in this article, *eg*, a condition that the use must not unreasonably prejudice the normal exploitation of the work (as in Japan), and requirements to implement technical measures that prevent AI from outputting content that simply replicates training data. The exception should also require that those engaging in TDM have lawful access to the source material (mirroring the EU and Singapore) – in other words, the data used must be obtained through legitimate means or public availability, not through hacking or breach of access controls. Crucially, any South Korean TDM exception should prohibit private contracts from overriding it, ensuring that a website's terms of service or a licensing contract cannot eliminate the user's right to perform TDM on lawfully accessed content. This framework preserves the public policy balance and prevents contractual boilerplate from nullifying the exception's effect. Finally, the law might consider a registration or opt-out mechanism for rights owners who *genuinely wish to exclude* their content (similar to the EU's opt-out under Art 4 of Directive (EU) 2019/790) – though such a mechanism should be designed carefully to avoid undermining the utility of the exception for AI training at scale.

57 In conclusion, by enacting a clear and balanced TDM exception, South Korea can provide the legal certainty and flexibility needed to foster its AI industry while respecting the interests of content creators. Such a provision would place South Korea in harmony with international best practices evidenced by Japan, Singapore, and the EU, and *sidestepping the protracted fair use litigation* seen in the US and the chilling effect of unclear rules. It would affirm that using copyrighted works to train AI is a socially and economically beneficial use that copyright law should *enable*, not inhibit. By codifying a TDM exception, South Korea can strike the proper normative balance: safeguarding authors' legitimate market

and moral interests and unleashing AI-driven creativity and knowledge for the greater public good. The analysis in this article thus culminates in a normative call to action: South Korea *must* update its copyright law to explicitly permit text and data mining for AI, under fair conditions, and to ensure that its legal system keeps pace with technological advancement and continues to promote progress in the arts and sciences in the age of AI.