

**P-24029/34/2025-IPR-VII**

भारत सरकार / Government of India

वाणिज्य एवम उद्योग मंत्रालय / Ministry of Commerce & Industry

औद्योग संवर्धन और आंतरिक व्यापार विभाग / Department for Promotion of Industry and Internal Trade

(आई.पी.आर.-प्रतिलिप्याधिकार, अभिकल्प और सीआईपीएएम अनुभाग/ IPR – Copyrights, Design and CIPAM Section)

वाणिज्य भवन, नई दिल्ली-110001 / Vanijya Bhavan, New Delhi-110 011

**Dated: December 8, 2025**

### **Notification**

I am directed to state that Department for Promotion of Industry and Internal Trade formed a committee to examine the intersection of *Generative Artificial Intelligence and Copyright* (hereinafter referred to as "**Committee**"). The Committee was tasked with evaluating whether the existing legal framework on copyright adequately addresses the issues raised by this new technology or amendments to the law are required, and to give its recommendations. The Committee was also tasked with preparing a working paper outlining the Committee's analysis and recommendations.

2. Accordingly, the Working Paper - Part I prepared by the Committee on the issues relating to the use of copyright-protected works for training of AI systems is enclosed herewith for consultation with public and stakeholders.

3. The comments/feedback, if any, may be provided to this Department on email id "**ipr7-dipp@gov.in**" within 30 days of the publication of this letter.

Encl. As above

**(सिमरत कौर /Simrat Kaur)**

निदेशक/Director

To

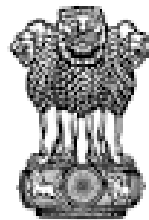
All the stakeholders

# **Working Paper on Generative AI and Copyright**

## **Part 1**

**ONE NATION ONE LICENSE ONE PAYMENT**

***Balancing AI Innovation and Copyright***



सत्यमेव जयते

DEPARTMENT FOR PROMOTION OF  
INDUSTRY AND INTERNAL TRADE  
MINISTRY OF COMMERCE & INDUSTRY  
GOVERNMENT OF INDIA

**DECEMBER 2025**

Table of Content	
Topic	Page
List of Abbreviations	–
Executive Summary	–
<b>1. Introduction</b>  1.1 What is Artificial Intelligence (AI)? 1.2 India and AI 1.3 GenAI and Copyright – Identification of Issues 1.4 Objectives of the working paper 1.5 Need for a balanced framework	1-13
<b>2. GenAI and Copyright Issues on the Input Side – The current legal framework on copyright</b>  2.1 Overview 2.2 Infringement 2.3 Fair Dealing	13-22
<b>3. Position in Other Jurisdictions</b>  3.1 United States 3.2 Japan 3.3 United Kingdom 3.4 European Union 3.5 Singapore	22-36
<b>4. Assessment of Various Regulatory Models</b>  4.1 Voluntary Licensing via Direct licensing agreements 4.2 Text and Data Mining Exception 4.3 Collective Licensing and Extended Collective Licensing	36-55
<b>5. Proposed Policy Framework</b>  5.1 Hybrid Model – Concept 5.2 Nature of license: Mandatory Blanket License 5.3 Collecting Entity: Copyright Royalties Collective for AI Training 5.4 Royalty Rates and Rate Setting Mechanism 5.5 Distribution of Royalties 5.6 Grievance Redressal and Monitoring	55-79

5.7 Burden of Proof 5.8 Injunction Remedy 5.9 Potential Benefits and Suitability of Hybrid Model	
<b>6. Conclusion</b>	79
<b>7. Annexure A</b>	80-81
<b>8. Annexure B</b>	82-83
<b>9. Annexure C</b>	84-88
<b>10. Annexure D</b>	89-91
<b>11. Annexure E</b>	92-115



## List of Abbreviations

AI	Artificial Intelligence
CMA	Cine Musicians Association
CL	Collective licensing
CMOs	Collective Management Organisations
CS	Copyright Society
CDA	Computational Data Analysis
CDPA	Copyright, Designs and Patents Act, 1988
CDSM Directive	Copyright in the Digital Single Market Directive
CRCAT	Copyright Royalties Collective for AI Training
CJEU	European Court of Justice
DNPA	Digital News Publishers Association
DPIIT	Department for Promotion of Industry and Internal Trade
DUAA	The Data (Use and Access) Act, 2025
EoI	Expression of Interest
ECL	Extended Collective Licensing
FRT	Facial recognition technology
Gen(AI) / GenAI	Generative Artificial Intelligence
IADI	IndiaAI Application Development Initiative
IBDF	Indian Broadcasting & Digital Foundation

IMI	Indian Music Industry
ICRIER	International Council for Research on International Economic Relations
IFRRO	International Federation of Reproduction Rights Organisations
JCO	Japanese Copyright Office
NCG	National Cancer Grid
NIELIT	National Institute of Electronics & Information Technology
NBDA	News Broadcasters & Digital Association
PSA	Principal Scientific Advisor
SDG1	Sustainable Development Goal 1
TPM	Technological Protection Measures
TDM	Text and Data Mining
WIPO	World Intellectual Property Organisation

## Executive Summary

### ONE NATION ONE LICENSE ONE PAYMENT

#### *Balancing AI Innovation and Copyright*

Generative Artificial Intelligence has immense potential to transform the world for better, underscoring the need for a regulatory environment that supports its development. However, the processes by which the AI Systems are trained, often using copyrighted materials without authorization from copyright holders and the nature of the outputs that they generate, have sparked an important debate around copyright law. The central challenge lies in how to protect the copyright in the underlying human-created works, without stifling technological advancement. To address this, a balanced regulatory architecture is required to preserve the integrity of the creative ecosystem in the country while encouraging AI innovation.

Recognizing the growing need for deliberations on emerging issues pertaining to AI Systems and copyright, Department for Promotion of Industry and Internal Trade (DPIIT) formed a committee on April 28, 2025, which was tasked with identifying the issues raised by AI Systems, examining the existing regulatory framework, assessing its adequacy, recommending changes if necessary, and preparing a working paper for consultation with stakeholders. The Committee identified and deliberated on the following legal issues: 1) the legal issues relating to the use of copyright-protected works as training data for GenAI systems; and 2) the copyrightability and authorship of GenAI-generated outputs, including the applicability of moral rights and attribution of liability for infringing outputs. It was decided that the first issue would be covered in Part 1 of the working paper, and the remaining issues would be addressed in Part 2.

The Committee consulted stakeholders from the tech/AI Industry where the majority of stakeholders advocated for a blanket Text and Data Mining (TDM) exception enabling training of GenAI on copyright-protected works. A small group of stakeholders expressed support for a TDM exception with an opt-out right for copyright holders. A separate consultation with the representatives of the content industry was held, wherein all the stakeholders unanimously advocated for a voluntary licensing model.

The Committee examined the global developments, focusing on the legal and policy frameworks in the United States, Japan, the UK, Singapore, and the European Union. The Committee also took note of the pending litigation before the Hon'ble Delhi High Court on this

issue, as well as litigations pending and judgements issued in other jurisdictions, and acknowledged that awaiting finality on such pending litigations may not be optimal.

Various regulatory models were explored by the Committee including Voluntary Licensing, Extended Collective Licensing, Statutory licensing, Text and Data Mining Exception (Blanket) and Text and Data Mining Exception (with an opt-out right for copyright holders). Upon undertaking a detailed analysis to understand the pros and cons of these regulatory options, the majority of the Committee members held the view that all of these options, when applied in traditional form, have significant suitability challenges, as explained in the paper.

The Committee had detailed deliberations on the TDM exception model recommended by the tech industry, however, this approach was not found to be a prudent policy approach. Allowing such an exception under law for commercial purposes would undermine copyright and it would leave human creators powerless to seek compensation for use of their works in AI Training. It was not found to be a wise policy choice, especially for a country like India which has a rich cultural heritage and a growing content industry with immense potential.

Even the exception with opt-out right for copyright holders was found to be insufficient in achieving the necessary balance. While it may offer some comfort to large content industry players who may take leverage from this model, it leaves small creators largely unprotected owing to lack of awareness to opt-out, bargaining power to negotiate, and the mechanisms to see if their content has been scraped despite opt-out. Moreover, opt-out functionality does not prevent downstream reuse once the content is stripped of its metadata and transformed, hence the control lost over the data is irrecoverable. This model may also limit the availability of broad and representative datasets for AI Training, especially if many rights holders choose to opt out. This could compromise the quality of AI Systems. Importantly, this model shifts the burden from content users to content creators and subordinates the exclusivity of rights conferred by copyright to a presumed utility for innovation, enforced through mechanisms that structurally favour large players over individual creators. It's important to recognize that an opt-out right for copyright holders is largely ineffective without a transparency requirement obligating AI Developers to fully disclose the data used for AI Training, with all relevant details. However, such a transparency obligation could be highly burdensome and may hinder AI innovation.



The Committee recognized that **access to large volumes of data and high-quality data is crucial** for AI development. Long negotiations and high transaction costs can hold back innovation, particularly for startups and MSMEs.


In light of all the above, it was decided to adopt a hybrid approach which ensures:


- 1) Availability of all *lawfully accessed* copyrighted content for AI Training as a **matter of right**, without the need for individual negotiations;
- 2) Reduced transaction costs for AI Developers;
- 3) Reduced compliance burden on AI Developers;
- 4) **Fair compensation to copyright holders**;
- 5) Judicial review over royalty rates established;
- 6) Easy and straightforward process of payment to rightsholders;
- 7) Mitigated risk of AI bias and hallucinations; and
- 8) Level playing field for all, including start-ups and small players.


Recognizing the above, a hybrid model was conceptualized by the Committee which was found to offer a balanced solution, addressing many of the concerns associated with the other models. With a majority view, the Committee decided to recommend a mandatory blanket license in favor of AI Developers for the use of all lawfully accessed copyright-protected works in the training of AI Systems, accompanied by a statutory remuneration right for the copyright holders. Under this framework, the rights holders will not have the option to withhold their works from use in the training of AI Systems. A centralized non-profit entity made by the rightsholders and designated by the Central Government under the statute would be responsible for collecting the payments from the AI Developers. This entity would have Copyright Societies and Collective Management Organizations (CMOs) as its members (one member for each class of works). These member organizations would further allocate the royalties to both their members and non-members provided they register their works for the purpose of receiving royalties related to AI Training. In other words, even the non-members would be eligible to receive royalties. Certain percentage of the revenue generated from AI Systems trained on copyrighted content would be payable as royalties. The royalty rates would be fixed by a committee appointed by the government. By preserving the right of the copyright owners to receive royalties and administering it through a single umbrella organization made by the rights holders and designated by the government, the model aims to provide an easy access to content for AI Developers for AI Training, simplify licensing procedures, reduce transaction costs, ensure fair compensation for rightsholders. It offers a single window for AI Developers to gain access to copyrighted works for AI Training.

Through its submissions to the Committee dated 17 August 2025, Nasscom expressed its dissent regarding the hybrid approach supported by other Committee members. In its submissions, Nasscom recommended *“Text and Data Mining (TDM) for both commercial and non-commercial purposes where access is lawful, and a good faith knowledge safeguard is met, solely for the training and input processing stage of machine learning. Rightsholders should be provided clear statutory protection against TDM in two complementary ways. For content that is publicly accessible online (freely accessible without paywalls, logins, or other access restrictions), rightsholders should be able to reserve their works from TDM through a machine readable opt out, at the point of availability. For content that is not publicly accessible, rightsholders should be able to reserve their works from TDM through contract or licence terms”*. The Committee members from Nasscom put forth these views.


Considering the perspectives of the majority of Committee Members who found alternative regulatory options, including the TDM model, insufficient to achieve the necessary balance and therefore endorsed the Hybrid Model detailed in the current working paper, the Committee presents the attached working paper (Part 1) for stakeholders’ feedback.


  
Ms. Himani Pande  
(Chairperson)

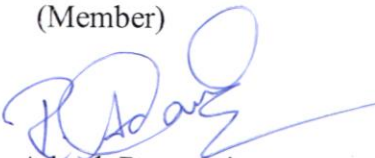
  
Ms. Simrat Kaur  
(Member & Convenor)

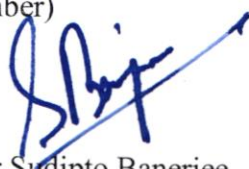
  
Mr. Ameet Datta  
(Member)

  
Prof. Raman Mittal  
Member

  
Mr Chockalingam M  
Member

  
Mr. Anurag Kumar  
(Member)

  
Mr. Adarsh Ramanujan  
(Member)

  
Mr Sudipto Banerjee  
Member

## 1. INTRODUCTION

Disruptive technologies that introduce profound changes to human life emerge only once in several decades. Since these technologies offer both opportunities as well as challenges, it becomes necessary to carefully assess their pros and cons, and regulate their use responsibly. This is how tech innovation puts existing legal frameworks to the test and copyright law is no exception. Copyright has had a complex relationship with technology. From the arrival of the printing press to the dawn of the dot-com era, copyright had to reinterpret its principles to remain relevant. However, throughout this journey of digital evolution, nothing has been able to shake the foundation of copyright. Its capacity to protect original human expression while enabling access to information has proven resilient. As generative artificial intelligence poses fresh challenges, the dialogue between copyright and technology enters a new phase, highlighting the need for careful deliberations. We need to embrace this new technology which carries huge potential, cultivate an environment conducive to its development, and address the issues that it poses to copyright without compromising the core values of copyright law. Confidence<sup>1</sup> in the enduring flexibility of copyright law that guided us during the digital revolution should continue to guide us in the present context as well.

### 1.1 What is Artificial Intelligence (AI)?

As per the white paper released by Niti Ayog of India in 2018, “*AI is a constellation of technologies that enable machines to act with higher levels of intelligence and emulate the human capabilities of sense, comprehend and act*”<sup>2</sup>. According to an issues paper published by the World Intellectual Property Organisation (WIPO), artificial intelligence is “*a discipline of*

---

<sup>1</sup> See the report on 1993 symposium on ‘*The Impact of Digital Technologies on Copyright Law*’ held at Harvard Law School, and sponsored by the World Intellectual Property Organization where the view that digital technologies pose no threat to copyright law received substantial support. Report is available at <https://tind.wipo.int/record/20125?ln=en&v=pdf>, last accessed on August 2, 2025.

<sup>2</sup> See the white paper titled ‘*National Strategy for Artificial Intelligence*’ available at <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>, last accessed on August 2, 2025. The paper further says – “*The natural language processing and inference engines can enable AI systems to analyse and understand the information collected. An AI system can also take action through technologies such as expert systems and inference engines or undertake actions in the physical world. These human capabilities are augmented by the ability to learn from experience and keep adapting over time. AI systems are finding ever-wider application to supplement these capabilities across enterprises as they grow in sophistication.*”



*computer science that is aimed at developing machines and systems that can carry out tasks considered to require human intelligence, with limited or no human intervention”*.<sup>3</sup>

Generative Artificial Intelligence (GenAI) is a sub-class of AI which relies on *“sophisticated machine learning models called deep learning models algorithms that simulate the learning and decision-making processes of the human brain”*<sup>4</sup>. In simple words, it is said to be *“catch-all name for a massive ecosystem of loosely related technologies, including conversational text chatbots like ChatGPT, image generators like Midjourney and DALL-E, coding assistants like GitHub Copilot, and systems that compose music, create videos, and suggest molecules for new medical drugs”*<sup>5</sup>.

## 1.2 India and AI

### 1.2.1 National Strategy on AI

Artificial Intelligence holds out the potential of immense benefits for mankind. It can revolutionise healthcare by enhancing patient care and accurate diagnosis. Quality education can be made more accessible and inclusive with the use of AI. It can support disaster management, boost productivity and research, streamline and enhance government service delivery to citizens, leading to a far more sustainable and efficient world. According to the United Nations, Sustainable Development Goal 1 (“SDG1”) ‘No Poverty - end poverty in all its forms everywhere’ is not on track and *“if current trend persists, estimated 575 million people will still be living in extreme poverty by 2030”*<sup>6</sup>. AI supported research and innovation may generate benefits which *“could trickle down to SDG1 via the creation of new products or*

<sup>3</sup> See the issues paper on IP Policy and AI available at [https://www.wipo.int/edocs/mdocs/mdocs/en/wipo\\_ip\\_ai\\_2\\_ge\\_20/wipo\\_ip\\_ai\\_2\\_ge\\_20\\_1\\_rev.pdf](https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf), last accessed on August 2, 2025.

Under Article 3(1) of the EU Artificial Intelligence Act, an 'AI system' is defined as a *“machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, inferring from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that influence physical or virtual environments”*. See [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689), last accessed on August 21, 2025.

Also see the definition by Google Cloud at <https://cloud.google.com/learn/what-is-artificial-intelligence>, last accessed on August 21, 2025 – *“Artificial intelligence (AI) is a set of technologies that empowers computers to learn, reason, and perform a variety of advanced tasks in ways that used to require human intelligence, such as understanding language, analyzing data, and even providing helpful suggestions.”*

<sup>4</sup> See <https://www.ibm.com/think/topics/generative-ai>, last accessed on August 21, 2025.

<sup>5</sup> Katherine Lee, A. Feder Cooper, James Grimmelmann, ‘Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain’, available at <https://arxiv.org/pdf/2309.08133>, pg. 4, last accessed on August 2, 2025.

<sup>6</sup> See ‘AI for Good Innovate for Impact Interim Report 2025’, available at <https://social.desa.un.org/sdn/ai-for-good-impact-report>, last accessed on August 10, 2025.



*services that are more affordable or better suited to the needs of the most vulnerable communities”.*<sup>7</sup> As per a Goldman Sachs report, GenAI could “drive a 7% (or almost \$7 trillion) increase in global GDP and lift productivity growth by 1.5 percentage points over a 10-year period”.<sup>8</sup>

Recognising the significant economic and social benefit potential that AI holds for the world and the instrumental role that India can play in the global AI landscape, Hon’ble Finance Minister, in the budget speech of 2018-19, mandated NITI Aayog to establish the National Program on AI. A paper on the National Strategy on Artificial Intelligence<sup>9</sup> was released by Niti Ayog in 2018, which articulated India’s vision on AI. It laid the roadmap to leverage AI in five public sectors, i.e. healthcare, education, agriculture, smart cities, and smart mobility, under the theme of #AIforAll. This paper emphasized the need for the adoption of safe, ethical and responsible AI. The necessity to boost research in AI and upskill the workforce was highlighted. It also recommended “building an attractive IP regime for AI innovation” in India.

Thereafter, stakeholder consultations were initiated in collaboration with the World Economic Forum, and an approach paper was released in two parts on Principles for Responsible AI. The first part<sup>10</sup> was published in February 2021, which examined the system considerations like security risk, privacy risks, difficulty in assigning responsibility, inherent bias of AI, etc., and the societal considerations like impact on jobs and “malicious psychological profiling”. It also identified seven principles, i.e. safety and reliability, inclusivity and non-discrimination, equality, privacy and security, transparency, accountability, and protection and reinforcement of positive human values, as a guiding framework for stakeholders to leverage and manage AI.

The second part<sup>11</sup> of the paper which was published in August 2021, elaborated on the mechanisms for operationalising the above seven principles. It highlighted the need for government intervention to strengthen below-mentioned pillars:

---

<sup>7</sup> *Ibid*

<sup>8</sup> See ‘Generative AI could raise global GDP by 7%’, available at <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent.html>, last accessed on August 21, 2025.

<sup>9</sup> See <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>, last accessed on August 2, 2025.

<sup>10</sup> See Approach Document for India - Part 1 ‘Principles for Responsible AI’ available at <https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf>, last accessed on August 2, 2025.

<sup>11</sup> See Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI available at <https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf>, last accessed on August 2, 2025.

- a) Regulatory interventions towards creating a trusted AI ecosystem
- b) Policy interventions to enable responsible AI adoption
- c) Awareness and capacity building on responsible AI in the public sector
- d) Facilitate alignment of procurement mechanisms with responsible AI principles

On the policy making, the paper said that the government needs to play the lead role to:

- a) Manage and update the principles for responsible AI in India
- b) Research into technical, legal, policy and social aspects of responsible AI in India
- c) Enable access to data, responsible AI tools and techniques
- d) Develop India's perspectives on responsible AI

In the year 2022, the third report<sup>12</sup> was released wherein Facial recognition technology (FRT) was taken as the use case for examining the responsible AI principles and operationalisation mechanism proposed in earlier papers.

### **1.2.2 IndiaAI Mission**

In March 2024, the Government of India approved an allocation of over Rs 10,300 crore for the **IndiaAI Mission**<sup>13</sup> marking a significant step towards bolstering India's AI ecosystem. The aim was to build a precise and cohesive strategy for the AI ecosystem catalysing AI innovation through strategic programs and partnerships across the public and private sectors. By democratizing computing access, improving data quality, developing indigenous AI capabilities, attracting top AI talent, enabling industry collaboration, providing startup risk capital, ensuring socially impactful AI projects and bolstering ethical AI, the mission aims to drive responsible and inclusive growth of India's AI ecosystem.

The IndiaAI Mission comprises of 7 pillars namely *IndiaAI Compute Capacity*, *IndiaAI Innovation Center*, *AIKosh IndiaAI Dataset platform*, *IndiaAI Application development initiatives*, *IndiaAI FutureSkills*, *IndiaAI Startup Financing*, and *Safe & Trusted AI*. Below paras capture the information about these seven pillars, updated till July 30, 2025<sup>14</sup>.

<sup>12</sup> See Responsible AI for All: Adopting the Framework – A use case approach on Facial Recognition Technology, available at [https://www.niti.gov.in/sites/default/files/2022-11/Ai\\_for\\_All\\_2022\\_02112022\\_0.pdf](https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf), last accessed on August 2, 2025.

<sup>13</sup> See the report available at <https://www.pib.gov.in/PressReleasePage.aspx?PRID=2012375>, last accessed on August 2, 2025.

<sup>14</sup> See more updated details about all the seven pillars at <https://indiaai.gov.in/>, last accessed on August 3, 2025.

The *IndiaAI compute pillar* has been building a high-end, scalable AI computing ecosystem to cater to the increasing demands of the rapidly expanding AI start-ups and research ecosystem in India. This is aimed at delivering Compute as a Service to address India's dedicated AI computing needs across various sectors. **38,231 GPUs** are being made accessible at subsidized rates through the Compute portal facilitating increased accessibility and affordability for tech developers. A significant development includes the proposed construction of a **government-controlled GPU cluster** set to house **3,000 next-generation GPUs**, which will cater primarily to sovereign and strategic requirements.

The *IndiaAI Application Development Initiative* (IADI) is an initiative to develop, scale and promote the adoption of diverse AI solutions with the potential for catalyzing large-scale socio-economic transformation. Launched in 2024, the first **Innovation Challenge** focused on pivotal sectors such as climate change, disaster management, healthcare, agriculture, governance, and assistive technologies for learning disabilities. 30 applications have reached the prototyping stage, with a second iteration of the challenge planned in conjunction with the Ministry of Education. Additionally, sector-specific hackathons have been organised with the Indian Cybercrime Coordination Centre, the Geological Survey of India and National Cancer Grid (NCG) to further encourage focused AI solution development.

*AIKosh* has been conceptualized as a unified data platform, integrating datasets from all existing data platforms as well as onboarding non-government data contributors and providing new-age AI-centric features. The platform has made strides in democratizing data access for AI development by onboarding **1,400 datasets and 217 AI models** from more than **34 entities** across **20 sectors**. A notable feature of this platform is its permission-based access, allowing data contributors to maintain control over downloads, thereby balancing data sharing with privacy concerns. The library, featuring over **23 use cases**, along with **13 toolkits** providing development utilities for project integration, has seen more than 17,000 downloads. MEITY is engaging with multiple ministries/departments/state government agencies to onboard datasets into the IndiaAI Datasets Platform. Sectoral workshops as well as the signing of data sharing agreements are underway.

*IndiaAI Foundation Models* pillar is working to catalyse AI development by supporting **15 to 20 startups** in **developing LLMs and SLMs**. These startups receive full compute usage and additional funding covering up to 25% of compute expenditure. The financial support is structured as a mix of grants - 40% of compute costs and 60% equity to be realized at future valuation rounds. Over **500 proposals** were received, of which four - **Sarvam AI, Soket AI,**

**Gnani AI, and Gan AI** - were selected in the first phase to develop India's foundation models. Further proposals are under consideration.

The *FutureSkills pillar* is a cornerstone initiative under the IndiaAI Mission that focuses on building India's AI talent ecosystem and democratizing access to AI education nationwide. This comprehensive program aims to create a robust pipeline of AI-skilled professionals through strategic interventions at various educational levels - from undergraduate to doctoral studies.<sup>15</sup>

**34 institutions** have partnered with IndiaAI to **onboard PhD students**, with 26 PhD applications approved thus far. Additionally, **570 Data and AI labs will be established** across the country, of which **20 labs** in collaboration with the National Institute of Electronics & Information Technology (NIELIT) are ready for launch. **174 ITIs/Polytechnics** across 27 States/UTs have been approved for setting up Data and AI Labs. These labs will deliver foundational and sectoral AI courses tailored to India's diverse development needs.

*IndiaAI Startup Financing* addresses the critical need for risk capital across the entire lifecycle of AI startups from prototyping to commercialization. IndiaAI Mission in collaboration with Station F (Paris, France) and HEC Paris, announced an acceleration program for Indian AI startups. This program aims to support **10 Indian AI startups** in scaling globally by leveraging the European market's vast opportunities. A call for applications for Indian Startups working in the domain of AI was launched in March 2025. From this, 10 startups have been shortlisted and initiated into an acceleration program in April. A call for proposals to establish state-level Centers of Excellence in AI alongside industry, academia, and research partners, received 29 submissions from 22 states and union territories, aiming to further strengthen local AI ecosystems.

Recognising the need for adequate guardrails to advance the responsible development, deployment, and adoption of AI, **Safe & Trusted AI** pillar is working on Responsible AI Projects to enable the development of indigenous instruments of AI governance, which are contextualised to India's social, cultural, economic, and linguistic diversity. The first Expression of Interest (EoI) was floated with 10 thematic priorities for Responsible AI projects

---

<sup>15</sup> Number of Fellowships have been enhanced to support 500 PhD fellows, 8,000 undergraduates, and 5,000 postgraduate students. IndiaAI Fellowships have been expanded to include eligible students from Medicine, Law, Commerce, Business, and Liberal Arts with funding support of Rs 1 lakh for undergraduates in their final year and Rs 2 lakhs for Master's students for 2 years. The IndiaAI Research Fellowships for PhD scholars are aligned with the Prime Minister Research Fellowship, offering support of up to ₹55 lakh per fellow. As part of the initiative, around 200 students from B.Tech and M.Tech programs have already received fellowships in the first year.

**and 8 projects** across various themes were selected. **The 2nd round of Expression of Interest (EoI) launched in December 2024**, covering themes such as Watermarking & Labelling, Ethical AI Frameworks, AI Risk Assessment & Management, Stress Testing Tools and Deepfake Detection Tools, received 400+ applications. Final evaluations are currently in progress.

IndiaAI has setup the **IndiaAI Safety Institute** under the Safe and Trusted Pillar to address AI risks & safety challenges. The Institute, incubated by IndiaAI Mission will be set up on a hub and spoke model with various research and academic institutions, private sector partners joining the hub and taking up projects under the Safe and Trusted AI Pillar of IndiaAI Mission. The EoI to onboard the partners has received 90+ applications, currently under consideration.

Also, India will be hosting the **Artificial Intelligence (AI) Impact Summit** in February 2026. This Summit will be the fourth in the series of Global AI Summits, following the AI Safety Summit at Bletchley, the AI Seoul Summit, and the AI Action Summit, co-chaired by France and India. The Summit marks a “*defining global inflection point — transitioning from dialogue to demonstrable impact*”<sup>16</sup>. Anchored in the “*principles of People, Planet, and Progress, it envisions a future where AI advances humanity, fosters inclusive growth, and safeguards our shared planet*”<sup>17</sup>. There would be 7 flagship events. More than 100 countries have been engaged. 200+ pre-summit events have been held so far. Deliberations at the Summit will be organized by working groups around seven interconnected themes (Chakras)<sup>18</sup>.

### **1.2.3 AI Governance**

Pursuant to the need to develop a suitable regulatory framework, an Advisory Group was constituted under the chairmanship of the Principal Scientific Advisor (PSA) of India, with representatives from different ministries. This group was tasked to undertake development of an ‘AI for India-Specific Regulatory Framework’. Under the guidance of this Advisory Group, Ministry of Electronics and Information Technology, on November 9, 2023, constituted a sub-committee to examine key issues related to AI governance in India, conduct a gap analysis of existing frameworks, and offer recommendations for a comprehensive approach to ensure the trustworthiness and accountability of AI Systems in India.

<sup>16</sup> See <https://impact.indiaai.gov.in/>

<sup>17</sup> *Ibid*

<sup>18</sup> See <https://impact.indiaai.gov.in/working-groups>

The aforesaid sub-committee released its report<sup>19</sup> in 2024 (hereinafter referred to as “Subcommittee Report of 2024”). Public consultation on the report received over 2,500 submissions from government bodies, academic institutions, think tanks, industry associations, private sector organisations, and individual stakeholders. Thereafter, a drafting committee was formed to develop India’s new AI governance framework. In November 2025, MEITY published India AI Governance Guidelines<sup>20</sup> with a dual purpose “*to maximise the developmental and economic gains from AI by fostering innovation and adoption at scale, and to mitigate associated risks in a manner that safeguards individuals, protects societal interests, and upholds democratic values*”. The Guidelines provide “*a framework for the development and deployment of safe, trustworthy, responsible, inclusive and accountable AI systems, such that cutting-edge AI can be harnessed in concert with other transformative technologies to anchor the long-term growth, resilience and sustainability of India’s digital ecosystem*”. Part 1 of the Guidelines sets out seven sutras of Trust, People First, Innovation over Restraint, Fairness & Equity, Accountability, Understandable by Design and Safety, Resilience & Sustainability which are designed to be “*technology agnostic and applicable across all sectors*”. Part 2 examines issues and offers recommendations across three domains - enablement (infrastructure, capacity building), regulation (policy & regulation, risk mitigation) and oversight (accountability, institutions). Part 3 gives action plan to operationalize those recommendations, and Part 4 provides practical guidelines for industry and regulators.

### **1.3 GenAI and Copyright – Identification of issues**

The Subcommittee Report of 2024 had identified two primary legal concerns concerning AI and Copyright - first, the use of copyrighted material as input for AI training; and second, the copyrightability of works generated by AI models. It raised key policy questions around whether AI Systems<sup>21</sup> should be permitted to train on large datasets containing copyrighted works without prior consent from rightsholders, and under what conditions such use could be deemed lawful. It also highlighted the lack of clarity regarding the rights of creators concerning AI training and output, and the potential need to interpret or redefine the scope of such rights.

<sup>19</sup> Report on AI Governance Guidelines Development, available at <https://indiaai.s3.ap-south-1.amazonaws.com/docs/subcommittee-report-dec26.pdf>, last accessed on August 2, 2025.

<sup>20</sup> Available at <https://indiaai.s3.ap-south-1.amazonaws.com/docs/guidelines-governance.pdf>, last accessed on November 19, 2025.

<sup>21</sup> Notable generative AI tools are ChatGPT, DALL·E; ERNIE Bot by Baidu; Copilot by Microsoft; Stable Diffusion by Stability AI; Gemini by Google; Midjourney, Boomy, and Suno.

The report underscored that resolving these issues is necessary to improve legal certainty and enable consistent application of the law by users and public authorities.

In line with these recommendations, the Department for Promotion of Industry and Internal Trade formed a committee constituted of government officials and external experts to examine the intersection of *generative artificial intelligence and copyright* through a legal and policy lens (hereinafter referred to as “**Committee**”/ “**DPIIT Committee**”). The Committee was tasked with assessing whether the existing legal framework on copyright sufficiently addresses the issues raised by this new technology or whether amendments to the law are needed, and to give its recommendations. The Committee examined the extant framework, global developments and internationally published literature on the subject. The Committee held detailed consultations with various stakeholders. List of the stakeholders who were consulted is provided at **Annexure A**.

#### **1.4. Objective of this working paper**

The present working paper has been prepared based on the deliberations of the Committee and its recommendations. It builds on the findings of the Subcommittee Report of 2024 and issues identified therein. It assesses the adequacy of existing legal provisions, undertakes an analysis of comparative international approaches, and proposes potential regulatory reforms that align with India’s constitution, the country’s international obligations, the goals of the IndiaAI Mission, and the principles of natural justice.

The objective is to facilitate a balanced and transparent framework that safeguards the rights of content creators while enabling responsible Gen AI innovation and equitable access to technology. The working paper examines the following two main issues:

a) Use of copyrighted materials as input

Legal and policy considerations regarding the use of copyrighted content in training of AI Systems.

b) Copyright claims over AI-generated output

- a. Determining the copyrightability of AI generated works
- b. Identifying authorship in AI-generated content
- c. Applicability and scope of moral rights in relation to AI-generated works
- d. Attribution of liability for infringing outputs

The paper is organised into **two** parts. The first part examines the legal and regulatory issues pertaining to the large-scale unlicensed use of copyrighted materials for training of generative artificial intelligence tools (hereinafter referred to as “**AI Systems**”<sup>22</sup>). The second part of the paper will cover the copyright status of AI-generated outputs, assessing whether such works are eligible for copyright protection, who should be the author of copyright therein, whether moral rights should apply, and also the attribution of liability in the event of infringement.

The paper seeks to initiate a dialogue on these issues and explore potential policy and legal responses that promote a fair, balanced, transparent, and innovation-friendly framework that safeguards the rights of creators while supporting the continued growth of India’s generative AI ecosystem.

### 1.5 Need for a balanced framework

The benefits of generative AI, as outlined in section 1.2.1 of this paper, highlight its huge potential. A report by McKinsey released in 2023 states that “*Generative AI’s impact on productivity could add trillions of dollars in value to the global economy. Our latest research estimates that generative AI could add the equivalent of \$2.6 trillion to \$4.4 trillion annually across the 63 use cases*”<sup>23</sup>. Alongside Gen AI, it is crucial to note the importance and contributions of the creative economy. As per UNCTAD’s global survey, the economic contributions of the creative economy, across different countries, range from 0.5% to 7.3% of GDP, employing between 0.5% to 12.5% of the workforce.<sup>24</sup> Citing FICCI’s report, this UNCTAD report states that India’s film industry recorded box office revenues of about US\$ 1.4 billion in 2023.<sup>25</sup> The M&E sector in India crossed US\$ 29.4 billion in 2024, contributing 0.73% to India’s GDP. The sector is set to grow at over 7% annually reaching INR 3.07 trillion (US\$ 36.1 billion) by 2027, which is more than India’s GDP growth rate.<sup>26</sup> A recent report

---

<sup>22</sup> Any reference to ‘**AI Systems**’ in this paper refers specifically to Generative Artificial Intelligence Tools/Systems. Any reference to ‘**AI training**’/ ‘**AI Training**’ shall mean training of such AI Systems, and **AI Developers** shall mean the developers of such AI Systems.

<sup>23</sup> Report on ‘The economic potential of generative AI’, available at <https://www.mckinsey.com/~media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf>, last accessed on August 9, 2025.

<sup>24</sup> Creative Economy Outlook 2024, available at <https://unctad.org/publication/creative-economy-outlook-2024>, last accessed on August 3, 2025.

<sup>25</sup> *Ibid*

<sup>26</sup> See ‘A Studio Called India’, available at <https://www.ey.com/content/dam/ey-unified-site/ey-com/en-in/pdf/2025/ey-a-studio-called-india-content-and-media-services-for-the-world.pdf>, last accessed on September 7, 2025.



released by an Indian consulting firm states that *“In 2021, the creative economy contributed \$43 billion to the national GDP, representing 4.3% of economic output. This sector is projected to grow at a CAGR of 13% between 2021 and 2026 making it \$80 billion USD economy.”*<sup>27</sup>

Human creativity and AI development are, therefore, both vital pillars of the Indian economy.

India has always celebrated human creativity, drawing inspiration from its rich cultural heritage and young population. With the rapid growth of the middle class, improved access to education, and heavy digital penetration<sup>28</sup>, content creation has seen a significant rise across the country in the last few years. From regional storytelling to digital art and music, there is a massive creative surge that reflects the deep cultural diversity and creativity found across different parts of India.

Importantly, India’s creative economy extends beyond formal creative sectors. The informal sectors, comprising folk arts, folk music, and crafts, also thrive, which are preserved through oral traditions passed down to young generations, and contribute significantly. According to a recent report<sup>29</sup> by the International Council for Research on International Economic Relations (ICRIER), India’s informal music industry alone employs over 1.4 crore people.

Both the organised and unorganised sectors, thus, constitute a vast creative landscape that offers livelihood to millions of people. At the heart of this ecosystem lies human creativity, and copyright law is its foundational pillar. It is therefore crucial to recognise and reward human creators, especially those from underrepresented sectors, to achieve long-term sustainability of the country’s creative economy.

Sharing his vision for India’s creative industry, Hon’ble Prime Minister, in his address at the WAVES Summit in Mumbai stated, *“This is the period of rise of Orange Economy in India.*

<sup>27</sup> See ‘Shaping Education to nurture the \$80 Billion Creative Economy’, available at [https://primuspartners.in/docs/documents/Shaping%20%20Education%20to%20nurture%20the%20\\$80%20Billion%20Creative%20Economy\\_updated.pdf](https://primuspartners.in/docs/documents/Shaping%20%20Education%20to%20nurture%20the%20$80%20Billion%20Creative%20Economy_updated.pdf), last accessed on August 16, 2025. As per this report, “creative economy represents the intersection of culture, technology, and innovation.” It encompasses industries such as media and entertainment, design and visual arts, gaming and animation, cultural heritage and performing arts.

<sup>28</sup> According to an IAMAI and Kantar (a market research firm) report, “India is set to have over 900 million internet users by 2025, with a majority from rural areas”, See <https://economictimes.indiatimes.com/tech/technology/india-to-cross-900-million-internet-users-this-year-says-iamai-report/articleshow/117290089.cms?from=mdr>

Also see [https://www.business-standard.com/industry/news/india-to-exceed-900mn-internet-users-by-2025-125011600669\\_1.html](https://www.business-standard.com/industry/news/india-to-exceed-900mn-internet-users-by-2025-125011600669_1.html), last accessed on August 28, 2025.

<sup>29</sup> The Untold Potential of India’s Informal Music Industry, available at <https://icrier.org/research/assessing-the-economic-impact-of-the-recorded-music-industry-in-the-unorganised-sector-in-india/>, last accessed on August 16, 2025.

*Content, Creativity and Culture - these are the three pillars of the Orange Economy. India, along with its billion plus population, is also a country of billion plus stories. Two thousand years ago, when Bharat Muni wrote Natya Shastra, his message was - “Natyaam Bhavayati Lokam”. It means, art gives emotions, feelings to the world. Centuries ago, when Kalidasa wrote Abhijnana-Shakuntalam, India gave a new direction to classical drama. Every street in India has a story, every mountain has a song, every river hums something or the other. If you go to more than 6 lakh villages in India, each village has its own folk, has its own special style of storytelling. This is the right time to Create in India, Create for the World. Today, when the world is looking for new ways of storytelling, India has a treasure of its stories dating back thousands of years. And this treasure is Timeless, Thought-Provoking and Truly Global.*<sup>30</sup>

While Gen AI offers significant benefits and fostering a robust AI development ecosystem is crucial for India, it is equally important to reflect on what we might risk losing in the process, if systems like copyright, which reward human creativity, are compromised. Gen AI development relies on learning from a vast amount of human-created content. If AI Developers continue to train their models on the works created by humans without paying them for the same, we may risk seeing a sharp decline in human-created works, which would, over the years, affect the richness of our cultural and creative landscape. This would, in the long term, impact the AI ecosystem as well, because it risks entering into a self-reinforcing loop where machines will be trained on machine-generated content, lacking depth and emotional resonance. True public good lies in supporting innovation and human creativity together. Societies like India, which have a strong cultural identity and creative spirit, need a balanced framework that enables AI innovation as well as respects creators’ rights. India has the power of data, creativity, technological innovation; therefore, we are uniquely positioned to leverage all these assets in pursuit of the goal of Viksit Bharat by 2047. Hon’ble Prime Minister, during his independence-day speech on August 15, 2025 stated that “*Today is the era of IT, we have the power of data, isn't it the need of the hour? From operating systems to cyber security, from deep tech to artificial intelligence, everything should be our own, on which the strength of our own people is concentrated, we should introduce the power of their capabilities to the world.*”<sup>31</sup>

<sup>30</sup> See the press release available at <https://www.pib.gov.in/PressReleaseDetail.aspx?PRID=2125749> , last accessed on August 21, 2025.

<sup>31</sup> See the press release available at <https://www.pib.gov.in/PressReleaseDetail.aspx?PRID=2156749> , last accessed on August 21, 2025.

On the consumption front, India is adopting generative AI at a great pace. OpenAI CEO Sam Altman recently stated that *“India is our second-largest market in the world after the US and it may well become our largest market”*.<sup>32</sup> Since India represents a significant market for AI Systems and directly contributes to the revenues of AI Developers, there is an added rationale that a portion of such revenue be shared with the creators from India whose works are used in the training of such AI Systems. In light of this, this paper advocates for the development of a balanced framework, and puts forward a proposal in this regard.

## 2. GENAI AND COPYRIGHT ISSUES ON THE INPUT SIDE: EXAMINING THE CURRENT COPYRIGHT FRAMEWORK

### 2.1 Overview

AI Systems ingesting mass volumes of data, including copyrighted works, as training materials, to generate outputs raises complex questions around originality, authorship, copyright ownership, infringement and fair dealing.

On the input side, the key concern is that the copyrighted content is often used in AI training without the license from the rightsholders. This practice risks weakening the economic incentives for humans to develop creative works. To support such an incentive, authors/copyright owners should be able to prevent unauthorized exploitation of their works for commercial purposes. It is also essential to recognise that AI Systems, as they exist today and what may be expected in the future, are here to stay, and access to extensive and diverse materials is necessary to develop effective AI Systems, reduce hallucinations<sup>33</sup> and mitigate bias<sup>34</sup>.

There is an inherently reciprocal relationship between GenAI and copyright. Unregulated use of copyright protected content may devalue human creativity and lead to underproduction of

<sup>32</sup> See <https://timesofindia.indiatimes.com/technology/tech-news/openai-ceo-sam-altman-at-gpt-5-launch-india-is-our-second-largest-marketbut-what-users-are-doing-with/articleshow/123178437.cms>, last accessed on August 24, 2025.

<sup>33</sup> IBM defines AI hallucination as “a phenomenon where, in a large language model (LLM) often a generative AI chatbot or computer vision tool, perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate”. See <https://www.ibm.com/think/topics/ai-hallucinations>, last accessed on September 6, 2025.

<sup>34</sup> See ISO/IEC 22989 available at <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en>, last accessed on September 9, 2025, where ‘Bias’ is defined as “systematic difference in treatment of certain objects, people or groups in comparison to others”.

human-created content, and excessive restrictions can hamper AI innovation, harming overall public benefit. Therefore, finding the right balance is pivotal.

Section 14<sup>35</sup> of the Copyright Act, 1957 (“Act”) provides for exclusive rights to a copyright owner, including the right to reproduction including storage, translation, adaptation, communication to the public, issuing copies of the work to the public, etc. Doing any of these acts without obtaining the license from the copyright owner amounts to infringement under Section 51 of the Act<sup>36</sup>, unless it is covered by any of the exceptions embodied under Section 52 of the Act. There is currently no specific exception under copyright law for text and data mining, nor is there any other exception specifically designed to exempt all AI training activities from potential copyright infringement.

The other critical question is whether the “fair dealing” exception under Section 52(1)(a)<sup>37</sup> of the Act, applies to AI training. Section 52(1)(a) exempts from infringement the fair dealing of a copyright work for the limited purposes such as private or personal use, including research; criticism or review; and reporting of current events. This exception under Section 52(1)(a) is narrowly defined and purpose-specific.

This question of whether an exception to infringement applies, alongside the broader issue of infringement, is currently pending before the Delhi High Court at the interim stage in India’s first major AI and copyright litigation in India – *ANI Media Pvt. Ltd. v. Open AI Inc*<sup>38</sup>. In this case, the ANI sued OpenAI for copyright infringement, *inter alia*, alleging that the latter used ANI’s content to train its AI model without ANI’s permission. This case is currently pending disposal before the Delhi High Court.<sup>39</sup>

---

<sup>35</sup> See Copyright Act, 1957 available at [https://copyright.gov.in/Copyright\\_Act\\_1957/chapter\\_xi.html](https://copyright.gov.in/Copyright_Act_1957/chapter_xi.html), last accessed at August 3, 2025.

<sup>36</sup> *Ibid*

<sup>37</sup> The following acts shall not constitute an infringement of copyright, namely:

(a) a fair dealing with any work, not being a computer programme, for the purposes of—

(i) private or personal use, including research;

(ii) criticism or review, whether of that work or of any other work;

(iii) the reporting of current events and current affairs, including the reporting of a lecture delivered in public.

*Explanation* - The storing of any work in any electronic medium for the purposes mentioned in this clause, including the incidental storage of any computer programme which is not itself an infringing copy for the said purposes, shall not constitute infringement of copyright.

See Copyright Act, 1957 available at [https://copyright.gov.in/Copyright\\_Act\\_1957/chapter\\_xi.html](https://copyright.gov.in/Copyright_Act_1957/chapter_xi.html), last accessed at August 3, 2025.

<sup>38</sup> CS(COMM) 1028/2024

<sup>39</sup> Court has framed following issues in this case:

## 2.2 Infringement

The central question is whether training AI Systems involves exercising any of the exclusive rights given to copyright holders under Section 14 of the Act. One must extract information from millions or billions of works to train AI.<sup>40</sup> A recent survey of text datasets for Large Language Models (LLMs) found that they included over 774.5 terabytes of data from more than 444 datasets.<sup>41</sup> These datasets rely on various sources, including creative works like books, articles, paintings, photographs, and software.<sup>42</sup>

Using expressions such as “datasets” or stating that the training involves extracting “information” from works, often obscures the steps involved. There is a collection phase, so to speak, where books, articles, etc., are sourced from internet scraping or licensing. These are stored at a particular space or location, typically a large server farm. This collection of works or “raw data” then undergoes data preparation, which involves transforming the raw data into formats suitable for machine learning. This process includes organising the data into formats that can be analysed using machine learning algorithms. This results in the creation of training datasets. This process may also involve adding, deleting, or changing parts of the training set.

Building on the idea/expression dichotomy<sup>43</sup>, some argue that training AI and creating datasets do not infringe on copyright because they focus on non-expressive elements of creative works. Legal precedents suggest that no protection be afforded for non-expressive elements of a

---

1) Whether the storage by the defendants, of plaintiff’s data (which is in the nature of news and is claimed to be protected under the Copyright Act, 1957) for training its software i.e., ChatGPT, would amount to infringement of plaintiff’s copyright.

2) Whether the use by the defendants of plaintiff’s copyrighted data in order to generate responses for its users, would amount to infringement of the plaintiff’s copyright.

3) Whether the defendants’ use of plaintiff’s copyrighted data qualifies as ‘fair use’ in terms of Section 52 of the Copyright Act, 1957.

4) Whether the Courts in India have jurisdiction to entertain the present lawsuit considering that the servers of the defendants are located in the United States of America.

<sup>40</sup> Pamela Samuelson, Christopher Jon Sprigman, and Matthew Sag, ‘Comments In Response To The Copyright Office’s Notice Of Inquiry On Artificial Intelligence And Copyright’, at page 6-7, available at <https://ssrn.com/abstract=4976391> or <http://dx.doi.org/10.2139/ssrn.4976391>, last accessed on September 9, 2025.

<sup>41</sup> Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, Lianwen Jin, ‘Datasets for Large Language Models: A Comprehensive Survey’, available at <https://arxiv.org/pdf/2402.18041v1>, pg. 4, last accessed on September 9, 2025.

<sup>42</sup> *Ibid* pg. 10-15.

<sup>43</sup> R.G. Anand v M/s Delux Films & Ors. (1978) 4 SCC 118, at page 140, para 46(1.) - “There can be no copyright in an idea.....and copyright in such cases is confined to the form, manner and arrangement and expression of the idea by the author of the copyrighted work”); Eastern Book Company & Ors. v D.B. Modak & Anr. (2008) 1 SCC 1, at page 90, para 15 (“It is a well-accepted principle of copyright law that there is no copyright in the facts per se, as the facts are not created nor have they originated with the author of any work which embodies these facts”), also page 112, para 57 (“Copyright Act is not concerned with the original idea but with the expression of thought”)

work.<sup>44</sup> The contention is that the training aims to learn statistical relationships, such as patterns and structures, from the works, not to memorize the creative works. The purpose is to enable AI Systems to create new content that mirrors human generated works. The value of training thus lies not in replicating existing data, but in developing an adaptable foundation that enables AI Systems to create entirely new and contextually appropriate content across various tasks and settings.<sup>45</sup> Training of AI Systems cannot constitute infringement as the purpose is to “learn” the underlying features and associations within the training data and not to memorise the snippets of the original expression from the individual works.<sup>46</sup> Thus, it is argued that the process involves extracting information *from* the work rather than the work itself.

Arguably, by law, infringement happens when a work is reproduced or stored, in this case, regardless of the intention behind it. Under this logic, training AI Systems involves copying and storing data, which constitutes infringement. Accordingly, some scholars opine that training requires copying and storing data while it is being prepared, which constitutes infringement.<sup>47</sup> Even on these so-called “statistical relationships”, a US District Court in *Kadrey vs Meta Platforms, Inc.* observed that even these “statistical relationships” are the product of creative expression.<sup>48</sup> Therefore, it is unclear if the purpose of being limited to deriving “statistical relationships” implies non-infringement.

Interestingly, the Report of the U.S. Copyright Office on Copyright and Generative Artificial Intelligence highlights that aside from the initial copy created, the AI training process may involve creating multiple temporary copies as well:

*“[t]he steps required to produce a training dataset containing copyrighted works clearly implicate the right of reproduction. Developers make multiple copies of works by downloading them; transferring them across storage mediums; converting them to*

<sup>44</sup> Mark A. Lemley and Bryan Casey, Fair Learning, 99 Tex. L. Rev. 743, at 774-775, available at <https://ssrn.com/abstract=3528447> and <http://dx.doi.org/10.2139/ssrn.3528447>, last accessed on September 9, 2025.

<sup>45</sup> Written Testimony of Christopher Callison-Burch, before The U.S. House of Representatives Judiciary Committee; Subcommittee on Courts, Intellectual Property, and the Internet; Hearing on Artificial Intelligence and Intellectual Property: Part I Interoperability of AI and Copyright Law; May 17, 2023; available at <https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf>, last accessed on September 9, 2025.

<sup>46</sup> Matthew Sag, ‘Copyright Safety for Generative AI’, 61 HOUS. L. REV. 295, at 302, 313 available at: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4438593](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593), last accessed on September 3, 2025.

<sup>47</sup> Andres Guadamuz, ‘A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs’, GRUR International, Vol. 73, Issue 2, at Pg. 4, available at <https://ssrn.com/abstract=4371204> or <http://dx.doi.org/10.2139/ssrn.4371204>, last accessed on November 19, 2025.

<sup>48</sup> *Kadrey vs Meta Platforms, Inc.* 2025 U.S. Dist. LEXIS 121064, pg. 14.



*different formats; and creating modified versions or including them in filtered subsets. In many cases, the first step is downloading data from publicly available locations, but whatever the source, copies are made—often repeatedly.”<sup>49</sup> The report further says that “The training process also implicates the right of reproduction. First, the speed and scale of training requires developers to download the dataset and copy it to high-performance storage prior to training. Second, during training, works or substantial portions of works are temporarily reproduced as they are “shown” to the model in batches. Those copies may persist long enough to infringe the right of reproduction, depending on the model at issue and the specific hardware and software implementations used by developers.”<sup>50</sup>*

The above view suggests that reproduction is not merely a one-time reproduction but may involve several copies being reproduced, even if some are only temporary. The suggestion from the U.S. Copyright Office appears to be that even these temporary copies could amount to infringement unless exempt under the fair use provision. Even the issue of temporary copies in training raises another debate. Some US cases hold that in order to prove infringement, the infringing copy must fulfil a certain ‘fixation’ standard, i.e., the alleged infringer must know whether the copy is being made (as opposed to knowledge that it is infringing) and the period for which the copy is retained must be something more than ephemeral.<sup>51</sup> Some scholars have followed this jurisprudence to opine that temporary copies created during training may not fulfil this ‘fixation’ standard.<sup>52</sup>

A study commissioned by the EU Parliament, published in July 2025, goes a step further when stating the following:

*“Generative AI models encode expressive works during training, transforming them into vector spaces and model weights. This internalisation allows for later output that*

<sup>49</sup> See <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 26, last accessed on August 21, 2025.

<sup>50</sup> *Ibid* pg. 27

<sup>51</sup> *CoStar Grp., Inc. v. LoopNet, Inc.*, 373 F.3d 544, 546-47 (4th Cir. 2004), at 551. *Cartoon Network LLP, v. CSC Holdings, Inc.*, 536 F.3d 121, 129-30 (2d Cir. 2008). In the audiovisual context, the Second Circuit held in *Cartoon Network LLP* that a brief embodiment of downloaded content for 1.2 seconds was not sufficiently fixed to constitute a copy. *Cartoon Network* argued that the 1.2-second storage of the content was unauthorized copying constituting infringement. The Second Circuit opinion emphasized that fixation requires not only that a work is embodied in a medium from where it can be later retrieved or reproduced, but also that it is embodied for more than a transitory duration.

<sup>52</sup> Jessica Gillotte, ‘Copyright Infringement in AI-Generated Artworks’, July 23, 2019, *UC Davis Law Review*, Vol. 53, No. 5, 2020, available at <https://ssrn.com/abstract=3657423>, pg. 2678, last accessed on August 17, 2025.

*mimics protected expression. Empirical studies confirm that models can memorize and reproduce content verbatim. This process constitutes a functional equivalent of partial reproduction, even where the output is not identical. Even compressed and abstracted representations in model weights can amount to reproductions if they enable the reconstitution of protected elements. This reflects the technology-neutral and functional interpretation of ‘reproduction’ under EU law.”<sup>53</sup>*

While scope of the present paper is limited to examining the process of training of AI Systems (as opposed to the output generated), like the EU study mentioned above, certain other studies draw a link between the training data and model outputs and suggest that the process of training AI Systems involves reproductions and retention of those copies, leading to infringement concerns. It is argued that a model can be said to have “*memorized a piece of training data when it is possible to reconstruct from the model a (near-)exact copy of a substantial portion of that specific piece of training data*”<sup>54</sup>. Models exhibit “*associative memory behaviors, recalling specific training examples when prompted with certain inputs, raising concerns about the legal implications of such behaviours, especially in the context of copyrighted data*”<sup>55</sup>.

Very recently, 42nd Civil Chamber of the Munich I Regional Court<sup>56</sup> ruled in favour of GEMA and against OpenAI companies saying the memorization of training data in the AI Systems and the reproduction of lyrics of songs in the outputs encroaches on the rightsholders’ copyright in such lyrics, and such use is not covered by exceptions on text and data mining. Court held that the recognisable outputs constitute infringements of the right of reproduction and the right of making available to the public.

---

<sup>53</sup> Available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST\\_STU\(2025\)774095\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf), pg. 45, last accessed on August 17, 2025.

<sup>54</sup> A. Feder Cooper and James Grimmelman, ‘The Files are in the Computer – On Copyright, Memorization and Generative Ai’ available at <https://arxiv.org/pdf/2404.12590>, pg. 142, last accessed on August 10, 2025. Also see Matthew Sag and Peter K Yu, ‘Globalization of Copyright Exceptions for AI Training’, Emory Law Journal, Vol. 74, available at [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4976393](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4976393), pg. 1174, last accessed on August 21, 2025. The authors state – “*If someone asks ChatGPT to tell a bedtime story about a hard-boiled detective bear living in Helsinki, the resulting story will owe something to the works of Dashiell Hammett, but it will not be an infringing copy of The Maltese Falcon because literary style and genre are not protectable aspects of copyright. Nevertheless, some researchers have found generative AI models to memorize significant parts of particular works in the training data or learn enough about the subject matter protected by copyright law at a higher level of abstraction, such as copyrightable fictional characters. Such memorization suggests a real possibility of copyright infringement.*”

<sup>55</sup> Tim W. Dornis and Sebastian Stober, ‘Generative AI Training and Copyright Law’, available at <https://arxiv.org/pdf/2502.15858>, pg. 5, last accessed on August 10, 2025.

<sup>56</sup> See <https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php>, last accessed on November 19, 2025.



It also bears mentioning that some scholars believe the issue of infringement in such situations is “not a question that has a yes or no answer” and “there is not a blanket rule which determines so” – each AI system and each set of facts may have to be decided independently.<sup>57</sup>

## 2.3 Fair Dealing

The other question to be considered is whether the existing “fair dealing” exception under Section 52(1)(a) of the Copyright Act, 1957, can be interpreted to protect the training of AI Systems. The discussion on the statutory exception to copyright infringement begins with a distinction between the narrower test of “fair dealing” mentioned in the Indian statute as well as other common law jurisdictions (UK<sup>58</sup>, Canada<sup>59</sup> etc.), on one hand and the broader test of

---

<sup>57</sup> Katherine Lee, A Feder Cooper and James Grimmelmann, “Talkin’ ’Bout AIGeneration: Copyright and the Generative-AI Supply Chain”, available at <https://arxiv.org/pdf/2309.08133>, pg. 148, last accessed on August 16, 2025.

The study says – “Our conclusion is simple. “Does generative AI infringe copyright?” is not a question that has a yes-or-no answer. There is currently no blanket rule that determines which participants in the generative-AI supply chain are copy right infringers. The underlying technologies and systems are too diverse to be treated identically, and copyright law has too many open decision points to provide clear answers. Our hope is that the supply-chain framing provides a clear and precise mechanism for understanding this diversity and, in turn, for reasoning about the various legal consequences.”

<sup>58</sup> 29. Research and private study.

(1) Fair dealing with a... work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement....

(1C) Fair dealing with a...work for the purposes of private study does not infringe any copyright in the work... available at <https://www.legislation.gov.uk/ukpga/1988/48/section/29>, last accessed on November 19, 2025.

30. Criticism, review quotation] and news reporting.

(1) Fair dealing with a work for the purpose of criticism or review, of that or another work or of a performance of a work, does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement unless this would be impossible for reasons of practicality or otherwise and provided that the work has been made available to the public.

available at <https://www.legislation.gov.uk/ukpga/1988/48/section/30>, last accessed on November 19, 2025.

<sup>59</sup> Fair Dealing

Research, private study, etc.

29 Fair dealing for the purpose of research, private study, education, parody or satire does not infringe copyright.

Criticism or review

29.1 Fair dealing for the purpose of criticism or review does not infringe copyright if the following are mentioned:

(a) the source; and

(b) if given in the source, the name of the

(i) author, in the case of a work,

(ii) performer, in the case of a performer’s performance,

(iii) maker, in the case of a sound recording, or

(iv) broadcaster, in the case of a communication signal.

News reporting

29.2 Fair dealing for the purpose of news reporting does not infringe copyright if the following are mentioned:

(a) the source; and

(b) if given in the source, the name of the

(i) author, in the case of a work,

(ii) performer, in the case of a performer’s performance,

(iii) maker, in the case of a sound recording, or

(iv) broadcaster in the case of a communication signal.

‘fair use’ codified under the US copyright law.<sup>60</sup> From the plain language of the Indian statute, the “fair dealing” exception is not worded as broadly as the “fair use” standard employed in the US statute. Canada also has a similar, narrower “fair dealing” exception, and the Canadian Supreme Court has rejected the applicability of the broader American “fair use” test.<sup>61</sup> Viewed this way, it can be opined that one cannot seek refuge under S.52(1)(a) solely by establishing fulfilment of the American “fair use” test without establishing how the use is for purposes specified in Section 52(1)(a). Some High Court rulings in India suggest that the US four-factor test for fair use may only serve as a supplementary tool, rather than replacing the statutory requirements of Section 52 under Indian law.<sup>62</sup> However, other judicial decisions by Division Benches of the Delhi High Court have treated “fair dealing” and “fair use” as interchangeable.<sup>63</sup> This working paper does not intend to address this specific point in further detail except to highlight this debate and emphasise that the developer of an AI System must establish how and why the training process is directed to the purposes mentioned in Section 52(1)(a) as a threshold question.

Proceeding further, *if* one were to apply the US four-factor test for fair use, some US scholars appear to be of the view that the training of AI Systems is likely to constitute “fair use”<sup>64</sup>, and

available at <https://laws-lois.justice.gc.ca/eng/acts/C-42/page-6.html#docCont>, last accessed on November 19, 2025.

<sup>60</sup> Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phone records or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

- (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
- (2) the nature of the copyrighted work;
- (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- (4) the effect of the use upon the potential market for or value of the copyrighted work.

The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

available at <https://www.wipo.int/wipolex/en/text/130040>, last accessed on November 19, 2025.

<sup>61</sup> *Society of Composers, Authors and Music Publishers of Canada v. Bell Canada*, 2012 SCC 36, at paras 25-26, the Supreme Court of Canada, citing the difference in the statutory scheme, held that: (i) the American four-factor test could not be blindly adopted into Canadian law; (ii) the assessment of “fairness” does not arise until and unless it is shown that the defendant’s dealing is for one of the allowable purposes mentioned in the provision.

<sup>62</sup> *Super Cassettes Industries Ltd. v. Hamar Television Network Pvt. Ltd. & Anr.*, ILR (2010) V Delhi 230, at para 11.1, *Rupendra Kashyap v. Jiwan Publishing House*, 1996 (38) DRJ, at para 21. *Syndicate of Press of University of Cambridge & Anr. v. Kasturi Lal & Sons*, 2005 SCC OnLine Del 1448, at paras 6(v)(a) and 10).

<sup>63</sup> *India TV Independent News Service Pvt. Ltd. vs Yashraj Films Pvt. Ltd.* 2012 SCC OnLine Del 4298 at para 6; *Syndicate of the Press of the University of Cambridge on Behalf of the Chancellor, Masters and School v. B.D. Bhandari & Anr.*, 2011 SCC OnLine Del 3215, at paras 35-37; *The Chancellor, Masters & Scholars of University of Oxford & Ors. v. Rameshwari Photocopy Services & Ors.*, 2016 SCC OnLine Del 6229, at paras 31-35, 66.

<sup>64</sup> Fair use in US is tested by the four-factor test given under 17 U.S. Code § 107. These are: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the

thus the training of AI Systems may not constitute infringement.<sup>65</sup> Some disagree and have highlighted instances where the existing fair use exception may not cover the training of AI Systems. This contrary view, for instance, highlights cases in which the training process might cross the line from fair use to infringement, particularly when AI Systems "memorize" the training data rather than simply "learning" from it.<sup>66</sup>

The inherent open-endedness of the "fair use" concept, or any concept of "fairness" for that matter, can lead to divergence of views. For instance, in both *Kadrey vs Meta Platforms, Inc.*,<sup>67</sup> and *Bartz vs PBC*,<sup>68</sup> a summary judgment of "fair use" was issued in favour of the developer of AI System, yet the legal findings were partially inconsistent. In *Kadrey vs Meta Platforms, Inc.*,<sup>69</sup> the Northern District of California held that two of the four factors favoured the developer of AI System but emphasised that the works copied (memoirs, books, autobiographies) were highly expressive and thus, the second factor was in favour of the copyright owner. On the fourth factor, the Court acknowledged a market dilution of the plaintiff's work by 'indirect' instead of direct substitution.<sup>70</sup> Conversely, a different judge from the same Court in *Bartz vs PBC*<sup>71</sup>, while agreeing that the copied works reflected the copyright owner's creativity, ruled against the copyright owner on the fourth factor, provided the first copy of a copyrighted work was purchased legitimately. In that case, no market dilution would occur since the copyright owner is compensated from the first legitimate access.<sup>72</sup>

It must be emphasised that these are district court judgments from a foreign jurisdiction under a statute that does not precisely mirror Indian law, and these rulings are still subject to further

---

copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.

<sup>65</sup> See Pamela Samuelson, 'Fair Use Defenses in Disruptive Technology Cases', (November 8, 2023). 71 UCLA L. Rev. 1484 (2024), available at <https://ssrn.com/abstract=4631726>, last accessed on November 19, 2025.

<sup>66</sup> Matthew Sag, 'Copyright Safety for Generative AI', December 3, 2023, Houston Law Review, Vol. 61, No. 2, 2023, available at SSRN: <https://ssrn.com/abstract=4438593> or <http://dx.doi.org/10.2139/ssrn.4438593>, pg. 301, last accessed on November 19, 2025.

<sup>67</sup> *Kadrey vs Meta Platforms, Inc.* 2025 U.S. Dist. LEXIS 121064

<sup>68</sup> *Bartz vs PBC*, 2025 U.S. Dist. LEXIS 118989

<sup>69</sup> *Kadrey vs Meta Platforms, Inc.* 2025 U.S. Dist. LEXIS 121064

<sup>70</sup> *Ibid.* at page 42 has held that training of AI systems do not fulfil the fourth factor of fair use, ("No matter how transformative LLM training may be, it's hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books."). (It is important to note that this was only a summary judgment, which the court granted the Defendant on the fair use defence, as the Plaintiff did not plead any evidence of market dilution)

<sup>71</sup> *Bartz vs PBC*, 2025 U.S. Dist. LEXIS 118989

<sup>72</sup> *Ibid.* at page 31, it has been held that training of AI Systems constitutes fair use ("The copies used to train specific LLMs were justified as a fair use. Every factor but the nature of the copyrighted work favors this result. The technology at issue was among the most transformative many of us will see in our lifetimes.").

appeal. However, the key takeaway is that even within the broadest framework of "fair use" in US law, courts may not always arrive at consistent conclusions, and there remains a lack of clarity for stakeholders.

As discussed above, there are strong reasons to believe that training AI Systems may raise concerns about copyright infringement. Simultaneously, there is an ongoing debate regarding whether the "fair dealing" exception can be applied to the process of training of AI Systems.

Many litigations are still pending on whether using copyrighted content to train AI Systems violates the law. Relying on case law to resolve these complex issues may take time, whereas the need for clarity on the underlying issues is much more urgent. Even the Delhi High Court is seized of the questions about infringement and fair dealing for interim injunction purposes. It is a well-established principle of Indian law that judgments on interim injunctions, even detailed ones, carry no precedential value. This decision may also be subject to further appeals. Even after a full trial, appeals could delay a final adjudication, and the other High Courts in the country are not bound to take the same approach.

This working paper recognises that the creative industry and technology sector may face long-term uncertainty about what is legally allowed. Direct legislative intervention may offer all stakeholders legal certainty and strike the right balance between competing interests.<sup>73</sup> The recommendations in this working paper represent an initial step in this direction. This working paper, therefore, does not attempt to resolve these questions or offer definitive conclusions on whether infringement is made out and/or the "fair dealing" exception applies. Instead, the goal is to propose a forward-looking legal and policy framework. This framework would support innovation while safeguarding creative rights.

### 3. POSITION IN OTHER JURISDICTIONS

Other jurisdictions offer useful reference points as India shapes its legal and policy framework to address copyright issues in the context of generative AI. Countries such as the European Union, Singapore, and Japan have introduced copyright exceptions for text and data mining (TDM), each with varying scope and conditions. Australia is also considering updates to its copyright law, however, according to a recent media release, the Australian government has

<sup>73</sup> In UK, Communications and Digital Committee appointed by the House of Lords says in its report that "Govt has a duty to act" and "Govt. cannot sit on its hands for the next decade until sufficient case law has emerged." See <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/54.pdf>, para 246, pg. 69

decided against introducing a TDM exception, and plans to focus on finding workable solutions to foster innovation while ensuring compensation to copyright holders.<sup>74</sup> In the United States, the broad fair use doctrine may extend to certain TDM activities, subject to judicial interpretation. These international examples offer valuable insights for guiding the vision and policy direction.

### 3.1 United States

The Copyright Law of the United States<sup>75</sup> grants exclusive rights to the owner of copyright in relation to reproduction, preparation of derivative works, distribution of copies or phonorecords of copyrighted works, public performance, and public display.<sup>76</sup>

The US copyright law does not provide an express text and data mining exception. However, it provides a broad “fair use” exception that permits use of copyright-protected material, in certain cases. To assess whether a specific use of a work qualifies as fair use or not, the following four factors<sup>77</sup> are taken into account:

- The purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
- The nature of the copyrighted work;
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- The effect of the use upon the potential market for or value of the copyrighted work.

Developers of AI Systems rely on the principle of “transformative use”, which is considered to assess if the work adds something new with a purpose or character that does not substitute the original work, for the use to be considered fair. They argue that the nature of the use of vast datasets to train AI Systems is highly transformative and non-expressive, which does not substitute or compete with the original works, constituting fair use<sup>78</sup>. This proposition is contested, though. Professor Matthew Sag points out that GenAI systems developed with

---

<sup>74</sup> See the report available at <https://ministers.ag.gov.au/media-centre/albanese-government-ensure-australia-prepared-future-copyright-challenges-emerging-ai-26-10-2025>, last accessed on November 19, 2025.

<sup>75</sup> US Code Title 17- Copyrights (“17 USC”), available at <https://www.copyright.gov/title17/title17.pdf>, last accessed on August 21, 2025.

<sup>76</sup> Section 106, 17 USC

<sup>77</sup> Section 107 17 USC

<sup>78</sup> See OpenAI LP’s comment before USPTO, available at [https://www.uspto.gov/sites/default/files/documents/OpenAI\\_RFC-84-FR-58141.pdf](https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf), last accessed at August 9, 2025.

“care”, “will qualify as non-expressive use” and would thus be “*strong candidates for fair use protection*”, however, this is not “*the be-all and end-all of fair use analysis*” and courts may consider whether “*the challenged use undermines the economic incentives that copyright is designed to create*” to assess fair use.<sup>79</sup> It is also argued that fair use doctrine is “*unpredictable and has been stretched beyond its limits in new technological contexts*” and there is a need for a data mining safe harbour to be enacted.<sup>80</sup>

The US Copyright Office launched consultations in March 2023 to “examine the copyright law and policy issues raised by artificial intelligence, including the scope of copyright in works generated using AI tools and the use of copyrighted materials in AI training”<sup>81</sup>. Upon hosting public listening sessions, a notice of inquiry was issued in the Federal Register in 2023 which received over 10,000 comments.<sup>82</sup> After considering the stakeholder comments, the Office released its report in three parts. First two parts were related to digital replicas<sup>83</sup> and copyrightability of outputs generated by AI Systems<sup>84</sup>. Part 3 of the report pertains to generative AI training and the pre-publication version thereof was released in May 2025. According to this report<sup>85</sup>, some uses of copyrighted works for AI training may not qualify as fair use. The report analyses the four-factor test in detail and concludes as follows:

*“The Office expects that some uses of copyrighted works for generative AI training will qualify as fair use, and some will not. On one end of the spectrum, uses for purposes of non-commercial research or analysis that do not enable portions of the works to be reproduced in the outputs are likely to be fair. On the other end, the copying of expressive works from pirate sources in order to generate unrestricted content that competes in the marketplace, when licensing is reasonably available, is unlikely to qualify as fair use.”*

<sup>79</sup> Matthew Sag, ‘Fairness and Fair Use in Generative AI’, Fordham Law Review, Volume 92, Issue 5, available at <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=6078&context=flr>, pg. 1917, last accessed on August 9, 2025.

<sup>80</sup> Jenny Quang, ‘Does Training AI violated Copyright Law’, Berkley Technology Law Journal, Vol 36, available at <https://btj.org/wp-content/uploads/2023/02/0003-36-4Quang.pdf>, pg. 1407-08, last accessed on August 21, 2025.

<sup>81</sup> Copyright Office Launches New Artificial Intelligence Initiative, available at <https://copyright.gov/newsnet/2023/1004.html>, last accessed on August 9, 2025.

<sup>82</sup> See <https://www.copyright.gov/ai/>, last accessed on August 21, 2025.

<sup>83</sup> Copyright and Artificial Intelligence Part 1: Digital Replicas, available at <https://copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf>, last accessed on August 9, 2025.

<sup>84</sup> Copyright and Artificial Intelligence Part 2: Copyrightability, available at <https://copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf>, last accessed on August 9, 2025.

<sup>85</sup> Copyright and Artificial Intelligence Part 3: Generative AI Training [pre-publication version], available at <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, last accessed on August 9, 2025.



In the case of *Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.*<sup>86</sup>, the court held that ROSS had infringed Thomson Reuters' copyrights by using Westlaw headnotes in its AI training data. As per the findings of the court, LegalEase had access to and copied portions of the headnotes. Out of a huge number of headnotes, the court granted summary judgment for Thomson Reuters on 2,243 headnotes, finding actual copying. Issues regarding the remaining headnotes were left for trial. The court rejected the fair use defence, stating that ROSS' use was not transformative, and ROSS' product was deemed a market substitute for Westlaw threatening both the original and derivative markets.

The US has seen two more landmark court rulings this year on whether usage of copyright material for AI training constitutes fair use. In its summary decisions in the cases of *Bartz v. Anthropic*<sup>87</sup> and *Kadrey v. Meta Platforms*<sup>88</sup>, the Court of the Northern District of California found that the use of copyrighted works to train AI models was transformative and fair use, in the facts of these respective cases.

In *Bartz v. Anthropic*, the allegations pertained to the use of copyrighted books by Anthropic to train its AI models, some acquired through purchases and others downloaded from pirate sites, without permission from the rights holders. The U.S. District Court, Northern District of California held that using lawfully acquired books to train Anthropic's large language models (LLMs) constituted fair use since the use was transformative, and training of LLMs serves a fundamentally different purpose from the original books and the training did not compete directly with or replace the original works in the marketplace. The court, however, distinguished between lawfully obtained and pirated materials, and held: *"This order grants summary judgment for Anthropic that the training use was a fair use.... But, it denies summary judgment for Anthropic that the pirated library copies must be treated as training copies. We will have a trial on the pirated copies used to create Anthropic's central library and the resulting damages, actual or statutory. That Anthropic later bought a copy of a book it earlier stole off the internet will not absolve it of liability for the theft but it may affect the extent of*

---

<sup>86</sup> Thomson Reuters Enterprise Centre GmbH et al v. ROSS Intelligence Inc., No. 1:20-cv-613-SB. available at <https://cases.justia.com/federal/district-courts/delaware/dedce/1:2020cv00613/72109/547/0.pdf?ts=1695742638>, last accessed on September 7, 2025.

<sup>87</sup> Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson v. Anthropic PBC, No. C-2405417 WHA. available at <https://docs.justia.com/cases/federal/district-courts/california/candce/3:2024cv05417/434709/231>, last accessed on August 21, 2025.

<sup>88</sup> Richard Kadrey v. Meta Platforms, Inc., Case No. 23-cv-03417-VC, available at <https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv03417/415175/598/>, last accessed on August 21, 2025.

*statutory damages.*” In July 2025, the case was certified as a class-action lawsuit.<sup>89</sup> Subsequently, it was reported that Anthropic has reached a settlement in this case, which includes payment terms of US\$1.5 billion.<sup>90</sup>

Similarly, in *Kadrey v. Meta Platforms Inc.*<sup>91</sup>, the court held that training a generative AI model on copyrighted books, even ones sourced from pirated datasets, qualified as “fair use” under the facts presented since Meta’s use was “highly transformative,” and the AI system analyzed text to learn language patterns rather than replicate the expressive content of the books. The court expressly noted that the decision went in favour of Meta because the plaintiffs failed to present credible evidence of significant harm to the market for their books, for the pivotal market harm factor.

The following observations of the court merit attention:

*“Fair use is a fact-specific doctrine that requires case-by-case analysis that is sensitive to new technologies and their potential consequences. No previous case has involved a use that is both as transformative and as capable of diluting the market for the original works as LLM training is. So, no previous case answers the question whether Meta’s copying was fair use. That question must be answered by flexibly applying the fair use factors and considering Meta’s copying in light of the purpose of copyright and fair use: protecting the incentive to create by preventing copiers from creating works that substitute for the originals in the marketplace.”*

*“In cases involving uses like Meta’s, it seems like the plaintiffs will often win, at least where those cases have better-developed records on the market effects of the defendant’s use. No matter how transformative LLM training may be, it’s hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books. And some cases might present even stronger arguments against fair use.”*

---

<sup>89</sup> Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson v. Anthropic PBC, No. C-2405417 WHA, Order on Class Certification, available at [https://www.govinfo.gov/content/pkg/USCOURTS-cand-3\\_24-cv-05417/pdf/USCOURTS-cand-3\\_24-cv-05417-1.pdf](https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_24-cv-05417/pdf/USCOURTS-cand-3_24-cv-05417-1.pdf), last accessed on November 19, 2025.

<sup>90</sup> See <https://www.nytimes.com/2025/09/05/technology/anthropic-settlement-copyright-ai.html>, last accessed on November 19, 2025.

<sup>91</sup> Richard Kadrey v. Meta Platforms, Inc., Case No. 23-cv-03417-VC, available at <https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv03417/415175/598/>, last accessed on August 21, 2025.



*“In this case, because Meta’s use of the works of these thirteen authors is highly transformative, the plaintiffs needed to win decisively on the fourth factor to win on fair use. And to stave off summary judgment, they needed to create a genuine issue of material fact as to that factor. **Because the issue of market dilution is so important in this context, had the plaintiffs presented any evidence that a jury could use to find in their favor on the issue, factor four would have needed to go to a jury. Or perhaps the plaintiffs could even have made a strong enough showing to win on the fair use issue at summary judgment. But the plaintiffs presented no meaningful evidence on market dilution at all. Absent such evidence and in light of Meta’s evidence, the fourth factor can only favor Meta.** Therefore, on this record, Meta is entitled to summary judgment on its fair use defense to the claim that copying these plaintiffs’ books for use as LLM training data was infringement.”*

Notably, the court in the Meta case noted that the case was dismissed due to the plaintiffs' failure to present sufficient factual evidence of market harm. The court observed that in future litigations, plaintiffs could succeed under similar facts and circumstances if they could adduce evidence demonstrating a tangible negative impact on the market for the copyrighted works.

Although these cases represent a significant step in shaping how fair use is applied to AI training in the US, these are district court decisions subject to appeal and not binding on other district courts or the appellate courts. The reasoning provided in these cases may be persuasive, however, judges in future cases with similar facts could arrive at different decisions. Until a higher court provides definitive guidance, the legal landscape remains open for alternative interpretations and approaches to this issue in the US.

The position in the US on the applicability of the fair use defence to AI training in the US is thus still evolving.

### 3.2 Japan

Japan deals with copyright concerns related to AI training under its Copyright Act, 1970,<sup>92</sup> (“*Japanese Copyright Act*”), which provides exclusive rights to the author of a work, *inter alia*<sup>93</sup> the rights of reproduction; stage performance rights and musical performance rights; on-screen presentation; right to transmit to the public; recitation rights; exhibition rights;

<sup>92</sup> Copyright Act, 1970 (“*Japanese Copyright Act*”), available at <https://www.wipo.int/wipolex/en/legislation/details/22612>. Translated version (Updated as on 19th January 2023) available at <https://www.cric.or.jp/english/clj/index.html>

<sup>93</sup> Section 3, Sub-section 3, Article 21-28, Japanese Copyright Act.

distribution rights; right of transfer; right to rent out; translation, adaptation and other rights.

Japan was the first country to introduce a TDM exception in 2009<sup>94</sup> which was amended in 2018 to broaden the scope. The current provision permits TDM, if the following conditions are met:<sup>95</sup>

- Exploitation should not be for the purpose of “enjoying” or causing another person to “enjoy” the thoughts or sentiments expressed in the copyrighted work;
- Exploitation should not unreasonably prejudice the interests of the copyright owner in light of the nature and purpose of the copyrighted work.

The Japanese Copyright Office (“JCO”) has clarified that the term ‘enjoyment’ refers to the act of *“obtaining the benefit of having the viewer’s intellectual and emotional needs satisfied through using the copyrighted work.”*<sup>96</sup> It was also clarified that - *“The financial benefits that copyright holders receive from their works are generally considered rewards for meeting intellectual and emotional needs. Meanwhile, the exploitation of works for non-enjoyment purposes, which may occur without the consent of the copyright holder, is generally regarded as not harming the financial interests of the copyright holder. Therefore, in such cases acquiring permission for use of the copyrighted works from the copyright holder is not deemed to be required pursuant to Article 30-4 of the Act.”*<sup>97</sup>

The JCO has also clarified that to determine whether the exploitation of the copyrighted work ‘unreasonably prejudices the interests of copyright owner’, it must be considered “whether it will compete in the market with the copyrighted work” and “whether it will impede the potential sales channels of the copyrighted work in the future.” This assessment should be done by taking various factors into account, such as “technological advancements” and “changes in the way the copyrighted work is used.”<sup>98</sup>

---

<sup>94</sup> Article 47-7 prior to the enactment of the 2018 amendment

<sup>95</sup> See Article 30-4, Japanese Copyright Act, available at <https://www.japaneselawtranslation.go.jp/en/laws/view/4207>

<sup>96</sup> Overview on “General Understanding on AI and Copyright in Japan”, available at [https://www.bunka.go.jp/english/policy/copyright/pdf/94055801\\_01.pdf](https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf), pg. 5, para 2, last accessed on November 19, 2025.

<sup>97</sup> *Ibid*, Para 3, Page 5.

<sup>98</sup> *Ibid*, Para 1, Page 10.

TDM exception under the Japanese Copyright Act is available for both commercial and non-commercial purposes, so long as the exploitation satisfies the “non-enjoyment” requirement and does not unreasonably prejudice the interests of the copyright owner, as mentioned above.

On the lawful access requirement in Japan, there seems to be lack of enough clarity in the absence of an explicit provision on the same. Having said that, JCO says *“If an AI developer or AI service provider collects training data for their AI from a website that they know contains pirated or infringing content, there is a high possibility that the business will be held responsible for any copyright infringement caused by the generative AI developed using the training data taken from the website”*.<sup>99</sup>

### 3.3. United Kingdom

In the UK, copyright and rights of copyright owners are governed by the Copyright, Designs and Patents Act, 1988 (“**CDPA**”)<sup>100</sup>, which grants exclusive rights including copying the work;<sup>101</sup> issuing copies of the work to the public;<sup>102</sup> renting/lending the work to the public;<sup>103</sup> public performance;<sup>104</sup> communicating the work to the public;<sup>105</sup> making an adaptation of the work etc.<sup>106</sup>

The CDPA provides a TDM exception, which is limited to computational text and data analysis of anything recorded in the copyrighted works for non-commercial research purposes only. Under this provision, the person making the copy must have lawful access to the work and a sufficient acknowledgement must accompany the copy made unless it is impossible for reasons of practicality or otherwise.<sup>107</sup>

In 2021, the UK Government released the National AI Strategy<sup>108</sup>, setting out ambitions for the UK Government to make the UK one of the leading locations for the development and

---

<sup>99</sup> *Ibid*, Page 11.

<sup>100</sup> Copyright Designs and Patents Act, 1988 (“**CDPA**”), available at: <https://www.legislation.gov.uk/ukpga/1988/48/contents>

<sup>101</sup> Section 17, CDPA

<sup>102</sup> Section 18, CDPA

<sup>103</sup> Section 18A, CDPA

<sup>104</sup> Section 19, CDPA

<sup>105</sup> Section 20, CDPA

<sup>106</sup> Section 21, CDPA

<sup>107</sup> Section 29A, CDPA

<sup>108</sup> National AI Strategy, available at [https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National\\_AI\\_Strategy\\_-\\_PDF\\_version.pdf](https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf), last accessed on August 9, 2025.

commercialisation of AI. In October 2021, consultations were launched on AI and copyright<sup>109</sup>. After considering the submissions, the Government published its response in 2022, saying the Government has decided to introduce an exception allowing the use of lawfully accessed content for TDM for any purpose, with no option for rightsholders to opt-out of the exception.<sup>110</sup>

In around April 2023, Govt. dropped<sup>111</sup> the proposal of a broad TDM exception for commercial purposes. Another consultation was launched<sup>112</sup> on Artificial Intelligence and Copyright in December 2024, giving various options, and the below produced option providing for a broad TDM exception with an opt-out option to rightsholders on EU lines as the Government's "preferred option"<sup>113</sup>:

*"Option 3: A data mining exception with a rights reservation mechanism. This would permit TDM for any use by anyone, but rights holders would be able to opt-out individual works, sets of works or all of their works that they do not want them to be mined for commercial purposes. Where such works are made available online they would need to be accompanied by a machine readable method to reserve rights, so that systems data mining significant numbers of works can easily identify works that can be lawfully mined. This option appears to have the potential to meet our objectives of control, access, and transparency."*

<sup>109</sup> Artificial Intelligence and IP: Consultation on copyright and patents legislation, available at <https://www.gov.uk/government/news/artificial-intelligenceand-ip-consultation-on-copyright-and-patents-legislation?>, last accessed on August 9, 2025.

<sup>110</sup> See Consultation outcome - Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation, available at <https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation>, last accessed on August 9, 2025.

The govt. stated as follows: *"The Government has decided to introduce a new copyright and database right exception which allows TDM for any purpose....Rights holders will no longer be able to charge for UK licences for TDM and will not be able to contract or opt-out of the exception. The new provision may also affect those who have built partial business models around data licensing. However, rights holders will still have safeguards to protect their content. The main safeguard will be the requirement for lawful access. That is, rights holders can choose the platform where they make their works available, including charging for access via subscription or single charge. They will also be able to take measures to ensure the integrity and security of their systems."*

<sup>111</sup> See <https://committees.parliament.uk/publications/39303/documents/192860/default/>, last accessed on August 23, 2025.

<sup>112</sup> See [https://www.gov.uk/government/news/uk-consults-on-proposals-to-give-creative-industries-and-ai-developers-clarity-over-copyright-laws?utm\\_medium=email&utm\\_campaign=govuk-notifications-topic&utm\\_source=a96c8ea2-c315-4647-8d0d-53648f61321b&utm\\_content=immediately](https://www.gov.uk/government/news/uk-consults-on-proposals-to-give-creative-industries-and-ai-developers-clarity-over-copyright-laws?utm_medium=email&utm_campaign=govuk-notifications-topic&utm_source=a96c8ea2-c315-4647-8d0d-53648f61321b&utm_content=immediately). Also see [Copyright and Artificial Intelligence - GOV.UK](#), last accessed on August 23, 2025.

<sup>113</sup> Summary Assessment of Options, available at <https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fassets.publishing.service.gov.uk%2Fmedia%2F676060d87d90cafb3f7d5e4%2FSummary-assessment-of-options.docx&wdOrigin=BROWSELINK>, last accessed on August 9, 2025.

The consultations attracted 11500<sup>114</sup> responses. The TDM exception with an opt-out option proposal received intense opposition from the creative industry, with the industry running a “MAKE IT FAIR” campaign<sup>115</sup> in the UK. Industry stakeholders argue that training of Gen AI models using books, music, audio-visual content, images etc. without prior permission and fair payment, “is unfair and threatens their livelihoods”<sup>116</sup>. Debates in Parliament<sup>117</sup> also highlighted that certain members did not agree with the Government’s proposal for an opt-out system and called it an “unworkable” solution.

The response of the Government to these consultations which were launched last year is still awaited. Expert working groups will now work towards finding solutions, as per the press release<sup>118</sup> of the UK Government dated July 16, 2025.

Additionally, The Data (Use and Access) Act, 2025 (“DUAA”)<sup>119</sup> received royal assent in the UK on 19th June 2025. DUAA governs data protection in addition to the UK GDPR. It has introduced certain obligations for the UK Government on the issue of AI development and the use of copyrighted material.<sup>120</sup> The obligations include the requirement that the Secretary of State publish and present to Parliament both an economic impact assessment of various policy options for handling copyright and AI, and a detailed report on the use of copyright works in AI development, within nine months of DUAA’s commencement. The DUAA also requires information gathering, stakeholder consultation, and careful review before any substantive legal reforms regarding AI and copyright are enacted.

<sup>114</sup> See <https://www.gov.uk/government/news/creative-and-ai-sectors-kick-off-next-steps-in-finding-solutions-to-ai-and-copyright?>, last accessed on August 23, 2025.

<sup>115</sup> See <https://newsmediauk.org/make-it-fair/>, <https://newsmediauk.org/blog/2025/02/25/uk-creative-industries-launch-make-it-fair-campaign/>, last accessed on August 23, 2025.

<sup>116</sup> See ISM’s response to UK Govt. on AI and Copyright consultation (available here <https://www.ism.org/news/copyright-ai-consultation-ism-submission/>) where they rely on the govt. figures to say that UK Creative industries have a contribution of £124.6 billion to the UK economy in 2023 and the “UK must not risk its thriving and respected creative industries for uncertain investment promises from large multinational tech companies”. Figures on which the ISM placed reliance are given in a report available on <https://lordslibrary.parliament.uk/creative-industries-growth-jobs-and-productivity/>

<sup>117</sup> See <https://hansard.parliament.uk/commons/2025-04-23/debates/DE191BB4-49C1-44E0-9378-A247893A71CA/IntellectualPropertyArtificialIntelligence>, last accessed on August 23, 2025.

<sup>118</sup> See the press release “Creative and AI sectors kick-off next steps in finding solutions to AI and copyright” available at [https://www.gov.uk/government/news/creative-and-ai-sectors-kick-off-next-steps-in-finding-solutions-to-ai-and-copyright?utm\\_source=chatgpt.com](https://www.gov.uk/government/news/creative-and-ai-sectors-kick-off-next-steps-in-finding-solutions-to-ai-and-copyright?utm_source=chatgpt.com), last accessed on August 23, 2025. It says “Representatives of the creative and AI sectors will now gather in London in the first of a series of regular planned meetings, with the groups made up of key industry figures. They include representatives of: News Media Association, Alliance for IP, Sony Music Entertainment, Publishers Association, The Guardian, Open AI, Amazon, Meta”

<sup>119</sup> Data (Use and Access) Act 2025 (“DUAA”), available at <https://www.legislation.gov.uk/ukpga/2025/18/contents>

<sup>120</sup> Section 135-137, DUAA

### 3.4 European Union

European Law provides for text and data mining exceptions. Article 2(2) of the Directive on Copyright in the Digital Single Market (CDSM) Directive<sup>121</sup> defines Text and Data Mining (TDM) as “*any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations.*” Article 3<sup>122</sup> and Article 4<sup>123</sup> of the CDSM Directive embody exceptions and limitations to copyright and database rights. Article 3 allows reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have *lawful access*. Any copies made for this purpose must be securely stored and can be kept for future scientific use, including verifying research results. While rights-holders can implement measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted, however, such measures must be proportionate and limited to what's necessary. Member States are also encouraged to promote the development of best practices between rights-holders, researchers, and institutions to support the responsible and effective use of TDM. Article 4 extends TDM rights more broadly, allowing reproductions and extractions of lawfully accessible works and other subject matter

<sup>121</sup> See <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng>, last accessed on August 17, 2025.

<sup>122</sup> **Article 3: Text and data mining for the purposes of scientific research**

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.
2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.
3. Rightsholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.
4. Member States shall encourage rightsholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.

<sup>123</sup> **Article 4: Exception or limitation for text and data mining**

1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.
2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.
3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.
4. This Article shall not affect the application of Article 3 of this Directive.



for the purposes of text and data mining, in any case. Such reproductions and extractions can be retained for as long as needed to carry out the TDM activity.

Simply put, the CDSM Directive establishes two separate TDM exceptions: a limited one under Article 3, specifically for research organisations and cultural heritage institutions, and a general exception under Article 4 for all, which is being invoked in relation to training of AI Systems.

In 2024, the EU adopted the Artificial Intelligence Act (“EU AI Act”), which addresses the use of TDM in the context of generative AI by referencing the TDM exceptions set out in the CDSM Directive. Recital 105<sup>124</sup> says that TDM techniques may be used extensively in the context of training of AI Systems for the retrieval and analysis of such content, which may be protected by copyright and related rights.

A recent Study commissioned by EU Parliament<sup>125</sup> released in July 2025, however, questions the applicability of these TDM exceptions to AI training saying AI training processes go beyond what EU defines as TDM under Article 2(2) of CDSM Directive. Relying on various studies of legal scholars, the study explains the difference as below:

*“TDM belongs to the field of Data Science, which is primarily about analysing existing information. It involves using software to process large volumes of text, images, or other data in order to find patterns—for example, tracking how often a certain term appears in scientific articles. The goal is to extract knowledge from what already exists. By contrast, Generative AI falls within the broader field of Artificial Intelligence, and more specifically, Machine Learning. Rather than merely analysing data, generative AI systems are engineered to process large datasets and algorithmically synthesise outputs—such as textual sequences, visual renderings, or audio patterns—based on statistical correlations. While both TDM and GenAI rely on large-scale data, they use it in fundamentally different ways. A simple way to remember the difference is: TDM finds patterns; GenAI synthesises new expressions. This difference has*

<sup>124</sup> General-purpose AI models, in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos, and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. See <https://artificialintelligenceact.eu/recital/105/#:~:text=General%2Dpurpose%20AI%20models%2C%20in,%2C%20distributed%2C%20used%20and%20consumed.>, last accessed on August 22, 2025. Also see [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689), pg. 27.

<sup>125</sup> See [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST\\_STU\(2025\)774095\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf), last accessed on August 17, 2025.



*significant legal implications. Under EU copyright law, TDM may fall within certain exceptions—particularly when used for research purposes. But GenAI, because it can synthesise outputs that resemble or incorporate protected works, raises more complex and unsettled legal questions.”*

Notably, Article 4 of CDSM directive as detailed out above, has an opt-out option, for the copyright owners. In other words, copyright holders may exercise their opt-out right and bar the use of their copyright protected materials for training of AI Systems, if they wish to. EU AI Act also embodies a transparency<sup>126</sup> obligation and obligates AI Developers to draw up and make publicly available a sufficiently detailed summary about the content used for training<sup>127</sup>. The effectiveness of this opt-out regime has faced criticism which is detailed out in section 3 of this paper.

### 3.5 Singapore

Copyright issues with respect to AI training are dealt with under the Singapore Copyright Act 2021 (“Singapore Copyright Act”),<sup>128</sup> which grants exclusive rights to the copyright owner, *inter alia*, the right to make a copy<sup>129</sup>; to publish the work if it is unpublished<sup>130</sup>; to perform the work in public<sup>131</sup>; to communicate the work to the public<sup>132</sup>; to make adaptation of the work<sup>133</sup>; and to enter into a commercial rental arrangement in case of computer program<sup>134</sup>. The Act, however, includes an exception allowing for ‘computational data analysis’ (“CDA”) of a work or a recording of a protected performance.<sup>135</sup> CDA<sup>136</sup> in relation to a work or a recording of a protected performance includes:

<sup>126</sup> Transparency means making AI system information made available to relevant stakeholders in a comprehensive, accessible and understandable manner. See ISO/IEC 22989:2022(E), available at <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en>, last accessed on September 6, 2025.

<sup>127</sup> See Article 53 of EU AI Act, available at <https://artificialintelligenceact.eu/article/53/>, last accessed on August 22, 2025. Also see [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689), pg. 84.

<sup>128</sup> Copyright Act 2021 (“*Singapore Copyright Act*”), available at <https://sso.agc.gov.sg/Act/CA2021>

<sup>129</sup> *Ibid*, Section 112(1)(a).

<sup>130</sup> *Ibid*, Section 112(1)(b).

<sup>131</sup> *Ibid*, Section 112(1)(c).

<sup>132</sup> *Ibid*, Section 112(1)(d).

<sup>133</sup> *Ibid*, Section 112(1)(e).

<sup>134</sup> *Ibid*, Section 112(1)(g).

<sup>135</sup> *Ibid* Section 243.

<sup>136</sup> Illustration provided under Section 243 is as below:

An example of computational data analysis under paragraph (b) is the use of images to train a computer program to recognise images.

- Using a computer program to identify, extract and analyse information or data from the work or recording; and
- Using the work or recording as an example of a type of information or data to improve the functioning of a computer program in relation to that type of information or data.

The Singapore Copyright Act allows making of a copy of a work or recording of a protected performance (i.e. reproduction of works made for the purpose of CDA) if the following conditions are complied with<sup>137</sup>:

- The copy is made for CDA, or preparing the work or recording for CDA;
- The copy is not used for any other purpose;
- The copy is not supplied to any other person other than for the purpose of verifying the results of CDA or for collaborative research or study relating to the purpose of the CDA;
- There must be lawful access to the material (first copy) from which the copy is made; and
- The first copy should not be an infringing copy, or if it is an infringing copy, the person making a copy for CDA should not know this, or the first copy is obtained from a flagrantly infringing online location which the person making a copy for CDA did not know or could not have reasonably known.

While the term ‘lawful access’ has not been defined under the Singapore Copyright Act, the Act indicates that the same does not allow access to the first copy *via* circumvention of technological protection measures.<sup>138</sup>

The Singapore Copyright Act allows CDA exception for both commercial and non-commercial purposes, so long as the usage of copyrighted works fulfils the prescribed conditions mentioned above. Notably, the law in Singapore also does not include opt-out mechanisms or allow any contractual override<sup>139</sup> to exclude the exception for CDA reservation.

Simply put, while the lawful access to works remains a pre-requisite, Singaporean regime does not provide for an opt-out option for the copyright holders. Given the above, it can be said that

<sup>137</sup> Section 244, Singapore Copyright Act

<sup>138</sup> See the Illustration (a), Section 244(d), Singapore Copyright Act

<sup>139</sup> Section 187(1), Singapore Copyright Act

Singapore provides for wide exceptions, covering both the non-commercial as well as commercial purposes, for training of AI Systems.

#### 4. ASSESSMENT OF VARIOUS REGULATORY MODELS

There is an ongoing debate around the usage of copyrighted materials for training of AI Systems. Globally, content industries are demanding that the use of content for training of AI Systems should be subject to “*consent and compensation*”<sup>140</sup> and allowing AI training on copyrighted works “*without permission or fair remuneration poses an existential threat to the creative industries*”<sup>141</sup>. More than 50000 copyright holders signed a statement<sup>142</sup> saying “*the unlicensed use of creative works for training generative AI is a major, unjust threat to the livelihoods of the people behind those works, and must not be permitted.*”

Arguments in favour of voluntary licensing<sup>143</sup> are made saying it is a “win-win for everybody in the value chain”<sup>144</sup>. A growing number of licensing agreements between AI Developers and the content industry are being cited as evidence that voluntary licensing is not only feasible but could also become the standard practice. The agreements between Open AI and Associated

<sup>140</sup> See Submission of the Media, Entertainment and Arts Alliance (MEAA) to the ‘Inquiry into the opportunities and impacts for Australia arising out of the uptake of Ai technologies’ to Select Committee on Adopting Artificial Intelligence (AI), Submission 137, in Australia, May 2024.

<sup>141</sup> See the response of Independent Society of Musicians (ISM) to consultation of UK Govt. on AI and Copyright, available here <https://www.ism.org/news/copyright-ai-consultation-ism-submission/>, last accessed on August 15, 2025.

<sup>142</sup> See <https://www.aitrainingstatement.org/>, last accessed on August 23, 2025.

<sup>143</sup> **ANI**, in its written submissions to DPIIT Committee says that “*Global AI developers are already entering into commercial licensing agreements with foreign news organisations such as The Financial Times and Associated Press. Indian content creators must not be excluded from this evolving licensing landscape.*”

**Indian Music Industry** in its written submissions to DPIIT Committee recommends to “*maintain India’s existing robust copyright regime, which enables free market negotiations for the use of copyrighted works and allows vibrant licensing markets to develop without downgrading this through the introduction of TDM exceptions or other limitations to copyright*”

**Indian Broadcasting and Digital Foundation**, vide its written submissions to DPIIT Committee says that “*the law should mandate that the use of copyrighted material for training and/or utilization of AI models must be premised on a voluntary, opt-in licensing framework, obtained through explicit consent and involving mutually acceptable consideration and proper attribution*”.

<sup>144</sup> See ‘CCC develops collective licensing solution for internal AI systems’, available at <https://ifro.org/page/article-detail/ccc-develops-collective-licensing-solution-for-internal-ai-systems/?k=e20240717810367678>, last accessed on August 23, 2025.

Press<sup>145</sup>, The Atlantic<sup>146</sup>, Shutterstock<sup>147</sup>, News Corp<sup>148</sup>, Conde Nast<sup>149</sup>, Le Monde<sup>150</sup>; Google and Associated Press<sup>151</sup>; Microsoft and Axel Springer<sup>152</sup> are some of the examples of content partnerships, with different scope and features.

The tech industry claims that training AI Systems on copyrighted materials is fair use<sup>153</sup>, and should be exempted from copyright infringement. A scholarly study by Prof. Edward Lee argues in support of this claim, saying “*the use of copyrighted works to train AI models serves a highly transformative purpose in creating innovative new technology*”<sup>154</sup>. He further argues that fair use is “*a doctrine in which the ends justify the means and it is somewhat analogous to the public necessity doctrine in tort law in which the ends justify the means—i.e., destroying someone else’s property or engaging in activity that would otherwise constitute a trespass to property may be justified if it serves a countervailing, larger public interest*”<sup>155</sup>.

In India too, the tech industry is advocating an express exemption from copyright infringement for the training of AI Systems. Notably, Nasscom, in its written submissions to Government of India has recommended a text and data mining exception (with an opt-out right to the copyright owner) for both commercial and non-commercial purposes stating that the “*basic purpose of copyright law has been to protect original expression from unfair copying and sharing with the public, but training use is incidental, non-expressive and computational*”<sup>156</sup>. Business

<sup>145</sup> See <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/>, last accessed on August 15, 2025.

<sup>146</sup> See <https://www.theatlantic.com/press-releases/archive/2024/05/atlantic-product-content-partnership-openai/678529/>, last accessed on August 15, 2025.

<sup>147</sup> See <https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year>, last accessed on August 15, 2025.

<sup>148</sup> See <https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba>, last accessed on August 15, 2025.

<sup>149</sup> See <https://openai.com/index/conde-nast/> and <https://www.wired.com/story/conde-nast-openai-deal/>, last accessed on August 15, 2025.

<sup>150</sup> See <https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/>, last accessed on August 15, 2025.

<sup>151</sup> See <https://apnews.com/article/google-gemini-ai-associated-press-ap-0b57bcf8c80dd406daa9ba916adacfaf>, last accessed on August 15, 2025.

<sup>152</sup> See <https://news.microsoft.com/source/2024/04/29/axel-springer-and-microsoft-expand-partnership-across-advertising-ai-content-and-azure-services/>, last accessed on August 15, 2025.

<sup>153</sup> See Open AI’s statement to the US Copyright office, available at <https://openai.com/index/openai-and-journalism>, last accessed on August 15, 2025, where they claim that “training AI models using publicly available internet materials is fair use”.

<sup>154</sup> Edward Lee, “Fair use and the origin of AI Training”, available at [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=5253011](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5253011), last accessed on August 23, 2025.

<sup>155</sup> *Ibid*

<sup>156</sup> Nasscom says - India should permit Text and Data Mining (TDM) for both commercial and non-commercial purposes where access is lawful, and a good faith knowledge safeguard is met, solely for the training and input processing stage of machine learning.

Software Alliance<sup>157</sup> (“BSA”) has also urged the Government of India to introduce an explicit exception and argued that “*relying solely on direct or statutory licensing for AI training data may be impractical and may not yield the best outcomes*”<sup>158</sup>. BSA further says that “*using only smaller sets of licensed or public domain material can limit the effectiveness of AI models and, ironically, increase the risk that outputs simply reflect trends and biases of the limited training data sets, and clear TDM exception can help address this challenge by supporting the development of high-quality models while supporting rights holders’ interests*”<sup>159</sup>.

The economic argument, often made in favour of allowing the free use of copyrighted content for AI training, is that requiring payment from AI Developers would merely transfer windfall profits from them to content creators. It is argued that “*copyright is an economic policy instrument to promote innovation. Any extension of the scope of copyright protection beyond what is needed for the production of innovation only strengthens the bargaining position of copyright holders and generates windfall profits, without any increase in consumer surplus or welfare for society as a whole*”<sup>160</sup>. It is argued that this would be a “*pure monopolistic windfall profit because it is not associated with the original purposes and main business models*”<sup>161</sup> of production of content.

On the above argument, a study<sup>162</sup> commissioned by the International Federation of the Phonographic Industry states that this argument overlooks two key issues. First, while AI

- TDM exception should be without prejudice to applicable laws that protect specific categories of data, including personal data and confidential data.
- The good faith knowledge safeguard should protect the TDM user on the lawful access test, if the user “does not know” that a source is infringing. However, in cases where the source of data used for TDM presents a heightened risk of unlawful access, the protection should only be available if the user “has no reasonable grounds to suspect that the source is infringing

For content that is publicly accessible online (freely accessible without paywalls, logins, or other access restrictions), rightsholders should be able to reserve their works from TDM through a machine readable opt out, at the point of availability.

<sup>157</sup> BSA’s members include Adobe, Alteryx, Amazon Web Services, Asana, Atlassian, Autodesk, Bentley Systems, Box, Cisco, Cloudflare, Cohere, Dassault Systemes, Databricks, Docusign, Dropbox, Elastic, EY, Graphisoft, HubSpot, IBM, Informatica, Kyndryl, MathWorks, Microsoft, Notion, Okta, OpenAI, Oracle, PagerDuty, Palo Alto Networks, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Trend Micro, TriNet, Workday, Zendesk, and Zoom Communications Inc.

<sup>158</sup> See <https://www.bsa.org/files/policy-filings/07162025dpiitcopyrightai.pdf>, last accessed on August 23, 2025.

<sup>159</sup> *Ibid*

<sup>160</sup> See B Martens, ‘Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs’, Working Paper 09/2024, Bruegel, available at [https://www.bruegel.org/system/files/2024-04/WP%2009%20040424%20Copyright%20final\\_0.pdf](https://www.bruegel.org/system/files/2024-04/WP%2009%20040424%20Copyright%20final_0.pdf), last accessed on August 23, 2025.

<sup>161</sup> *Ibid*

<sup>162</sup> Jorge Padilla and Kadambari Prasad, ‘Generative AI Models at the Gate - Licensing frameworks for the effective and efficient protection of copyright protected content in an AI world’, available at <https://compass-lexecon.files.svdcn.com/production/editorial/2025/04/Generative-AI-Models-at-the-Gate-Report-for-IFPI-Compass-Lexecon.pdf?dm=1743758320>, last accessed on August 23, 2025.

training itself might not impact current revenues, AI-generated music can directly compete with and “cannibalize” creators’ core income sources. This *substitution effect* threatens creators’ earnings and, consequently, their future investment capacity. Second, training AI Systems need to be continuously updated with fresh content, and this creates a dynamic incentive challenge. Without fair compensation, creators would lack motivation to produce high-quality works, which are crucial for improving AI Systems over time. Thus, free use risks undermining content creation incentives, the quality of AI tools and reducing overall social welfare. Other organisations also echo this argument.<sup>163</sup>

According to data from the Harvard Business Review<sup>164</sup>, the launch of ChatGPT led to a 30% decline in writing jobs and a 20% reduction in coding roles. Also, AI image generators resulted in a 17% drop in image creation jobs. This is cited as evidence to say that it is ‘*self-evident*’ that generative AI will compete with its training data.<sup>165</sup>

In light of the above arguments from both sides, the dual challenge for policymakers is to design a policy that offers adequate protection to content creators, while preserving their incentives to invest time and money in the creation of high-quality content, without erecting a barrier for the development of AI Systems.

Amending the fair dealing provision under Section 52(1)(a) of the Copyright Act, 1957, will neither help strike a balance, nor will it effectively address the legal exposure of AI Developers under copyright law for the following reasons:

(a) The fair dealing provision is primarily an “exception” to copyright infringement, in the nature of a defence and not an enabling provision such as the exclusive rights provisions of copyright owners under Section 14.<sup>166</sup>

---

<sup>163</sup> News Broadcasters and Digital Association, vide its written submissions to DPIIT Committee says “AI systems that scrape and repurpose such news content often generate real-time summaries or responses, which directly compete with news broadcasters and publishers’ revenue streams, causing irreparable financial harm to their market, which has a short monetization window due to its time-sensitive nature.”

<sup>164</sup> See Ozge Demirci, Jonas Hannane and Xinrong Zhu, ‘AI and machine learning Research: How Gen AI Is Already Impacting the Labor Market’, available at <https://hbr.org/2024/11/research-how-gen-ai-is-already-impacting-the-labor-market>, last accessed on August 23, 2025.

<sup>165</sup> See Ed Newton-Rex, “The UK’s AI & copyright proposals would irreparably harm the country’s creators” available at <https://fairtraining.substack.com/p/the-uks-ai-and-copyright-proposals>, last accessed on August 23, 2025.

<sup>166</sup> “From jurisprudential point of view, intellectual property rights work as a two-way sword. On the one hand, there is a growing awareness that such protection is a sine qua-non of the motivational factor underlying the creation of an intellectual work; however, on the other hand, granting an absolute protection to the intellectual work can be detrimental to the further progress of humanity. In order to balance the rights of the author on the one hand and the society on the other hand certain limitations have been made a part and parcel of the IPR

(b). Protection under the fair dealing provision is contingent on meeting specific criteria, namely,

- i) Any person or entity seeking refuge or the defence available under the fair dealing provision bears the burden of showing how its acts fall within the provision's narrow purposes enumerated under the provision.
- ii) Assuming the above requirement is satisfied, the AI Developer will also have to show that apart from serving the specific purposes, the acts of the AI Developer amount to "fair dealing". Issues such as market substitution (of the market for the original copyright work by an AI System), extent of commercial use, extent of copyrighted work used, etc., would be weighed by a court. This would be a factual analysis, which would differ in each situation/case.

As a consequence, any legislative amendment to the fair dealing provision to facilitate AI Training will not reduce the legal uncertainty faced by AI Developers.

In light of all the above, Committee reflected in detail to work towards recommending a regulatory architecture that would be best suited to offer legal certainty to AI Developers, uphold the basic principles of copyright and strike a balance. The suitability of all existing models was evaluated by the Committee. The section below provides such an assessment.

#### **4.1 Voluntary Licensing via Direct licensing Agreements**

Under this proposed framework, developers of AI Systems employing copyrighted works for model training must engage in the process of seeking, negotiating, and finalising a voluntary licensing agreement with each respective copyright owner, leading to very *high transaction costs*. This approach prioritises the rights of copyright holders, enabling content creators to assert their rights regarding the reproduction of copyrighted materials used in the training of AI Systems. Additionally, it affords copyright owners the flexibility to negotiate appropriate licensing fees for such usage.

---

*statutes around the globe. In the realm of copyright laws, one of the more important limitations is the doctrine of "fair use." This doctrine of Fair dealing/Fair use is a limitation and exception to the exclusive right granted by copyright law to the author of a creative work, and it allows limited use of copyrighted material without acquiring permission from the rights holders.*" *Syndicate of The Press of The University of Cambridge on Behalf of The Chancellor, Masters and School and Ors. vs. B.D. Bhandari and Ors. [2011:DHC:3895-DB]*



Several stakeholders representing the copyright owners have suggested a framework of direct voluntary licensing. The submissions of the Indian Music Industry (IMI) <sup>167</sup> to DPIIT Committee emphasised the right of copyright owners to authorise or deny the use of their protected content and that AI Developers should be required to remunerate copyright holders when they train on their protected content. The Indian Broadcasting & Digital Foundation (IBDF) <sup>168</sup> has suggested the development of innovative licensing platforms wherein AI Developers could easily license a large body of works directly from copyright owners.

However, this model of voluntary licensing, which works on the principle of consent and gives full control to copyright owners over their works, may prove impractical. Data indicates that, by 2025, the internet is projected to encompass 175 zettabytes of information, <sup>169</sup> likely including content from potentially a billion different copyright owners. The inherent transaction costs and logistical complexities involved in this exercise render it exceedingly arduous and challenging. Furthermore, this approach appears overly optimistic, as much of the internet's data has already been scraped and aggregated by various third-party entities, such as *Common Crawl*<sup>170</sup>, *LAION-5B*<sup>171</sup>, *The Pile*<sup>172</sup>, *3I Data Scraping*<sup>173</sup>, *Zyte*<sup>174</sup>, *Scrapy*<sup>175</sup>, and *Apify*.<sup>176</sup> Therefore, advocating for voluntary licenses under these prevailing conditions may not solve the problem because AI Developers could simply rely on these datasets.

Moreover, implementing such a licensing system would create substantial entry barriers for new entrants aiming to develop their own AI Systems, as established AI Systems, which have amassed the requisite data, would not be subject to the same licensing requirements.

The US Copyright Office has opined that voluntary licensing may have worked in some industries, but it is not scalable in the context of training AI Systems.<sup>177</sup> Similarly, a study

<sup>167</sup> Written Submissions of IMI to DPIIT Committee dated July 7, 2025

<sup>168</sup> Written Submissions of IBDF to DPIIT Committee, proposing direct licensing framework

<sup>169</sup> 1 Zettabyte = 1 billion Terabytes and 1 Terabyte = 1 billion Bytes. Seagate Technology and IDC, Data Age 2025: The Digitization of the World – From Edge to Core (IDC White Paper Doc No US44413318, November 2018), available at <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>, last accessed on July 20, 2025

<sup>170</sup> See Common Crawl, <https://commoncrawl.org/>

<sup>171</sup> See LAION-5B, <https://laion.ai/blog/laion-5b/>

<sup>172</sup> See The Pile, [The Pile](#)

<sup>173</sup> See 3I Data Scraping, <https://www.3idatascraping.com/>

<sup>174</sup> See The Zyte, <https://www.zyte.com/>

<sup>175</sup> See Scrapy, <https://www.scrapy.org/>

<sup>176</sup> See Apify, [https://apify.com/store?utm\\_term=apify&utm\\_campaign=TOP-](https://apify.com/store?utm_term=apify&utm_campaign=TOP-)

<sup>177</sup> Copyright and Artificial Intelligence Part 3: Generative AI training, available on <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 103, last accessed on July 29, 2025.

commissioned by the European Parliament has cautioned that relying on voluntary licensing risks leaving less commercially visible works unlicensed and underserved.<sup>178</sup>

Scholars have echoed similar issues, highlighting the scalability issue of entering into separate licensing agreements, given that the data on which AI Systems train may be owned by thousands, if not millions, of individual owners.<sup>179</sup> Academic commentators have criticised voluntary licensing mechanisms for their potential to create entry barriers and the possibility of industry consolidation. It has been argued that incumbents would be advantageously placed vis-à-vis newer entrants, as the former may use their existing AI Systems to generate synthetic data and train future models on a combination of already trained models and public domain data.<sup>180</sup>

It has also been pointed out that since the training of AI Systems entails weight assignment of the sources that a model is exposed to while being trained, often the weight assignment is done in a subtle, complex, and non-linear fashion<sup>181</sup>. It has been argued that this requires a licensing mechanism to quantify the remuneration proportionate to the degree to which their work contributes to generating output by trained AI Systems, which is a very challenging task.<sup>182</sup>

Also, in a voluntary model, not all rights holders may license their content, leading to gaps in data coverage, which can impair the quality and bias-resilience of AI Systems. Quality data from diverse sources is needed to develop good AI Systems, apart from the high volume of

---

<sup>178</sup> Generative AI and copyright: training, creation, regulation. In European Parliament (Report PE 774.095), available at

[https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST\\_STU\(2025\)774095\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf), pg. 81, last accessed on November 19, 2025.

Also see Damle, S. (2023). “*Statement of SY Damle before the U.S. House of Representatives Committee on the Judiciary, Subcommittee on Courts, Intellectual Property, and the internet on “artificial intelligence and intellectual property: Part I — interoperability of ai and copyright law.”*” at page 12, available at <https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/damle-testimony.pdf> - Sy Damle, former Assistant Registrar of Copyright at the US Copyright Office, testified before the US House of Representatives Committee that voluntary licensing negotiations are challenging due to the unlimited training data supply. This would mean that no individual rightsholder will be able to demand more than nominal compensation for the use of its works. He warned that it is unfeasible for AI developers to negotiate with every copyright owner of the billions of data points needed for model training. Using only licensed or public domain material would reduce model effectiveness. It may inadvertently increase the risk of regurgitating training data at the output stage, raising a separate copyright infringement angle.

<sup>179</sup> Lemley, Mark A. and Casey, Bryan, Fair Learning, 99 Tex. L. Rev. 743, at 770, available at <https://ssrn.com/abstract=3528447> or <http://dx.doi.org/10.2139/ssrn.3528447>, last accessed on July 29, 2025.

<sup>180</sup> Matthew Sag, Pamela Samuelson, and Christopher Jon Sprigman, ‘Comments In Response To The Copyright Office’s Notice Of Inquiry On Artificial Intelligence And Copyright’, October 04, 2024, Published response to the Copyright Office NOI, Response to NOI 13, at pg. 27-28, available at <https://ssrn.com/abstract=4976391> or <http://dx.doi.org/10.2139/ssrn.4976391>, last accessed on November 19, 2025.

<sup>181</sup> *Ibid*, Response to NOI No. 12 at pg. 26.

<sup>182</sup> *Ibid*

data.<sup>183</sup> It is said that “AI systems are designed to think the way we do. But who exactly is the “we” these AI systems are being modelled on. Whose values, ideals, and worldviews are being taught? The short answer is not yours-and also not mine. Artificial intelligence has the mind of its tribe, prioritizing its creators’ values, ideals, and worldviews”<sup>184</sup>. This highlights the need to train AI Systems on diverse content, ensuring they do not represent any single sect, region, or community and do not carry biases. In sectors like health, security etc., the need of bias resilient AI models becomes even more important. Niti Ayog’s paper on Responsible AI also emphasizes that “ensuring that AI systems are inclusive and non-discriminatory is important, especially in high-risk use cases, and it requires availability of high quality and representative datasets”<sup>185</sup>.

Given these considerations, the model of voluntary licensing is not the optimal solution due to the risks of holdouts by copyright owners, risk of AI bias in the circumstances where enough material is not made available for training, elevated transaction costs, practical challenges, and the consequent establishment of barriers for new market entrants.

#### 4.2 Text and Data Mining Exception

The second model under consideration is creating an exception under Section 52 of the Copyright Act of 1957 that allows the reproduction or other exploitation of copyrighted works for training AI Systems. Such a proposed exception often carries the nomenclature of the Text Data Mining (“TDM”) exception. TDM exceptions predate the contemporary discussions on AI and were formulated as statutory exceptions in several jurisdictions long before the emergence of contemporary AI Systems. Although the exact verbiage of the TDM exception differs across jurisdictions, as evidenced in **Annexure B** of this working paper, it is broadly similar in spirit and intent.

As the name suggests, TDM exceptions contemplate exploiting copyrighted works to mine information/data from the same. It is often referred to as “data analysis”. For instance, in Japan, the statute allows such exploitation for the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from many works or a large volume of works. The TDM exception in Singapore is directed to

<sup>183</sup> See <https://www.bsa.org/files/policy-filings/07152025bsaaitraining.pdf>, last accessed on August 18, 2025.

<sup>184</sup> Amy Webb, ‘The Big Nine – How the Tech Titans & Their Thinking Machines could Warp Humanity’, PublicAffairs New York, Hachette Book Group, 2019, pg. 53.

<sup>185</sup> See <https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf>, last accessed on August 18, 2025.

“computational data analysis”, which is inclusively defined to cover the use of a computer program to identify, extract, and analyse information or data from the work. The UK TDM exception is directed to “computational analysis”. These terminologies appear to be synonymous. “Data analysis” in the ordinary sense is understood to cover the process of systematically collecting, cleaning, transforming, describing, modelling, and interpreting data, generally employing statistical techniques<sup>186</sup>.

The discussions for TDM arose when copyrighted works were turned into machine-readable data for automatic processing for various purposes, but without displaying the work to the public without express permission, which typically occurs in mass digitization of works.<sup>187</sup> Such discussions gained steam in relation to the Google Books project. The purpose behind such mining can be as simple as indexing for bibliographic purposes, or as complex as understanding or developing relationships among works or within a body of literature. TDM is understood to be a “*research technique to collect information from large amounts of digital data through automated software tools*”<sup>188</sup>. It is understood to work by:

1. *“identifying input materials to be analysed, such as works, or data individually collected or organised in a pre-existing database;*
2. *copying substantial quantities of materials—which encompasses*
  - a. *pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted and*

<sup>186</sup> Eldridge, Stephen. “data analysis”. Encyclopedia Britannica, 19 Sep. 2025, available at <https://www.britannica.com/science/data-analysis>, last accessed on November 19, 2025.

<sup>187</sup> See, generally, Maurizio Borghi and Stavroula Karapapa, ‘Non-Display Uses of Copyright Works: Google Books and Beyond’, Queen Mary Journal of Intellectual Property, Vol. 1, No. 1, April 2011, available at <https://ssrn.com/abstract=2358912> or <http://dx.doi.org/10.2139/ssrn.2358912>, last accessed on November 19, 2025.

<sup>188</sup> Christophe Geiger, Giancarlo Frosio, and Oleksandr Bulayenko, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects’, (March 2, 2018). Centre for International Intellectual Property Studies (CEIPI), Research Paper No. 2018-02, pg. 5-6, available at SSRN: <https://ssrn.com/abstract=3160586> or <http://dx.doi.org/10.2139/ssrn.3160586>, last accessed on November 19, 2025.

Also see a ‘Study on the legal framework of text and data mining’ funded by EU Commission, pg. 28, available at <https://op.europa.eu/da/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en>, last accessed on November 19, 2025.

- b. possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed;*
3. *extracting the data; and*
4. *recombining it to identify patterns into the final output.*<sup>189</sup>

Also, many jurisdictions advocating for a TDM exception do not allow copyright holders the right to “opt out” of this framework. This perspective is reflected in the practices of Japan and Singapore.<sup>190</sup> Conversely, the European Union's approach features an “opt-out” right, whereby copyright owners may take reasonable measures to prevent the developers of AI Systems from accessing their copyrighted works for reproduction or training purposes.<sup>191</sup> Copyright holders can explicitly restrict the use of their works by others for such data analysis, including through machine-readable means. Companies or entities engaged in TDM must adhere to such reservations and refrain from reproducing copyrighted works for training objectives.

There is considerable ambiguity regarding the implementation of the “opt-out” mechanism due to its lack of guidance on how copyright owners should reserve their rights. In its 2021 ruling, the European Court of Justice (CJEU) concluded that copyright owners can only opt-out and restrict text and data mining (TDM) of their copyrighted works if they implement “effective technological measures.” It clarified that the term “effective technological measure” would have to be interpreted within the meaning of Articles 6(1)<sup>192</sup>, (3)<sup>193</sup> of the Infosoc Directive

<sup>189</sup> Christophe Geiger, Giancarlo Frosio, and Oleksandr Bulayenko, ‘The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects’, (March 2, 2018). Centre for International Intellectual Property Studies (CEIPI), Research Paper No. 2018-02, pg. 5-6, available at SSRN: <https://ssrn.com/abstract=3160586> or <http://dx.doi.org/10.2139/ssrn.3160586>, last accessed on November 19, 2025.

<sup>190</sup> See Sr. No. (4) in Annexure B, which tabulates the contours of TDM exceptions across various jurisdictions.

<sup>191</sup> See Article 4, 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, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790>, last accessed on November 19, 2025.

<sup>192</sup> **Article 6:** Obligations as to technological measures

(1). Member States shall provide adequate legal protection against the circumvention of any effective technological measures, which the person concerned carries out in the knowledge, or with reasonable grounds to know, that he or she is pursuing that objective

(3). For the purposes of this Directive, **the expression ‘technological measures’ means any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject matter, which are not authorised by the right holder of any copyright or any right related to copyright as provided for by law or the sui generis right provided for in Chapter III of Directive 96/9/EC. Technological measures shall be deemed ‘effective’ where the use of a protected work or other subject matter is controlled by the right holders through application of an access control or protection process, such as encryption, scrambling or other transformation of the work or other subject-matter or a copy control mechanism, which achieves the protection objective.**

<sup>193</sup> **Article 3:** Right of communication to the public of works and right of making available to the public other subject-matter:

2001/29/EC. The CJEU opined that any interpretation to the contrary would make it difficult for individual users to know whether copyright owners intended to reserve the right to perform TDM activities on their copyrighted works<sup>194</sup>.

If the content is publicly available online, copyright owners can appropriately reserve their rights using such machine-readable methods, or even by the terms and conditions of a website or accompanying metadata. When content is unavailable online, copyright owners may choose to reserve their rights through contractual agreements or a unilateral declaration.<sup>195</sup> A popular technological measure is “robots.txt” or the robot exclusion protocol. This protocol is attached to the root of a website, and indicates which pages should not be accessed. It could be likened to a sign saying “no trespass” or “no access” as used in the brick-and-mortar world.<sup>196</sup>

Some scholars have opined that such TDM exceptions cover AI training because the process of training AI Systems is to derive statistical relationships between words in the language.<sup>197</sup> However, there are practical and legal uncertainties on whether TDM exceptions are sufficient to truly enable AI training:

- a) For one, many jurisdictions employing the TDM exception, such as the United Kingdom and Switzerland, impose a condition that the exception is limited to non-commercial use only, while jurisdictions such as Japan, the EU, and Singapore<sup>198</sup> do not impose such a limitation. The United States represents a more dynamic scenario, as commercial nature of use is one of several factors considered in a holistic “fair use” analysis; however, current regulations allow TDM solely for non-commercial

---

(1). Member States shall provide authors with the exclusive right to authorise or prohibit any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them.

<sup>194</sup> Judgement of March 9, 2021, Case C-392/19, *VG Bild-Kunst vs Stiftung Preußischer Kulturbesitz*, ECLI:EU:C:2021:181, para 45 and 46.

<sup>195</sup> Gina Maria Ziaja, ‘The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightsholders or a suffocating blanket for European artificial intelligence innovations?’, *Journal of Intellectual Property Law & Practice*, Volume 19, Issue 5, May 2024, Pg. 456, available at <https://doi.org/10.1093/jiplp/jpae025>, last accessed on November 19, 2025.

<sup>196</sup> Maurice Schellekens, ‘Robot.txt: Balancing Interests of Content Producers and Content Users’ in Ronald Leenes and Eleni Kosta (eds), *Bridging Distances in Technology and Regulation* (Wolf Legal Publishers 2013), Pg No.175, available at <https://repository.tilburguniversity.edu/server/api/core/bitstreams/0fca36cd-8e9e-4a03-9f8e-ebf0b584f1c9/content>, last accessed on November 19, 2025.

<sup>197</sup> Written Testimony of Christopher Callison-Burch, Hearing on ‘Artificial Intelligence and Intellectual Property: Part I – Interoperability of AI and Copyright Law’, May 17, 2023, available at <https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf>, last accessed on November 19, 2025.

<sup>198</sup> See Sr. No. (2) in Annexure B, which tabulates the scope of beneficiaries of TDM exceptions across various jurisdictions.



purposes.<sup>199</sup> An exception that enforces a non-commercial usage limitation fails to provide sufficient incentive for AI Developers, inadvertently creating entry barriers for new industry participants. It is noteworthy that, though the current UK legislation is narrower in scope and permits a TDM exception only for non-commercial purposes, the UK government, in a consultation paper, as mentioned above, suggested a broad TDM exception covering even commercial uses of data, except in those instances wherein the copyright owners have expressly reserved their rights.

- b) Similarly, TDM exceptions were not crafted specifically with AI Systems in mind. Thus, it is unclear if the currently worded TDM exception that permit data analysis would be sufficient to exempt all the training steps of an AI System. As observed by the US Copyright Office, the AI training process may necessitate repetitive and temporary reproductions of works or significant portions thereof to present to the AI System in batches.<sup>200</sup> The justification of this aspect of the training process under the TDM exception remains ambiguous.
- c) A further aspect is that most jurisdictions require that the work must have been “lawfully accessed”. This condition is observed in the jurisdictions like EU and Singapore<sup>201</sup>. It is undeniable that AI training relies on datasets previously created by third parties, which often include pirated works.<sup>202</sup>
- d) Some scholars have also argued that even under the current EU and UK laws, unlicensed training of AI Systems does not fall under the scope of the application of copyright exception, not only those specific to TDM but also those carved out for research and teaching. This is because:
  - i. *Firstly*, while TDM may be part of the AI training process, it is neither synonymous with AI training nor is it all that AI training entails<sup>203</sup>. This is because a distinction needs to be drawn between the input/training and the output phase. While TDM, being an automated analytical technique aimed at analysing text and data in digital

---

<sup>199</sup> See Annexure B

<sup>200</sup> Copyright and Artificial Intelligence Part 3: Generative AI training, available on <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 27, last accessed on 29.07.2025

<sup>201</sup> See Annexure B

<sup>202</sup> Copyright and Artificial Intelligence Part 3: Generative AI training, available on <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 14, last accessed on 29.07.2025

<sup>203</sup> Eleonora Rosati, ‘Copyright Exceptions and Fair Use Defences for AI Training Done for ‘Research’ and ‘Learning’, or the Inescapable Licensing Horizon’, Pg No. 4, June 30, 2025, European Journal of Risk Regulation, available at <https://ssrn.com/abstract=5331405> or <http://dx.doi.org/10.2139/ssrn.5331405>, last accessed on November 19, 2025.

form, may cover the input stage, the output stage refers to the generation of content (text, audio, image or video) in response to the instructions (prompts) given by the users of the resulting AI models<sup>204</sup>. It is pertinent to note that Articles 3 and 4 of the CDSM Directive solely encompass acts of extraction and reproduction. Hence, scholars have argued that subsequent acts restricted by copyright are not within the exceptions and limitations of the EU Directive. Rosati argues that if a Gen AI model lawfully trained on third-party protected content subsequently reproduces and/or communicates/makes available to the public third-party protected content, such acts would not be covered by any Article of the EU directive.<sup>205</sup>

- ii. *Secondly*, TDM in itself is not one singular activity but a composition of several distinct activities. The creation of an AI training model may involve creating tools to scrape data, marketing or selling the tools to others, using the tools to scrape and copy-protected content, and storing scraped content in one or more data repositories or databases. The creation of AI training models involves such a chain of linked activities and interactions, and each link may or may not give rise to liability for copyright infringement. The present framework is silent on which of these activities within the TDM exception is covered under the present exception.<sup>206</sup>

On the other hand, even if one conceives of a carefully worded exception that allows commercial use, and language is devised to specifically cover AI training and address some of the concerns above, it is equally valid that such an exception devoid of an “opt-out” provision undermines the interests of copyright owners, as they would receive no compensation or benefits despite their works being reproduced during the training of AI Systems by AI Developers who may possess commercial motivations.

The US Copyright Office has taken a position that the “opt-out” mechanism is inconsistent with the basic principle of the Copyright Act, as it dispenses with the need for consent from copyright owners for uses which are within the scope of their statutory rights. It also highlights the limitation of AI companies honouring a platform-level flag, like robots.txt, because copyright owners may not always have control over the platforms where their works appear.

---

<sup>204</sup> E. Rosati, “Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law”, *European Journal of Risk Regulation*, Volume 16 , Issue 2 , June 2025, pg. 610, available at <https://doi.org/10.1017/err.2024.72>, last accessed on November 19, 2025.

<sup>205</sup> *Ibid*, pg. 611.

<sup>206</sup> Alissa Centivany, ‘A Window into Generative Artificial Intelligence Under Copyright Law & Policy in Canada’, available at <https://doi.org/10.29173/cais1951> , pg. 6, last accessed on November 19, 2025.

At the same time, a system-by-system or a company-by-company opt-out would be burdensome to monitor and implement<sup>207</sup>.

This issue is also addressed by some stakeholders in their comments to the present DPIIT Committee. The Digital News Publishers Association (DNPA) has argued that providing an opt-out functionality does not excuse past infringement. It has also been pointed out that the opt-out functionality is also riddled with concerns of transparency or the lack of an enforcement mechanism through which the copyright owners can verify whether their opt outs are being respected or not<sup>208</sup>. Similarly, the News Broadcasters & Digital Association (NBDA) has criticised the opt-out mechanism and has suggested an opt-in mechanism instead<sup>209</sup>.

A recent study commissioned by the European Parliament has criticised this system of TDM with opt-out functionality for its limitations:

- a) That inversion of the burden effectively treats silence as consent. It is akin to assuming that the contents of a book are freely reproducible unless the author prints “no copying” on every page, which undermines the intent of giving the owners a structure of exclusive rights. The current EU approach, legitimising a systemic extraction of protected content where there isn’t an explicit “opt-out”, converts the right into a default licensing regime<sup>210</sup>.
- b) Opt-out functionality does not prevent downstream reuse once the content is stripped of its metadata and transformed, and thus, control lost over the data is irrecoverable<sup>211</sup>.
- c) Small and independent creators bear a disproportionate administrative burden of complying, while the actual uptake of the functionality remains minimal. Article 4 represents not a mere exception but a paradigmatic shift in the nature of copyright protection. It subordinates the exclusivity of rights to a presumed utility for innovation, enforced through mechanisms that structurally favour large-scale, well-resourced users over individual creators<sup>212</sup>.

<sup>207</sup> Copyright and Artificial Intelligence Part 3: Generative AI Training [pre-publication version], available at <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. No. 105, accessed on August 9, 2025.

<sup>208</sup> Written Submissions by DNPA dated July 4, 2025.

<sup>209</sup> Written Submissions by NBDA dated July 4, 2025.

<sup>210</sup> ‘Generative AI and copyright: training, creation, regulation’, available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST\\_STU\(2025\)774095\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf), pg. 120, last accessed on August 17, 2025.

<sup>211</sup> *Ibid.*, pg. no. 121

<sup>212</sup> *Ibid.*, pg. no. 122

- d) The Risks of data laundering (whereby datasets are compiled under Article 3 for scientific research and are subsequently reused in commercial training under Article 4), remain present and circumvent the limit prescribed under both Article 3 and 4, allowing effective commercial use to benefit from research-based exemption<sup>213</sup>.
- e) Opt-out functionality conditions the exercise of rights on technological readiness, thereby creating an exclusionary effect against creators with fewer resources<sup>214</sup>.
- f) Lack of specific guidance and lack of harmonized standards as to what an appropriate opt-out is, has led many AI Developers to proceed under the assumption that in the absence of a valid opt-out, their actions are lawful<sup>215</sup>.

Given these considerations, the TDM exception, even with an opt-out facility, does not help ensure an adequate balance among the interests of all relevant stakeholders.

While members of the Committee from Nasscom expressed support for this model, the majority of the members of the Committee were strongly opposed to it. In practice, this model may limit data availability, especially if many rights holders choose to opt out. That could compromise the quality of AI Systems, as AI Developers may lack access to broad, representative datasets. It denies compensation to all those content rightsholders whose data is used in AI training because of their failure to exercise the right to opt out. Additionally, the practicality of implementing opt-outs, especially ensuring clear, standardised, and machine-readable notices across platforms, is still debated and poses serious compliance questions.

### 4.3 Collective Licensing and Extended Collective Licensing

The other options considered by the Committee were collective licensing (**CL**) and extended collective licensing (**ECL**). CL involves a collective organisation whose function is to facilitate the dissemination of works (music, dramatic works, films, etc.) at a lower transaction cost than with individual licensing and to administer the Royalties collected from the users and distributed to the copyright owners who are the members of the collective.<sup>216</sup> ECL is something similar, whereby Collective Management Organisations (CMOs) will perform collective

---

<sup>213</sup> *Ibid*, pg. no. 121

<sup>214</sup> *Ibid*, pg. no. 122

<sup>215</sup> *Ibid*, pg. no. 33-34

<sup>216</sup> European Broadcasting Union, 'Extended Collective Licensing: A valuable catalyst for the creative content economy in Europe', at pg. no. 4, available at [https://www.ebu.ch/files/live/sites/ebu/files/Publications/EBU-Extended-Collective-Licensing\\_EN.pdf](https://www.ebu.ch/files/live/sites/ebu/files/Publications/EBU-Extended-Collective-Licensing_EN.pdf), last accessed on November 19, 2025.

licensing, though it is extended by law to cover all non-member copyright owners of the same category who might not even be members of the concerned organisation<sup>217</sup>.

CL is not new to India. It is statutorily recognised in India under Section 33 of the Copyright Act, 1957<sup>218</sup>. A copyright society can issue or grant licences for any work in which copyright subsists or in respect of any other right given by the Copyright Act. The Copyright Act also addresses other provisions related to copyright societies, including the tariff structure of such societies (Section 33A) and the administration of these societies (Section 34). Presently, four registered copyright societies are exercising the right to collective licensing on behalf of their members<sup>219</sup>.

In the European Union, CL and ECL have been envisaged for satellite broadcasting and cable transmissions since 1993.<sup>220</sup> Similarly, the Directive on the harmonisation of certain aspects of copyright and related rights in the information society expressly acknowledges the validity of these systems.<sup>221</sup> ECL, though in other contexts, has also been recognised for many years in different jurisdictions, including Australia<sup>222</sup> and Germany<sup>223</sup>. The EU Broadcasting Union, a

---

<sup>217</sup> *Ibid.*, at pg. no. 4

<sup>218</sup> **33. Registration of Copyright society.**

(1) No person or association of persons shall, after coming into force of the Copyright (Amendment) Act, 1994 commence or, carry on the business of issuing or granting licences in respect of any work in which copyright subsists or in respect of any other rights conferred by this Act except under or in accordance with the registration granted under sub-section (3):

Provided that an owner of copyright shall, in his individual capacity, continue to have the right to grant licences in respect of his own works consistent with his obligations as a member of the registered copyright society:

Provided further that the business of issuing or granting license in respect of literary, dramatic, musical and artistic works incorporated in a cinematograph films or sound recording shall be carried out only through a copyright society duly registered under this Act.

<sup>219</sup> These societies are: (1) M/s Recorded Music Performance Limited (RMPL (sound recording works))(2) CINEFIL Producers Performance Ltd. (Cinematograph Works) (3) Indian Performing Right Society Limited (musical works and literary works associated with musical works) (4) M/s Screenwriters Rights Association of India (SRAI) (dramatic works and literary works associated with dramatic works) – See the list available at [https://copyright.gov.in/Documents/Copyright\\_Societies.pdf](https://copyright.gov.in/Documents/Copyright_Societies.pdf), last accessed on September 9, 2025.

<sup>220</sup> Articles 3(2), 3(4), Council Directive (EU) 93/83/EEC of 27 September 1993 on the coordination of certain rules concerning copyright and rights related to copyright applicable to satellite broadcasting and cable retransmission, available at <https://eur-lex.europa.eu/eli/dir/1993/83/oj/eng>, last accessed on November 19, 2025.

<sup>221</sup> Recital 26, Directive (EU) 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, accessible at <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:en:HTML>, last accessed on November 19, 2025.

<sup>222</sup> The Copyright Agency (Australia), Code of Conduct for Copyright Collecting Societies (Copyright Agency, 2023), available at <https://www.copyright.com.au/about-us/governance/code-of-conduct/>, last accessed on July 25, 2025.

<sup>223</sup> Act on Collective Management Organisations and Copyright Service Providers (Collecting Societies Act, VGG) (Federal Republic of Germany), available at [https://www.gesetze-im-internet.de/englisch\\_vgg/englisch\\_vgg.html](https://www.gesetze-im-internet.de/englisch_vgg/englisch_vgg.html), last accessed on July 25, 2025.

sector in which ECL was statutorily recognised, has suggested four characteristics of an ECL system:<sup>224</sup>

- b. *Firstly*, it must strongly represent a specific category of owners within its field. It must also be capable of remunerating the owners who are members of a foreign collective rights management organisation through agreements with the latter.
- c. *Secondly*, the transparency and effectiveness in licensing conditions, tariff setting, administrative costs, and rules for distributing remuneration are other essential requirements for such rights management organisations.
- d. *Thirdly*, all rights holders concerned, members and non-members are remunerated under equal treatment rules, and any non-member owners who would not wish to be part of this agreement are free to opt out at any time.
- e. *Fourthly*, CMOs must search for any author or owners who cannot be identified or located at the time when the licence for use is granted. The remuneration reserved for this copyright owner shall be kept aside for a minimum number of years so that all chances are given for the owner to be informed and paid.

In the CL and ECL frameworks, licenses are negotiated collectively and voluntarily without statutory mandates or government-set rates and terms, though some government oversight may exist. Compared to direct voluntary licensing, the collective model typically reduces transaction costs for both content creators and potential licensees due to its collective bargaining process.

The U.S. Copyright Office has suggested considering ECL as a fallback option for AI training purposes<sup>225</sup>, after observing strong interest in a collective licensing system among copyright owners and content creators.<sup>226</sup> The U.S. Copyright Office report further clarifies that while a CL system may have its own logistical and organisational challenges, it is a better alternative to a compulsory license system that will fix royalty rates, which become difficult to undo.<sup>227</sup>

---

<sup>224</sup> European Broadcasting Union, “Extended Collective Licensing: A valuable catalyst for the creative content economy in Europe”, at pg. no. 8, available at [https://www.ebu.ch/files/live/sites/ebu/files/Publications/EBU-Extended-Collective-Licensing\\_EN.pdf](https://www.ebu.ch/files/live/sites/ebu/files/Publications/EBU-Extended-Collective-Licensing_EN.pdf), last accessed on November 19, 2025.

<sup>225</sup> Copyright and Artificial Intelligence Part 3: Generative AI Training [pre-publication version], available at <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 107, accessed on August 9, 2025.

<sup>226</sup> *Ibid*, pg. no. 104.

<sup>227</sup> *Ibid*, pg. no. 104-105.



A study commissioned by the EU Parliament, published this year, has also suggested the creation of a new EU-level statutory exception for AI training purposes, coupled with an unwaivable right of equitable remuneration to copyright owners.<sup>228</sup> The study recommends collective licensing through CMOs to administer this remuneration right.<sup>229</sup> The study further suggested legislative amendments to clarify the CMOs' role,<sup>230</sup> and that remuneration should be allotted based on a data-driven approach<sup>231</sup>. The study has even proposed a model amendment provision.<sup>232</sup> This working paper will dwell on the aspect of remuneration distribution later.

In its recommendations to the present DPIIT Committee, the Cine Musicians Association (CMA) have proposed disbursing royalties to copyright owners through CMOs, which would function under mandatory AI-specific Royalty Disclosure protocols.<sup>233</sup> The International Federation of Reproduction Rights Organisations (IFRRO) has also supported collective licensing, taking cue from such organisations in various jurisdictions like the US (CCC), Germany (VG Wort), Japan (JAC) and Australia (CA).<sup>234</sup> Similarly, the Indian Reprographic Rights Organisation (IRRO) has advocated that India adopt a multi-tiered, institutional

<sup>228</sup> 'Generative AI and copyright: training, creation, regulation', available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST\\_STU\(2025\)774095\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU(2025)774095_EN.pdf), pg. 127,129, last accessed on August 17, 2025.

<sup>229</sup> *Ibid.* at pg. 128.

<sup>230</sup> *Ibid.*

<sup>231</sup> *Ibid.* at pg. no. 129-134.

<sup>232</sup> *Ibid.* at pg. no. 135-136

*"Article XX – Use of protected content in AI model training*

*1. Notwithstanding Articles 2 and 3 of Directive 2001/29/EC, the use of lawfully accessible works and other subject matter for the sole purpose of training generative artificial intelligence systems shall be permitted, provided that such use is accompanied by a fair and proportionate remuneration to the relevant rightsholders.*

*2. For the purposes of this Article, "training" includes initial training, re-training, fine-tuning, or any process in which protected works are ingested to adjust model parameters. It excludes the inference stage in which end-users interact with a pre-trained model.*

*3. The right to remuneration shall be unwaivable and exercised collectively through collective management organisations designated by the Member States.*

*4. The amount of remuneration shall take into account the scale of use, the nature of the works used, and the commercial value of the resulting AI system or model.*

*5. Providers of generative AI systems shall submit reports indicating the general categories, types, and sources of data used, in accordance with Article 53 of Regulation (EU) 2024/1689 (AI Act).*

*6. Member States shall ensure that appropriate procedures are in place for the distribution of remuneration to rightsholders, including fallback mechanisms in cases of unverifiable use.*

*7. Where the rightsholder has embedded a machine-readable opt-out signal in accordance with technical standards adopted under this Article, such content shall not fall under the obligation in paragraph 1, unless the metadata has been removed or ignored without justification.*

*8. The Commission shall be empowered to adopt implementing acts specifying the format, interoperability requirements, and technical means for communicating such opt-outs"*

<sup>233</sup> Written Submissions by CMA to DPIIT Committee as to how the government can audit CMOs in the context of AI licensing and royalty disbursement.

<sup>234</sup> Written Submissions of IFRRO dated June 19, 2025 to DPIIT Committee

licensing framework administered by sector-specific CMOs covering licensing for dataset creation, model training, downstream use, etc.<sup>235</sup>

Some scholars have suggested that collective licensing can play a critical role, enabling AI Developers to obtain a single license covering thousands or more copyrighted works without negotiating with each copyright owner individually. Certain training activities or general categories of outputs that *need access to diffuse copyrighted materials are considered to be good candidates for collective licensing*.<sup>236</sup>

CL and ECL, thus, solve two major issues, i.e., consent and remuneration from the copyright owners' perspective, and reduce transaction costs from the AI Developer perspective. Nevertheless, CL or ECL as a solution may also exacerbate hold-out situations where CMOs, leveraging their collective bargaining position, may hold out against licensing the works for an excessive royalty rate, stymying the development of AI Systems. This adversely affects the further training of existing AI Systems and would disproportionately affect new market entrants by creating a high entry barrier. Scholars have warned that ECL runs the risk of favouring incumbent AI Developers because newer entrants without deep pockets would be unable to get the vast amounts of data easily acquired by major players. The newer entrants may, therefore, look to use data that is readily available and the use of which involves low legal risk. The issue with using such data is that it is demonstrably biased. Thus, the systems developed by the newer entrants would be of a far inferior quality, making any meaningful competition non-existent.<sup>237</sup>

CL and ECL also present other challenges for enabling AI training. Samuelson<sup>238</sup> notes that past regimes regulated rights by industry, whereas AI Systems train on diverse content types. While some AI Systems use single datasets, others are multimodal (processing images, videos, and text). This diversity makes collective licensing for each data type impractical, and a

<sup>235</sup> Written Submissions of IRRO dated June 21, 2025 to DPIIT Committee

<sup>236</sup> Daniel Gervais, Haralambos Marmanis, Noam Shemtov and Catherine Zaller Rowland, 'The Heart of the Matter: Copyright, AI Training, and LLMs', pg. 27-28, available at SSRN: <https://ssrn.com/abstract=4963711> or <http://dx.doi.org/10.2139/ssrn.4963711> , last accessed on November 19, 2025.

<sup>237</sup> Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem', 93 Wash. L. Rev. 579 (2018) at pg. 609-610, available at <https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2> , last accessed on November 19, 2025.

<sup>238</sup> Pamela Samuelson, 'Fair Use Defenses in Disruptive Technology Cases', 71 UCLA L. Rev. 1484 (2024), 1565-1569, available at <https://www.uclalawreview.org/fair-use-defenses-in-disruptive-technology-cases/> , last accessed on November 19, 2025.

universal license is unfeasible due to varying input values. The multiplicity of CMOs may increase transaction costs and the complexity of negotiations.<sup>239</sup>

Another drawback of implementing CL and ECL for training AI Systems is their treatment of orphan works.<sup>240</sup> Given the scale of data required to train AI Systems, identifying orphaned works on a case-by-case basis would be prohibitively expensive and significantly delay the usage of these works for training AI Systems.<sup>241</sup>

Thus, even if CL and ECL are considered options, they must be well-rounded to incorporate some specific statutory mechanism that would disable or discourage CMOs or copyright owners from denying access to copyrighted works. In addition, the framework must address practical problems such as administering copyrighted works in unorganised sectors and those without current copyright societies, as well as orphan works, and reducing the burden involved with multiple CMOs.

## 5. PROPOSED POLICY FRAMEWORK

In light of the above analysis, the Committee decided that it would be most appropriate to follow a ‘no absolutes’ approach in this case.

A “zero-price license”<sup>242</sup>, i.e. an outright blanket exception under law in favour of use of copyrighted materials for AI training without any payment to copyright holders, can polarise the income in the value chain for AI, reducing the incentive to human creativity. This can, in the long term, lead to underproduction of human-generated creative content, and thereby harm the broader creative ecosystem. The substitution effect<sup>243</sup> problem, as detailed above,

<sup>239</sup> Comments from Authors Alliance dated October 30, 2023 to the Policy Study on Artificial Intelligence, Docket Number 2023-6, at pg. 17-18, available at [https://downloads.regulations.gov/COLC-2023-0006-8976/attachment\\_1.pdf](https://downloads.regulations.gov/COLC-2023-0006-8976/attachment_1.pdf), last accessed on November 19, 2025.

<sup>240</sup> Orphan works are works that are protected by copyright, but the author cannot be identified or found. US Library of Congress defines 'Orphan works' as “copyrighted works whose owners are difficult or even impossible to locate.”, Orphan Works, WIPO Seminar-May 2010, Lecture Summary, at pg. 3, available at [https://www.wipo.int/edocs/mdocs/sme/en/wipo\\_smes\\_ge\\_10/wipo\\_smes\\_ge\\_10\\_ref\\_theme11\\_02.pdf](https://www.wipo.int/edocs/mdocs/sme/en/wipo_smes_ge_10/wipo_smes_ge_10_ref_theme11_02.pdf), last accessed on November 19, 2025.

<sup>241</sup> Denise Troll Covey (Principal Librarian for Special Projects), ‘Response to Notice of Inquiry about Orphan Works, Federal Register, January 26, 2005, Vol. 70, No. 16: 3739-3743’, pg. 8, available at [https://www.andrew.cmu.edu/user/troll/Carnegie\\_Mellon%20Comments.pdf](https://www.andrew.cmu.edu/user/troll/Carnegie_Mellon%20Comments.pdf), last accessed on November 19, 2025.

<sup>242</sup> See Mark A. Lemley and Philip J. Weiser, ‘Should Property or Liability Rules Govern Information?’, available at <https://scholar.law.colorado.edu/cgi/viewcontent.cgi?article=1346&context=faculty-articles>, last accessed on August 3, 2025.

<sup>243</sup> News Broadcasters and Digital Association, vide its written submissions to DPIIT Committee stated that “AI Systems that scrape and repurpose such news content often generate real-time summaries or responses, which

strengthens the arguments against introducing a blanket exception. If works generated by AI Systems are allowed to negatively impact the sales, viewership and readership of original works, and the human creators of the original works are not compensated for the use of their works in AI training, it will create an imbalance. There would be little incentive for human creators to create something new when Gen AI, which is trained on their content, competes with them and undercuts their revenue. Copyright rewards human creativity, and substitution effect undermines this incentive.

The EU model of a TDM exception *with an opt-out mechanism for rightsholders*, while offering a compromise, raises several legal and technical challenges as detailed in the preceding sections of this paper. It shifts the burden from user of the works to copyright owners. Moreover, in the absence of full disclosure of all data on which the AI Systems are trained at a highly granular level, it becomes challenging for the rightsholders who choose to opt out of the TDM exception to enforce their copyright. In other words, an opt-out right in favour of rightsholders holds little value in the absence of a corresponding transparency obligation on AI Developers to make full disclosure of all training data. This is because such disclosure is important for rightsholders to identify any violations of their opt-out right. On the other hand, a full transparency obligation, if imposed on AI Developers, can create a significant compliance burden on them, impacting ease of doing business and also raises concerns about exposing trade secrets.

As regards the voluntary licensing system via direct licensing agreements, the huge transaction costs involved in it makes it logistically impractical, given the scale of AI training operations. Moreover, if AI Developers fail to navigate the negotiations with rightsholders, it can result in lack of access to crucial content. This “*exclusion of access to certain copyrighted source materials may create or promote biased AI systems*”, encouraging AI Developers to “*use easily available, legally low-risk sources of data for teaching AI, even when those data are demonstrably biased*”<sup>244</sup>.

Full control of the rightsholders over their content and the right to refuse licensing are the key features of the voluntary licensing model. This core feature of the model raises major public

---

*directly compete with news broadcasters and publishers’ revenue streams, causing irreparable financial harm to their market, which has a short monetization window due to its time-sensitive nature.”*

<sup>244</sup> Amanda Levendowski, ‘How Copyright Law Can Fix Artificial Intelligence’s Implicit Bias Problem’, Washington Law Review, Volume 93 Number 2, available at <https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=5042&context=wlr> , pg. 579, last accessed on August 10, 2025.

interest concerns, because access to large quantities and high-quality data is critical to the development of efficient AI Systems with mitigated risk of bias. Fairness is an important element embedded within the core principles of Gen AI development. Recently, the ‘Free AI Committee Report’ released by the Reserve Bank of India on ‘Framework for Responsible and Ethical Enablement of Artificial Intelligence in the Financial Sector’<sup>245</sup> has emphasised the importance of fairness, stating it is crucial to ensure that “*AI outcomes are unbiased and do not discriminate against individuals or groups*”.

In this context, to facilitate broader access to content for training and support bias mitigation, the model of Extended Collective Licensing (ECL) was considered, which has been recommended by the U.S. Copyright Office as an alternative mechanism<sup>246</sup>. However, ECL essentially remains a voluntary licensing framework with negotiable terms. In the context of Gen AI, this may amount to imposing a significant compliance burden on those seeking licences, due to the complexity of negotiations and the associated uncertainty. In ECL systems where a CMO issues licenses on behalf of its members as well as non-members, CMOs may use their bargaining power to demand disproportionately high royalties, creating legal and financial barriers that hinder Gen AI development particularly for startups and smaller players. This could give large tech companies an unfair advantage, as they are better equipped to navigate complex ECL systems, while newer entrants are left with limited access to high-quality content and are forced to rely heavily on public domain materials, increasing the risk of biased and lower-quality AI Systems. Moreover, if the ECL system is implemented at the CMO level, where each CMO representing a particular class / sector acts on behalf of all copyright owners within the relevant category of work, the sectors lacking established CMOs will still be excluded from the system. AI Developers would face fragmented and costly negotiations with multiple individual rights holders, further complicating access. If this restricts the data available for training, this will likely result in underperforming AI Systems and reinforcement of systemic biases.

In light of this, **statutory licensing** was evaluated as an alternative. Statutory licensing simplifies access to copyright-protected content. It does not necessitate prior permission from

<sup>245</sup> ‘Free AI Committee Report’, available at <https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF>, pg. 37, last accessed on September 9, 2025.

<sup>246</sup> Copyright and Artificial Intelligence Part 3: Generative AI training, available on <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>, pg. 107, last accessed on July 29, 2025.

rightsholders. Instead, users are required to pay a pre-determined royalty calculated on government or court fixed rates, as compensation to the rightsholders. Statutory licensing also reduces transaction costs and creates a predictable environment for licensees of works. While this model takes away the power of the copyright owners to refuse licensing or negotiate a fee, it guarantees them fair compensation.

Statutory licensing is not a new experiment. It has already been implemented in sectors such as broadcasting and streaming across various countries. The concept is also embodied in current copyright law in India under Sections 31C and 31D of the Copyright Act, 1957, which relate to cover versions and radio & television broadcasting, respectively.

As regards the rationale of statutory licensing, high transaction costs have historically been considered as one of the primary justifications of government intervention and introduction of statutory licenses<sup>247</sup>. Elevated transaction cost, as detailed above, is one of the major issues in data licensing for training of AI Systems, thereby providing a strong reason to explore statutory licensing as a viable policy option. Moreover, statutory licensing performs a critical balancing function where copyright imposes broader social costs. As a scholar aptly notes<sup>248</sup>, *“compulsory licensing regimes are informed by an impulse similar to fair use: to weigh the importance of public access to creative works against the incentive function provided by market-based licensing. In this respect, just as fair use provides a zero-price compulsory license in order to selectively rebalance the incentives/access tradeoff, so too an industry-wide compulsory license can price royalties at rates explicitly designed to foster access, sometimes meaning copyright owners are compensated less than they would be in open markets. In so doing, these regimes can, and, as have, attempted to mitigate some of the social costs imposed by copyright’s exclusive rights.”* In the context of Gen AI development, easy and fair access to data is critical to the development of effective AI Systems which serve the larger public interest. Statutory licensing can facilitate such access while preserving the right of the copyright owners to receive fair compensation. It would further help startups and small players to gain easy access to content for training of AI Systems at pre-determined rates.

---

<sup>247</sup> See Jacob Victor, ‘Reconceptualizing Compulsory Copyright Licenses’, Stanford Law Review, Volume 72, available at <https://review.law.stanford.edu/wp-content/uploads/sites/3/2020/04/Victor-72-Stan.-L.-Rev.-915.pdf>, pg. 915, last accessed on August 3, 2025.

<sup>248</sup> *Ibid.* pg. 936.



According to a recently published report<sup>249</sup> commissioned by the European Parliament's Policy Department for Justice, Civil Liberties and Institutional Affairs at the request of the Committee on Legal Affairs, *"the majority of welfare gains from AI accrue to consumers, and policy should be designed to preserve these gains while ensuring that creators continue to supply the fresh, high-quality data on which future progress depends"*. The report recommends that policymakers should *"avoid opt-out and adopt statutory licensing as the default framework"* stating it to be the *"most robust option for aligning private incentives with societal welfare"*. On the royalty rate, the report recommends that *"the royalty rate should be set at the lowest level that restores maintenance of creative supply, and it should be periodically reviewed"*. The report further stresses that *"statutory licensing, carefully designed and implemented, offers a pragmatic compromise: it secures broad access, sustains incentives, and minimises costs"*.

**Though statutory licensing seems to be the best suited because it ensures: (1) compensation to copyright holders, (2) availability of all lawfully accessed content for AI training to AI Developers as a matter of right without the need for negotiations, and (3) mitigates the risk of AI bias and hallucinations by making maximum content available for AI training, yet it is not likely to serve the purpose if applied in its traditional manner. This is because when applied in its traditional sense, such as in the case of broadcasting in India, where rate setting takes years and there is a requirement of identifying, intimating and paying numerous individual rightsholders, statutory licensing would pose its own challenges in the context of large-scale AI training. Most CMOs in India have incomplete membership coverage, making it difficult for AI Developers to identify, locate, and compensate all relevant creators, millions of whom are not represented by CMOs. The requirement of reaching millions of individual rightsholders and paying them makes such a licensing system logistically and administratively challenging.**

**What is needed is a model which fulfils the following objectives:**

---

<sup>249</sup> Available at [https://www.europarl.europa.eu/RegData/etudes/STUD/2025/778859/IUST\\_STU\(2025\)778859\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2025/778859/IUST_STU(2025)778859_EN.pdf). This report further explains – *"AI firms favour exceptions or no-royalty opt-outs that maximise their short-term profits, even if they risk underfunding future creation. Creators prefer statutory licensing with higher royalties, though their own surplus is maximised at an intermediate rate rather than at the highest possible one. From a social planner's perspective, a statutory licence with a modest positive royalty is usually optimal, as it balances representativeness and freshness against static distortions"*

- (a) **Availability of lawfully accessed content for training to AI Developers as a matter of right without the need for negotiations, ensuring high-quality AI development with mitigated risk of AI bias**
- (b) **Fair compensation to copyright holders**
- (c) **Rate setting via a quick, transparent process**
- (d) **Preserving the ability of affected parties to challenge/ask for review of rates established before a judicial forum**
- (e) **Facilitates both big players, start-ups and individuals in the AI tech as well as creative industry, creating no barriers to entry**
- (f) **Upholds the basic principles of copyright and rewards human creativity**
- (g) **Mitigates risk of litigation and disputes**

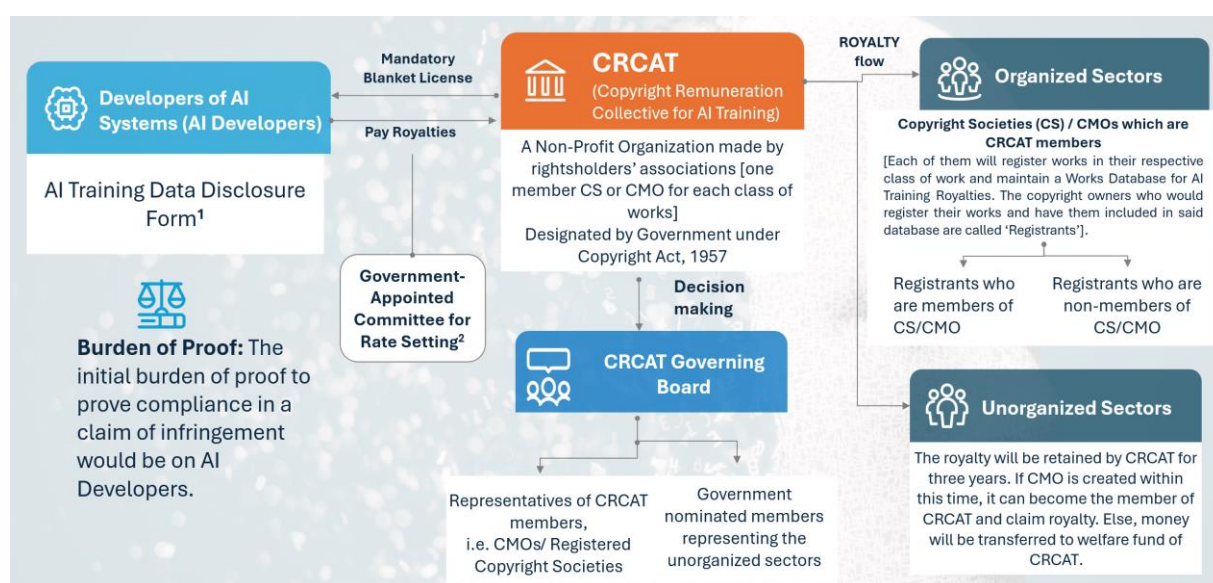
Recognising the above, a hybrid model (hereinafter referred to as “**Hybrid Model**”) was conceptualised by the Committee. This approach was found to offer a more balanced solution, addressing many of the operational and legal concerns associated with the other models.

Notably, Ministry of Electronics and Information Technology (MEITY) in its submission to DPIIT Committee, supported the Hybrid Model. Their submission is attached herewith as **Annexure D**. Nasscom, through its submissions to the DPIIT Committee dated 17.08.2025, lodged its dissent to the hybrid approach proposed by the majority of the Committee Members. In its submissions Nasscom recommended “*Text and Data Mining (TDM) for both commercial and non-commercial purposes where access is lawful, and a good faith knowledge safeguard is met, solely for the training and input processing stage of machine learning. Rightsholders should be provided clear statutory protection against TDM in two complementary ways. For content that is publicly accessible online (freely accessible without paywalls, logins, or other access restrictions), rightsholders should be able to reserve their works from TDM through a machine readable opt out, at the point of availability. For content that is not publicly accessible, rightsholders should be able to reserve their works from TDM through contract or licence terms.*” The Committee members from Nasscom put forth this dissenting view. Nasscom’s submission is enclosed as **Annexure E**. Rest of the members of the Committee opposed the recommendation to adopt a TDM model with opt-out right of copyright holders, citing the reasons set out under section 4.2 of this paper, and endorsed the Hybrid Model. Thus, Committee, with a majority view, recommends the below Hybrid Model.

## 5.1 Hybrid Model - Concept

A Hybrid Model is proposed wherein a mandatory blanket license with a **statutory remuneration right** for the creators and copyright holders would be established for the use of all lawfully accessed copyright-protected works in the training of AI Systems. The rightsholders will not have the option to withhold their works for use in the training of AI Systems. A centralised entity made by the rightsholders and designated by the Central Government under the statute, would be responsible for collecting the payments from the developers of the AI Systems. The details of the process of fixing the royalty, rate setting model, and distribution channels for dispensing the royalties to rightsholders are provided in the below paragraphs of this section.

The core philosophy of this approach is to address the challenges of large-scale data use by AI Developers while ensuring that creators are fairly compensated for the use of their works. It would ensure automatic availability of copyright-protected works for training of AI Systems and legal certainty, to help unlock the transforming potential of AI Systems for mankind. By preserving the right to fair compensation to copyright owners and administering it through a single umbrella organisation made by the rightsholders and designated by the Central Government, the Hybrid Model aims to simplify licensing procedures, reduce transaction costs, and ensure fair and easy access to content for the training of AI Systems. The chart set out below captures the proposed Hybrid Model:



**1. AI Training Data Disclosure Form** – to be submitted by AI Developers to CRCAT summarizing data types and sources, facilitating proportional royalty distribution among work categories.

**2. Government-Appointed Committee for Rate Setting** – A government-appointed committee, including legal, economic, technical experts, and representatives from CRCAT and AI Developers will set a Flat Rate.

## 5.2 Nature of license: Mandatory Blanket License

It is proposed that a **mandatory blanket license** for copyright protected works be introduced by statute (Copyright Act, 1957), ensuring that no copyright holder can withhold their works from use for training of AI Systems.

Lack of quality data leads to a high risk of AI bias and hallucinations. By mandating a blanket license, the proposed system ensures that all lawfully accessed copyright-protected content is available for use by AI Developers for training their AI Systems. This does away with the complexity involved in negotiating with multiple individual rightsholders and the possibility of hold outs hindering access to the requisite data for training.

It is proposed that *lawful access be the prerequisite* for the use of copyrighted works for the training of AI Systems. This means AI Developers would not be allowed to rely on the aforesaid mandatory license to bypass existing or future technological protection measures or to gain unauthorised access to works behind paywalls without making the necessary payment. However, once they obtain the access lawfully and make the payment for such access wherever it is required, the AI Developers will have the license to use such works for AI training without the need for seeking any further permission/consent from the rightsholders. Notably, as detailed above, even in countries like EU, Singapore, and Japan which provide TDM exceptions, lawful access is a pre-requisite for invoking such exceptions.

The lawful access requirement is proposed to be introduced with a prospective effect, and the past activities of the AI Developers can be dealt with as per the applicable existing legal framework, as interpreted by courts.

## 5.3 Collecting Entity: Copyright Royalties Collective for AI Training (CRCAT)

A single, umbrella collecting entity may be created which may be called the Copyright Royalties Collective for AI Training (CRCAT). It would be a nonprofit organization made by the associations of rightsholders and designated by the Central Government under the Copyright Act, 1957. The designation criteria can be provided in the statute, and the process can be laid by the Government under the Rules framed under the Copyright Act, 1957 (“Rules”).

### **5.3.1 Membership of CRCAT**

As regards the membership of CRCAT, only organisations can be its members. **Only one member per class of work would be allowed.** This would either be the copyright society for that category registered under Section 33 of the Copyright Act, 1957 (hereinafter referred to as “Copyright Society”), or a not-for-profit Collective Management Organisation formed by rights holders of a relevant class of work with broad representation (hereinafter referred to as “CMO”). The details of minimum representation from the relevant sector and other conditions which are required to be met for a Copyright Society or CMO to be a member of CRCAT can be prescribed by the Central Government under the Copyright Rules, 2013 (“Rules”).

All CMOs and Copyright societies which would be the members of CRCAT are hereinafter referred to as “**CRCAT members**”.

Since some unorganised sectors have no registered copyright societies or CMOs, CRCAT will not have membership from such sectors. CRCAT’s membership **can expand** with time as CMOs are created for such currently unorganised sectors. Such CMOs may become CRCAT members, if they meet the eligibility criterion provided in the Rules, and the representatives thereof can then be brought on the Governing Board of CRCAT. In the meantime, the Governing Board of CRCAT will have individual representatives of such sectors, as detailed in the following section.

### **5.3.2 Governing Board of CRCAT**

The Governing Board of CRCAT (“the Board”) will have balanced representation. It will consist of representatives from each member organization (Copyright Society or CMO). For categories of works which are currently not represented by registered copyright societies or CMOs in India, alternative mechanisms for representation on the Board would be established. To safeguard the interests of these sectors in the collection, administration, and distribution of royalties, the Central Government would consult relevant stakeholders from these sectors and nominate suitable representatives to the Board.

### **5.3.3 Role of CRCAT**

CRCAT would serve as a centralised facilitator that streamlines collection of royalties for use of copyrighted content for training of AI Systems, under a mandatory blanket license.

CRCAT would be empowered to:

- a) collect royalties from AI Developers based on rates fixed by a government appointed committee; on behalf of its member organisations and the unrepresented sectors which do not have CMOs and are represented by government-nominated representatives at the Governing Board of CRCAT.
- b) distribute these royalties to its member organisations (i.e. CRCAT members).

## **5.4 Rates and Rate Setting Mechanism**

### **5.4.1 Rate Setting Authority**

Royalty rate will be determined by a committee formed by the Central Government (“*Rate Setting Committee*”). The proposed Committee would consist of senior government officers, senior legal experts, financial or economic experts, and technical experts with expertise in emerging technologies. The Rate Setting Committee would also include a member from CRCAT and a representative of AI Developers. The composition would be designed to provide balanced and expert-driven guidance. The committee shall set rates in a time bound manner.

The Rate Setting Committee will:

- set royalty rates based on input from stakeholders, market data, and economic analyses
- ensure that rates are fair, predictable, and transparent
- review and adjust rates every 3 years, to reflect technological and market developments

Importantly, the rates set by the Rate Setting Committee may be challenged before the court and would be subject to judicial review.

Government departments have established mechanisms to manage pricing and tariff matters in key sectors. The Ministry of Railways reviews and proposes revisions to passenger fares and freight charges. The Department of Food & Public Distribution determines the Central Issue Prices (CIP) for essential food grains like rice, wheat etc. under the Public Distribution System (PDS). The Department of Fertilisers sets the Nutrient Based Subsidy (NBS) rates for phosphatic and potassic fertilisers. These are some examples among others.

### **5.4.2 Rate Setting Model**

As regards the criteria to set rates, it would be difficult to set rates basis the impact of the content and its value to the AI System. Any attempts to assess the value of each work to AI



System which is trained on such work would be highly challenging and the results may be speculative leading to multiple disputes. Traditional royalty rate formulae, such as per use / per play / per copy sold, may not help in this case because these methods rely on accurate trackable usage, like number of plays, downloads, copies, etc. The Shapley value method, which is used in some rate-setting models and is also suggested by some scholars<sup>250</sup> as a possible method to calculate royalties for training of AI Systems, may also not accurately capture the complex interactions within a neural network<sup>251</sup>, leading to potential disputes. Because training of AI Systems involves massive data consumption without direct interaction and no visible output linked to a specific data input, standard formulae of value assessment may not be helpful and are likely to fail in capturing the correct value of the works to AI Systems. In other words, accurate value assessment would be difficult as there appears to be “*no attribution technique which has been shown to work with complex Generative AI models*”<sup>252</sup>.

In light of the above, **flat rates** at this time seem to be the most appropriate model. The Committee can set a certain percentage of the **gross global revenue** (excluding taxes) earned by an AI Developer from the commercialisation of the AI System trained on such content. Consequently, there would be no need to pay any upfront remuneration / fee before using the works during the training phase, since, as with any revenue share model, payment will be due on generation of revenue, i.e., based upon the commercialisation of the AI System. Thereafter, the royalty would be payable annually on a recurring basis. On the other hand, a “up-front payment” model will likely only benefit larger established AI Developers while acting as a barrier for start-ups in India.

The rationale underlying a revenue share model is rooted in the idea of easing access to content challenges, making it possible for AI Developers to access all requisite content for training of AI Systems without first having to bear the burden of upfront payments.

<sup>250</sup> Jiachen T. Wang, Zhun Deng, Hiroaki Chiba-Okabe, Boaz Barak, Weijie J. Su, ‘An Economic Solution to Copyright Challenges of Generative AI’, available at <https://arxiv.org/html/2404.13964v2>, last accessed on August 17, 2025.

<sup>251</sup> See ISO/IEC 22989 where ‘artificial neural network’ is defined as “network of one or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output”, available at <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en>, last accessed on September 6, 2025.

<sup>252</sup> See Matthew Sag, Pamela Samuelson, and Christopher Jon Sprigman, ‘Comments In Response To The Copyright Office's Notice Of Inquiry On Artificial Intelligence And Copyright’, October 04, 2024, Published response to the Copyright Office NOI, Response to NOI 12, at pg. 26, available at <https://ssrn.com/abstract=4976391> or <http://dx.doi.org/10.2139/ssrn.4976391>, last accessed on November 19, 2025.

This would help create an environment supportive of AI development, and make it affordable to startups and small players, so that the cost of access to content does not act as a barrier for the development of the technology, which has significant potential public benefits. The suggested model will assist in affordable and inclusive AI development, while ensuring the copyright holders are duly and fairly compensated for the usage of their content, once the AI System in question generates profits.

Notably, revenue percentage-based fix rate approach already reflects in India's compulsory licensing regimes, such as the *Music Broadcast Pvt. Ltd. v. Phonographic Performance Ltd.*<sup>253</sup> case, where FM broadcasters were required to pay 2% of advertising revenue, to be distributed proportionately among music rights holders. The framework for access to biological resources and benefit sharing under the Biological Diversity Act, 2002, also serves as a helpful example. It imposes a benefit-sharing obligation for commercial utilisation of biological resources, requiring payment of a specified percentage of the annual gross ex-factory sale price of the product, excluding government taxes<sup>254</sup>.

MEITY's recommendation to prescribe a minimum revenue threshold above which royalty sharing will take effect was discussed by the Committee. The majority of members expressed that a minimum threshold is unnecessary, as the royalty is based on a percentage of revenue and is payable only upon commercialization. Startups with lower revenue would naturally pay lower royalties, and since this is a revenue-sharing model, there is no upfront cost that could act as a barrier to entry. Startups are not required to make payments before initiating AI training. Accordingly, it was decided not to recommend a minimum threshold.

#### **5.4.3 Retroactive application**

It is noteworthy that the obligation to pay royalties on the basis of revenue percentage-based rates set by the government shall apply *retroactively*. In other words, AI Developers who have already trained their AI Systems on copyrighted protected content and are earning revenues by commercialising such AI Systems would be obliged to pay the prescribed royalties.

The retroactive nature of the obligation to pay royalties is to address the past use of copyrighted materials in the training of AI Systems without prior license from the copyright owners. Many

---

<sup>253</sup> (2003) 26 PTC 70

<sup>254</sup> See Biological Diversity (Access to Biological Resources and Knowledge Associated thereto and Fair and Equitable Sharing of Benefits) Regulations, 2025 available at [http://nbaindia.org/uploaded/pdf/GNABSREG\\_2025.pdf](http://nbaindia.org/uploaded/pdf/GNABSREG_2025.pdf), last accessed on September 6, 2025, pg. 17.

AI Developers have built AI Systems which are commercially successful and generating huge revenues. In order to ensure fairness and accountability, such AI Developers must be required to pay royalties to copyright owners for past usage of their works. This is not a punitive measure, but a corrective mechanism to help create a balance in the creative ecosystem. In the absence of such retroactive application, new AI Developers will not enjoy a level playing field with the incumbent players.

## 5.5 Distribution of Royalties

### 5.5.1 Disclosure Obligation of AI Developers

Each developer of AI Systems (“AI Developer”) shall submit a declaration to CRCAT containing a Sufficiently Detailed Summary of the datasets used, to discharge their transparency obligation with respect to the data used. This declaration would be called “**AI Training Data Disclosure Form**”, and the format thereof would be prescribed by the Central Government under Rules.

The details to be provided under the ‘Sufficiently Detailed Summary’ would be prescribed by the Central Government, and the same must include following limited disclosures:

- **What is the category(s) and sub-category(s) of data under Section 14 of Copyright Act, 1957** (text, images, music, etc. divisible into classes of works as defined under the Copyright Act, 1957)
- **What is the source of data** (social media platforms, hard copy/ electronic publications, online libraries, public datasets, proprietary datasets, etc.)
- **What is the nature of data** (news, literature, entertainment, etc.)

It is important to note that the above transparency requirements pertain specifically to AI Developers in the context of copyright protection and creators’ royalties. They do not override or limit any additional transparency obligations that exist or may be introduced in future under other laws.

The Central Government may design a template to discharge transparency obligation on EU lines, yet keep it a ‘**light touch**’ format. The core philosophy underlying keeping the disclosure requirements simple and easy is to balance transparency with ease of doing business in India. The intention should be to seek only the necessary information to uphold copyright owners’ interests, without creating burdensome compliance requirements for AI Developers.

Notably, as provided in section 2 of this paper, the EU AI Act embodies a transparency obligation for Developers of AI Systems, obligating them to provide a sufficiently detailed summary of the training data. Recital 107 to EU AI Act says that *“in order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form.”*

A template<sup>255</sup> has been issued by EU Commission’s AI Office, which requires details like data sources, etc. The explanatory note to this template states that the information on data used in AI training is *“needed to facilitate the rightsholders to exercise their fundamental right to an effective remedy in the enforcement of their rights”*<sup>256</sup>. By the same reasoning, it is important for India to introduce at-least a light touch transparency obligation on AI Developers. It remains important even under the proposed framework, where the AI Developers have an obligation to pay a fixed share of their revenue, for the use of all copyrighted works in AI training. This is because the regulatory architecture suggested in this paper has a ‘lawful access’ requirement. The summary submitted by AI Developers of the training data would help rightsholders enforce their rights against any unlawful access of content. It would also help provide a basis for the distribution of royalties to different stakeholders.

<sup>255</sup> See <https://digital-strategy.ec.europa.eu/en/news/commission-presents-template-general-purpose-ai-model-providers-summarise-data-used-train-their> , last accessed on September 6, 2025.

<sup>256</sup> See the Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models, available at <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models> , last accessed on September 6, 2025, pg. 3, para 8.

### **5.5.2 Payment obligation of AI Developers**

Upon commercialisation of an AI System, each AI Developer shall, on an annual basis, set aside the total royalty amount calculated as per the flat revenue percentage-based rates fixed by the Rate Setting Committee. Such an amount will be paid by each AI Developer to the CRCAT annually.

The total royalty amount would then be apportioned proportionally by CRCAT among different classes of works, as categorized under Section 14 of the Copyright Act basis the proportion of their usage in AI training, as disclosed in the “**AI training data Disclosure Form**” by the AI Developer. This should reflect the relative extent to which each class of work was used in training the AI System, ensuring that categories of works more heavily utilised (for example, audiovisual content or music or literary publications) receive a correspondingly appropriate share of the royalty pool. This approach promotes fair compensation for creators across all types of content. The authors of underlying works in the case of cinematographic films and sound recordings, and the performers shall also be entitled to receive royalties.

In cases where the original AI Developer does not commercialise the AI System directly but instead *sells or transfers* the AI System or the training data to another party, the obligation to pay royalty will shift to the acquirer of the AI System or the training data, who must make annual payments going forward. Similarly, AI Developers (or AI System owners who have acquired the AI Systems or training data from the AI Developers) who *license out* their AI Systems or training data are expected to ensure that the royalty is paid either by themselves or through their licensees. They would need to include clear provisions in their licensing agreements addressing the royalty payment obligation.

This framework ensures that royalty is tied to actual economic benefit derived from the use of copyrighted content, promoting fairness while supporting innovation and early-stage AI development.

### **5.5.3 Works Database for royalty flowing from use of content in AI Training**

The CRCAT members would each establish and maintain an online system for registration of works, for the purpose of AI training-related royalties, named as “**Works Database for AI training royalties**”. Any copyright owner, who may or may not be a member of such CRCAT member, can register their works on such system. The process of registration may be very

simple and standardized. Only necessary details like name of author, title, host website / platform where the work is hosted, sub-category under literary works (like book, article etc.), and work's nature like news, fiction, biography, poetry, etc. may be sought. The time of creation of the work /or its date of publication can also be sought, as this information is crucial for the determination of eligibility for royalty. This would help ensure that the works which were not created or accessible at the time of the training of an AI System do not unjustly receive any share from the relevant royalty pool attributed to such AI System.

Depending on the class of works, details of any other unique identifiers which are used as a matter of normal course in a particular sector/industry for cataloguing and tracking of works may also be sought, such as:

- ISBN (International Standard Book Number)<sup>257</sup> and/ or ONIX (Books Products Information Format)<sup>258</sup> for books, audio books, e-books, mixed media publications, individual articles, art books or illustrated books, etc.,
- DOI for scholarly/research articles and for photographs as well<sup>259</sup>,
- ISRC Code for sound recordings<sup>260</sup>,
- ISWC Code for musical works<sup>261</sup>,
- IPI/CAE Numbers to identify creators and publishers of music<sup>262</sup>, and
- EIDR<sup>263</sup> or ISAN<sup>264</sup> for audio-visual content (Films, Television shows, podcasts, Radio shows, shorts, clips, edits, games, interactives).

Such members or non-members of 'CRCAT members' who register their works under the 'Works Database for AI training royalties' will hereinafter be referred to as "**Registrants**".

<sup>257</sup> The International Standard Book Number (ISBN) is a unique International Publisher's Identifier number, which is meant for monograph publications - See <https://isbn.gov.in>

<sup>258</sup> See ONIX overview at <https://www.editeur.org/83/overview/>

<sup>259</sup> Digital Object Identifier (DOI) is a unique alphanumeric string that is used to identify and provide a persistent link to digital objects, such as research articles, photographs, datasets, and other types of scholarly content - See <https://www.doi.org/the-identifier/what-is-a-doi/>

<sup>260</sup> The International Standard Recording Code (ISRC) enables sound recordings and music videos to be uniquely and permanently identified. See <https://isrc.ifpi.org/en/>

<sup>261</sup> The International Standard Musical Work Code (ISWC) is a unique, permanent and internationally recognised reference number for the identification of musical works - See <https://www.iswc.org/>

<sup>262</sup> Interested Parties Information/Composer, Author, and Publisher Codes (IPI/CAE Numbers) – International identification number assigned to songwriters and publishers to uniquely identify rights holders.

<sup>263</sup> The Entertainment Identifier Registry Association (EIDR) for audio visual works - See <https://www.eidr.org/about-us/>

<sup>264</sup> International Standard Audiovisual Number (ISAN) for audio visual works - See <https://www.isan.org>



Notably, CMOs, from different parts of the world, have the experience of managing massive databases and track work usage by monitoring where and when their members' works are used by various entities, through a combination of data-gathering methods and information-sharing agreements with other CMOs and users. They use identifying data to match usage information with specific works and copyright owners, collect usage data, and then distribute the resulting fees back to the rightful copyright holders<sup>265</sup>. CMOs operate across a diverse range of works and rights, including sound recordings, underlying lyrics and music, audiovisual works, drama, literature/books, and visual arts. CISAC – the International Confederation of Societies of Authors and Composers<sup>266</sup>, for example, has more than 228 member copyright societies spread across 111 countries and across music, audiovisuals, literature, drama, and visual arts.

#### ***5.5.4 Distribution to CRCAT members***

The royalties received by CRCAT will be distributed to each CRCAT member as per the share of each category of work disclosed in the “**AI training data Disclosure Form**”. CRCAT will publish on its website:

- a) a list of the primary classes of works under Section 14 and the sub-categories under each category
- b) the names of all CRCAT members
- c) the sectors having no CMOs, which are represented on the CRCAT Governing Board by members nominated by government.

#### ***5.5.5 Equal distribution***

Upon receipt of the royalties from CRCAT, each CRCAT member would distribute the same to Registrants, which may or may not be the members of that CRCAT member. It is important to note that only Registrants will be eligible to receive a share of the royalties. As regards the distribution formula, it may be decided by each CRCAT member and captured in their respective distribution policy governing the disbursement of royalty received for use of works in AI training (hereinafter referred to as “**AI Training Royalty Distribution Policy**”). Such a

---

<sup>265</sup> See the report commissioned by US Copyright Office on ‘Collective Rights Management Practices Around The World - A Survey of CMO Practices to Reduce the Occurrence of Unclaimed Royalties in Musical Works’, April 2020, by Susan P. Butler, Butler Business & Media LLC, available at <https://www.copyright.gov/policy/unclaimed-royalties/cmo-full-report.pdf>, last accessed on September 9, 2025.

<sup>266</sup> See <https://www.cisac.org/about/cisac-overview>

policy can be finalised and implemented only upon approval of a simple majority with a casting vote for chairman in the General Body meeting of the respective CRCAT member.

CRCAT members may choose to distribute the royalty to all Registrants on a pro-rata basis, or on the basis of the assessed value of works. These examples of methods of distribution are merely illustrative, and the CMOs may determine the most optimal solution. For instance, under the pro-rata approach, all incoming royalties will be distributed equally across all Registrants (calculation to be done basis the number of works of each Registrant). This is based on the logic that the AI System does not discriminate between the training data during the training process. On the other hand, the CMOs may also consider a value-based approach. Since the transparency obligations to be introduced vis-a-vis AI Developers would not require them to provide for the value of each work to the training of the AI System, the valuation would not reflect how the work was used in AI training itself and what was its value addition to the AI System in question, but rather the general standing or market value of the work independent of the AI context. Therefore, CRCAT members may choose to rely on an *analogy-based value assessment*, taking into account the relative value or recognition of each registered work. Sector specific indicators such as usage logs, market share, viewership data etc. can be used to assess the value of works.

For instance, if a CMO or Copyright Society for literary works category receives a total royalty pool of Rs. 10. All necessary details could be asked at the time of registration of the works, which would help calculate a work's score. For instance, weightage can be decided for factors such as those mentioned below to calculate the score.

Parameter	Weightage (%)	Description
Website Traffic	40%	Total unique views/downloads of the work from official sources
Licensing Instances	30%	Number of times the work has been licensed or published through third parties
Citations / Academic Uses	15%	Number of educational or scholarly citations of the work

Parameter	Weightage (%)	Description
Awards / Recognition	10%	Prestigious awards or institutional honours received by the work
Social Engagement	5%	Social mentions, reviews, and ratings received by the work

Works may be divided into three different categories based on the assigned scores – High value, Medium value and Low value works. Those works which score between 1 and 4 on a scale of 12 may be classified under high value category. The works which are assigned a score of 4 to 8 may be placed under medium value category, and the works carrying the score of 8 to 12 may be considered as low value works. The total royalties collected can then be divided into three royalty pools - 60% can be allocated to High Value works, 30% to Medium Value, and 10% to Low Value works.

The above is **merely an illustration**. Other value assessment methods used in a particular industry may be followed.

CRCAT members may introduce some reasonable registration requirements aligning with the principles of natural justice for inclusion of the works in ‘Works Database for AI Training Royalties’ in their Distribution Policy. The Registrants could be required to submit a self-declaration form, with the necessary details.

Importantly, the royalties will be calculated based on the number of registered works rather than the number of individual Registrants.

It may appear at first flush that collecting lump sum payments (expressed as a percentage of revenue) would present a problem or challenge that may contradict the nature of how CMOs function, i.e., collecting and distributing royalties on a ‘per-use’ or actual usage basis. However, CMOs across the world generally collect and distribute “logged royalties” and “unlogged royalties”. Unlogged royalties are not capable of being matched with specific works, because there is no usage data. They are distributed as a matter of international practice based

on a mix of survey evidence and statistically valid analogous data<sup>267</sup>. Accordingly, royalties collected under the process outlined above could be distributed to stakeholders on a best practices basis, which is reliable, fair and transparent.

#### *5.5.6 Challenges to address fake / duplicate works*

In order to ensure that genuine claimants/creators are rewarded, CMOs will be required to adopt and implement robust verification systems to filter fake or duplicate claims. Technology solutions like digital watermarking, blockchain, content fingerprinting etc. can be adopted and implemented to check the authenticity and originality of works. The distribution policies of CMOs must be transparent and based on unambiguous rules clearly reflecting the criteria using which the royalties are calculated and distributed. CMOs would need to evolve into technology enabled bodies with modern infrastructure and facilities in order to ensure fairness and accuracy in their systems. CRCAT members may face initial struggles to manage their ‘Works Database for AI training royalties’, however, these problems are not insurmountable. While the challenges are real, the same can, over time, be progressively addressed through innovative and technological solutions. There would be a need for CRCAT members to prioritise capacity building, upskill their staff, make digital infrastructure and technology investments, and undertake governance reforms.

Copyright creates livelihoods for creators. It would be a short-sighted policy decision to eliminate the creators’ right to receive royalties for their original works, owing to the implementation challenges in the distribution of royalties. The focus should rather be placed on establishing fair and efficient distribution systems, and not diluting the legitimate rights of creators. The CMOs may initially struggle to manage distribution; however, a glide path of a few years would help them grow competent enough to build the necessary capabilities to implement robust distribution policies.

---

<sup>267</sup> See Copyright Society/ CMO distribution policies – ASCAP (USA) - <https://www.ascap.com/~media/files/pdf/members/governing-documents/ascap-survey--distribution-rules--10322.pdf>, IPRS (India) - <https://iprs.org/wp-content/uploads/2025/01/Distribution%20Rules%20&%20Methods.pdf>, PRS for Music (UK) - <https://www.prsformusic.com/~media/files/prs-for-music/membership/membership-policies/prs-distribution-policy/distribution-cycles-and-concepts--prs-distribution-policy>, last accessed on September 9, 2025.

### ***5.5.7 Fine-Tuned Models trained on the specific nature of works***

The royalties flowing from developers of AI Systems which train their systems in a particular domain/nature of data and indicate the same in their 'AI training data Disclosure Form' to CRCAT shall flow to Registrants belonging to such domain of works only.

For instance, if an AI System is fine-tuned or augmented exclusively using news content within the category of literary works, and the developer of the AI System declares this in its 'AI training data Disclosure Form' submitted to CRCAT, in such a case, the entire share of revenue remitted to CRCAT as royalties will be allocated to the CRCAT members representing literary works sector. This CRCAT member will then distribute the royalties received only among those Registrants who have registered their works under the "news" category.

Notably, the royalties payable for fine-tuned model would be separate from the royalties payable in relation to the foundational model.

### ***5.5.8 Unclaimed Royalties***

CRCAT will retain the royalty collected for sectors or classes of works that are represented on its Governing Board by government-nominated members, for a period of three years. If, during this period, a copyright management organisation (CMO)/copyright society is established for any such sector, it may become a member of CRCAT and claim the royalty retrospectively. However, if no CMO is formed within three years, the unclaimed royalties for that sector shall be transferred to **CRCAT's welfare fund** for the benefit of rightsholders in those sectors. The welfare fund can be used for supporting a sector to establish a CMO or for other purposes such as setting up community studios / rehearsal spaces / co-working hubs, providing subsidized equipment or tools to creators who cannot afford them, organizing workshops or training programs to upskill creators in new technologies or trends etc. An exhaustive list of purposes for which this welfare fund can be used can be provided in detail in the by-laws of CRCAT.

## ***5.6. Grievance Redressal and Monitoring***

CRCAT, as well as each CMO or Copyright Society which is a member of CRCAT, will have a grievance redressal system in place. A grievance cell will be established, which will entertain all complaints on matters like non-receipt of royalties, conflicted copyright ownership claims etc. and resolve them within a specified time period, as prescribed under the Rules under the Copyright Act, 1957. All the decisions would be subject to judicial review.

It is strongly recommended by MEITY, in their written submissions annexed as Annexure D, to resolve issues in a manner that precludes disputes and litigation. To achieve this, clear rules and distribution policies at the Copyright society / CMO level are essential. Also, a robust grievance redressal system is needed at both the CRCAT and CMO levels.

### **5.7 Burden of Proof - Exclusion of proprietary data**

In cases where an AI Developer is sued for failing to offer the fixed royalties for the utilisation of copyrighted content, and the AI Developer denies training its AI System on any third-party content, the initial burden of proof to establish compliance would be on the AI Developer. The law would work on the presumption that the claim of the content owner is valid (subject to the usual requirement to demonstrate a *prima facie* case) until proved otherwise by the AI Developer.

Since the licensing framework is a mandatory blanket license based on a fixed percentage of revenue, AI Developers will usually have no disincentive to be transparent about the content used for training. However, there could be circumstances where AI Developers make false declarations. For instance, an AI Developer may refuse to pay on the grounds that he has only fine-tuned an existing lawfully licensed foundational model using its own proprietary content, or proprietary content obtained from a third party under a valid agreement for the purpose of training of AI Systems. The AI Developer may declare to CRCAT that no third-party content was used for fine tuning, and as far as the foundational model is concerned, he may rely on the licensing arrangement with the licensor of the foundational model according to which the licensor has the obligation to pay the royalty to CRCAT. In such a scenario, if a copyright holder has reasons to believe that the AI System has been fine-tuned using third-party data which includes his content, a legal action may be initiated against the AI Developer. In such a case, the burden of proof will lie on the AI Developer to demonstrate that only proprietary content was used and no infringement occurred. This presumption will favour the copyright owner unless successfully rebutted by the AI Developer.

In some cases, AI Developers may train their AI Systems partly on proprietary content and partly on lawfully accessed third-party data. To account for such mixed-use scenarios, the Rate Setting Committee could issue guidelines on how much of the AI Developer's global revenue should be paid as royalties. For instance, a standard rate (N%) could apply when 100% of the training data is non-proprietary. If a portion of the training data is proprietary, the payable



royalty would be proportionally reduced. A detailed guidance on this can be issued by Rate Setting Committee, which may devise a formula for the same.

## 5.8 Injunction remedy

An injunction is an essential remedy in a suit for copyright infringement. It was considered whether to specify the availability of injunctive relief (either perpetual or temporary) in this case and, if so, under what conditions. If both the copyright owner and the AI Developer are willing participants in the proposed framework, issues regarding injunctive relief may essentially become irrelevant. However, the question may still arise if the AI Developer refuses to comply with the framework or unlawfully accesses the copyrighted work.

Generally, most intellectual property laws in India do not prohibit injunctive relief, which is left to the discretion of the courts under the Specific Relief Act of 1963 and the Code of Civil Procedure of 1908. However, there are exceptions. For instance, Section 59 of the Copyright Act of 1957 related to architectural drawings, in effect, overrides the Specific Relief Act and prevents the granting of injunctive relief if it would require the demolition of a building or structure.

It was considered whether to propose a similar restriction in the context of using copyrighted works to train AI Systems. If both the AI Developer and the copyright owner are willing participants, courts are likely to take this into account when deciding on an injunction request. For instance, in cases involving Standard Essential Patents<sup>268</sup>, the willingness of both the patent owner and the implementer to enter a licensing agreement is a significant factor in determining whether injunctive relief is granted or denied.<sup>269</sup> Additionally, the likelihood of obtaining royalties through a license may indicate a lack of irreparable harm, a mandatory factor for such relief.

While a statutory restriction on injunctions could provide certainty and continuity to AI Developers, copyright owners may face challenges in enforcing their rights to seek royalty if the AI Developer refuses to comply with the proposed framework, particularly if they refuse to pay or gain illegal access to copyrighted works. This issue could be exacerbated when the AI Developer and the training are located outside of India.

<sup>268</sup> See <https://www.wipo.int/en/web/patents/topics/sep>

<sup>269</sup> See [https://icrier.org/pdf/Cellular-Standard-Essential-Patents\\_Policy-Brief-1.pdf](https://icrier.org/pdf/Cellular-Standard-Essential-Patents_Policy-Brief-1.pdf)

Accounting for these factors, it was concluded by the Committee that proposing a statutory bar on injunctive relief may be premature. The existing approach adopted by courts is flexible to accommodate all relevant factors and provides sufficient guardrails against potential abuses from either the AI Developer or the copyright owner.

## 5.9 Benefits and Suitability of the Hybrid Model

The suggested Hybrid Model addresses both creator royalties and AI scalability. Ensuring access to data through a mandatory blanket license and implementing fair royalty distribution via a centralised mechanism creates a predictable, efficient and inclusive framework. It ensures:

- **Legal Certainty:** AI Developers would be able to train their AI Systems on lawfully accessed copyright protected works without the fear of any legal repercussions. Payment of royalty would become due only upon commercialisation of the AI Systems which means there would no requirement to pay any upfront fee during the training phase. This would help startups, which would not need to incur any cost during the AI training phase for content access.
- **AI Innovation:** Provides a predictable legal environment for AI development. A prior-permission free data usage model for all lawfully accessed copyrighted content would help AI companies develop quality AI Systems. This will help mitigate the risk of AI bias and hallucinations.
- **Efficiency:** Blanket licensing reduces compliance burdens and reduces transaction costs. The only compliance burden for AI Developers is to submit a Disclosure form to CRCAT and pay the royalty calculated on the basis of rates set by a govt. appointed committee, which would be open to challenge/ review before judicial forums.
- **Fair Compensation:** Creators would be fairly compensated. All copyright owners, whether they are members of a CMO/copyright society, who register their works under “Works Database for AI training royalties” would be treated equally. They would receive royalty as per their entitled share from the total royalty pool received from AI Developers, without the need of any negotiation process with AI Developers.
- **Inclusivity:** Even unorganised or underrepresented sectors would receive a share of compensation. CRCAT will hold their royalties for three years and they can claim such share anytime by forming CMOs with broad enough representation and obtaining membership of CRCAT.

As AI Systems continue to grow in sophistication and impact, the above Hybrid Model, the features whereof are captured in a tabular form in Annexure C, offers a sustainable legal infrastructure which upholds the integrity of copyright while supporting responsible AI development.

## 6. CONCLUSION

Unlicensed use of copyrighted content to train AI Systems raises legitimate copyright related concerns. A policy framework supporting blanket statutory exception authorizing AI Developers to make use of copyrighted content without the consent of copyright owners would lean too heavily in favour of AI Developers, leading to underproduction of human generated content, in future. AI would also not be able to sustain itself without human created content as it would keep recycling the same datasets. Hence, in order to ensure cultural evolution, preservation of incentives for human creators, and development of efficient AI Systems, a balanced legal framework is required to be implemented. It would be most appropriate to craft a framework that ensures fair compensation to copyright holders, while enabling comprehensive data access for AI Developers. Hybrid Model proposed in this paper aims to achieve the same.

## ANNEXURE A

No.	Organisation
1	Indian Performing Right Society (IPRS)
2	Phonographic Performance Ltd (PPL)
3	Indian Reprographic Rights Organisation (IRRO)
5	Indian Singers and Musicians Rights Association (ISAMRA)
6	Indian Music Industry (IMI)
8	Recorded Music Performance Limited (RMPL)
9	Producers Guild of India
10	Indian Broadcasting & Digital Foundation (IBDF)
11	Digital News Publishers Association (DNPA)
12	The News Broadcasters and Digital Association (NBDA)
13	RM Radio – Uday Chawla
14	Federation of Indian Publishers (FIP)
15	Association of Publishers in India
16	CMA (Cine Music Ass.)
17	Motion Pictures Association
19	Google India
20	Open AI

<b>21</b>	Microsoft Corporation India Pvt. Ltd
<b>22</b>	Internet and Mobile Association of India
<b>23</b>	Amazon Web Services
<b>24</b>	Tata Consultancy Services
<b>25</b>	Business Software Alliance
<b>26</b>	CodeMate AI
<b>27</b>	Corrosion Intel
<b>28</b>	EcoRatings
<b>29</b>	Indika AI Private Limited
<b>30</b>	Revca India
<b>31</b>	VIDUR
<b>32</b>	BharatGen
<b>33</b>	Gnani AI
<b>34</b>	SocketAI
<b>35</b>	Meta (Facebook)

Note: Some other stakeholders also submitted representations including ANI, Culver Max Entertainment Private Limited (formerly Sony Pictures Networks India Pvt. Ltd.), Infosys, The Dialogue and Active Telugu Film Producers Guild (ATFPG).

## ANNEXURE-B

**Comparison Table of TDM Exception in Jurisdictions**

Sr. No	Criteria	United Kingdom	United States	Japan	Switzerland	EU	Singapore
1.	<b>Legal Basis</b>	CDPA Section 29A	Section 107 of Copyright Act on Fair Use	Art 47-7 (2009); amended in 2018 [Art 30(4)]	Art 24d of Federal Act on Copyright and Related Rights (CopA)	2019 CDSM Directive (Art 3 & 4)	Section 243 and 244 of Singapore Copyright Act.
2.	<b>Scope of Beneficiaries</b>	Allowed for Non-Commercial Research only.	Allowed for transformative, non-expressive use	Broad in scope – covers commercial activities too.	Scientific Research	Broad in scope Art 3: Research & Cultural Institutions; Art 4: Anyone	No limitation on beneficiaries
3.	<b>Lawful Access Requirement</b>	TDM allowed only if user has <b>lawful access</b> .	Not explicitly required, but <b>may matter in fair use assessment</b> .	no concrete information available	Must be lawfully obtained	Lawful access requirement for users + opt-out right of copyright holders	Lawful access requirement. TPMs cannot be circumvented ; pirated content not allowed



<b>4.</b>	<b>Commercial Use</b>	Not allowed u/s 29A	May be allowed only if deemed fair use (still to be clarified – jurisprudence is still evolving)	Yes commercial usage is allowed.	For basic + applied research	Art 3: Non-commercial only; Art 4: Commercial allowed unless the rightsholder opts out.	Yes, commercial usage is allowed.
<b>5.</b>	<b>Conditions, if any</b>	Must be for Non-commercial research with lawful access and lawful use.	Must be transformative, and pass the 4 factor fair use test.	Allowed only for non-enjoyment purposes	Reproduction only for technical process; lawful access required.	Lawful access required; Art 4 subject to opt-out	Lawful access required; no circumvention of technological protection measures

## ANNEXURE C

### Salient Features of the Hybrid Model

Component / Feature	Operational Approach / Structure	Intended Outcome / Effect	Problem that it addresses
Permission free access to content	A provision would be introduced in the Copyright Act allowing AI Developers to use <i>lawfully accessed</i> copyrighted content for training without prior permission or authorization from copyright owners.	Enables lawful, <b>permission-free access</b> to a broad base of works for AI training.	Resolves legal uncertainty around the use of protected works for AI Systems development.  Makes all lawfully accessed content available for AI Training, ensuring quality of AI tools is not compromised due to unavailability of content for training
Compensation Right	Copyright holders have a right to fair royalties when their works are used in AI training.	Ensures creators are fairly compensated without having to negotiate individually.	Addresses the value gap by ensuring that creators are not excluded from the AI value chain.
Rate Setting	A Committee formed by Government will determine standard licensing rates [ <i>revenue percentage based flat rates</i> ].	Rates set by a Govt. appointed committee removes negotiation burdens and prevents exploitative pricing by either party.	Ensures affordability and fairness; mitigates market failures and inflated licensing demands.
Collective Administration	A centralized collective entity (CRCAT) formed by rightsholder associations and designated by the Central Government, under the	Simplifies licensing through a single interface; reduces administrative burden on AI companies.	Resolves fragmentation of rights management and ensures structured, equitable representation across sectors.

	statute, will manage all royalty collection rights. It will include member organisations from each class of work. There will be one member from each class of work (it could either be a registered copyright society or a CMO made by rightsholders in the relevant class of work, having broad representation in the relevant sector)		
Representation of various sectors at CRCAT	<p>CRCAT would have one member from each class of works. The composition of its governing board will be as prescribed by Central Government.</p> <p>For categories of works not currently represented by registered copyright societies or copyright management organizations (CMOs) in India, alternative mechanisms for representation on the Governing Board would be established. To safeguard the interests of these sectors in the collection, administration, and distribution of royalties, the</p>	Balanced representation of all sectors.	The composition would be in accordance with the Rules prescribed by the Central Government, which would help achieve balanced representation of all sectors.

	Central Government would consult relevant stakeholders and nominate suitable representatives to the Governing Board.		
Distribution	<p>CRCAT will distribute royalty to its member CMOs or registered copyright societies (“CRCAT members”), who will further disburse them to <u>members and non-members</u> using their internal distribution policies. The royalty shall be payable only to those members or non-members who register their works with the CRCAT member. Each CRCAT member shall make a database of all registered works for the purposes of receiving royalty flowing for utilization of works in AI training. This database will be made public. The process of registration of works would be very simple requiring very limited details.</p> <p>The organisation would work on an automatic opt-in principle. For example – If an organisation ‘A’ is the</p>	Enables efficient flow of royalties across sectors.	Automatic opt-in for all rightsholders in a particular class ensures inclusiveness and removes all legal uncertainty for AI Developers.

	member of CRCAT for ‘AA’ class of works, then it would act for all rightsholders for AA class of works whether they are members of A or not.		
Unclaimed Royalties	<p>CRCAT will retain the royalties collected for sectors or classes of works that are represented on its Governing Board by government-nominated members, for a period of three years. If, during this period, a copyright management organization (CMO) is established for any such sector, it may become a member of CRCAT and claim the royalties. However, if no CMO is formed within three years, the unclaimed royalties for that sector shall be transferred to <b>CRCAT’s welfare fund</b> for the benefit of rightsholders in such <u>sector</u>.</p> <p><u>The welfare fund can be used for supporting a sector to establish a CMO or for other purposes such as setting up community studios / rehearsal spaces / co-working hubs, providing subsidized equipment or tools to creators who cannot afford them,</u></p>	Systemic usage of unclaimed royalties	This ensures the unclaimed royalties are utilised well.

	organizing workshops or training programs to upskill creators in new technologies or trends etc. <u>The purposes for which this welfare fund can be used can be provided in the detail in the by-laws of CRCAT.</u>		
--	---	--	--



## ANNEXURE D

### MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY

#### AI AND COPYRIGHT

Copyright has emerged as one of the most contentious legal and moral issues surrounding Artificial Intelligence. Copyright assigns tangible value to human creativity, protects the sanctity of labour, and rewards original thought in its varied expressions. The rise of increasingly autonomous technologies has expanded the scope and contracted the effort involved in the act of creation, placing at our disposal a spectrum of tools, resources, and increasingly, artistic styles and perspectives derived from wide-ranging samples.

This new paradigm of creativity recasts the relationship between effort and reward that underpins all human endeavour, shifting the locus of value from process to output. It questions traditional ideas of authorship, as generative models increasingly blur the boundary between originator and instrument. It also compels a new perspective on copyright doctrine, demanding a collectivist understanding of creative contribution in the vein of a shared resource, a digital topography that disrupts the language of individual ownership.

---

#### TECHNOLOGY AS AN INSTRUMENT OF PUBLIC GOOD

The Ministry has consistently encouraged the harnessing of technology as an instrument of inclusive and sustainable development, with a specific policy focus on broadening access to resources, promoting social uplift, and addressing systemic inequities.

Artificial Intelligence has shown early potential of being an accelerant on this mission, with AI-driven applications enhancing healthcare outcomes, improving agricultural productivity, expanding access to education, and strengthening public service delivery. Critical scientific and healthcare breakthroughs have been made possible by the aggregation and distillation of countless hours of human effort and expertise.

Where such demonstrable benefits can be widely deployed, there is a clear imperative to remove barriers to progress. At the same time, not all AI systems are created equal, nor are they designed with altruistic intent. Public good, while a worthy objective, cannot become a qualifying criterion for innovation, or unconditional immunity against legal challenges. The potential for societal benefit is the catalyst behind enablement of frontier technology, but the

governance of the technology follows from its use-cases, through the specific conditions of its development and deployment.

Therefore, even as there is a compelling case for the encouragement of Artificial Intelligence as a paradigm-shifting technology, there is, equally, a case for ensuring that the rights of those whose innovative strides, original thought and creative labour form its backbone are upheld. The challenge before us is to strike a balance: to ensure that the transformative potential of AI is fully realized while upholding the sanctity of creators' rights and contributions.

In order to build an indigenous, self-reliant AI ecosystem in India, we need to ensure access to wide-ranging, high-quality bulk datasets. This is imperative not only for improving the accuracy and efficacy of AI models, but also for ensuring fair representation and inclusivity. The latter is contingent on the holistic and enthusiastic participation of the people it is meant to represent and serve.

The key to facilitating this participation is striking a conscientious balance between AI training needs and copyright holder rights. This involves ensuring broad, unencumbered access to datasets with a low compliance burden on one hand, while ensuring fair recognition of the value of creative labour on the other.

For decades, governments around the world have grappled with Copyright in light of evolving technological paradigms, for the very mechanisms that protect creative work can also impede it, particularly as modes of production and distribution undergo radical change. The digital revolution brought many such disputes and questions to the fore, and AI is but an extension of the same technological continuum. As we enable the next iteration of cutting-edge technology, the frameworks and nuances of copyright protection and compensation may suitably evolve, but they remain fundamental concepts that sustain a society's creative impetus, and must be upheld in unequivocal terms.

## **REMARKS ON THE PROPOSED MODEL**

The Ministry is aware that the DPIIT committee has, through the course of several weeks, evaluated various models and mechanisms to ensure a fair balance between training needs and copyright holder rights.

What has been paramount to the Ministry is that its core objectives with respect to model effectiveness, bias mitigation, representation and valuing creative labour are met, in alignment with the broader impetus of enabling frontier technology for public good.

After careful examination of the hybrid model of statutory licensing with revenue-based compensation proposed by the Committee, the Ministry is of the opinion that the proposed model has the potential to equitably meet its multifaceted objectives across the domains of technological innovation and creative labour.

The model enables wide-ranging training access for AI developers while ensuring proportionate compensation for copyright holders. While operationalizing this framework, the Ministry is of the view that CRCAT should prescribe a minimum revenue threshold above which royalty sharing will take effect. Furthermore, it is recommended that CRCAT put in place a robust mechanism for revenue sharing with an anticipatory and proactive approach to potential questions on the quantum of allocations. Participatory and communicative leadership in this regard will help amicably resolve issues in a manner that precludes disputes and litigation.

In conclusion, the Ministry of Electronics and Information Technology is aligned with the recommendations made by the DPIIT Committee on AI and Copyright (Part 1).

## ANNEXURE E

**nasscom**

SUBMISSION TO THE DPIIT COMMITTEE ON COPYRIGHT AND AI

AUGUST 17, 2025

### Table of Contents

I. Recommendation: A balanced Text-Data-Mining framework for India .....	1
II. From Enabling TDM to Calibrating Safeguards: Evolving the Balance of Innovation and Rights.....	3
III. The Debate: Is TDM a real copyright infringement or (at most) a technical infringement?.....	4
IV. Order of Policy Options: Where is the place for Statutory Licensing? .....	6
V. Notes on Recommendations .....	9

Strengthening the generative AI ecosystem requires access to diverse Indian datasets and holds the potential to generate broad societal benefits across healthcare, agriculture, education, financial inclusion, defence and justice. Evidence indicates that some of the most widely relied upon sources of public knowledge and discourse, which are essential for building reliable AI models at scale, are also among those that most frequently impose restrictions on access. This highlights the importance of an equitable framework that provides legal certainty for rightsholders while ensuring training data remains sufficiently broad and representative. Updating copyright law to reflect the realities of AI training is an essential step towards achieving this balance.

#### I. Recommendation: A balanced Text-Data-Mining framework for India

- **Core permission:** India should permit Text and Data Mining (TDM) for both commercial and non-commercial purposes where access is lawful, and a good faith knowledge safeguard is met, solely for the training and input processing stage of machine learning.
  - TDM exception should be without prejudice to applicable laws that protect specific categories of data, including personal data and confidential data.
  - The good faith knowledge safeguard should protect the TDM user on the lawful access test, if the user “does not know” that a source is infringing. However, in cases where the source of data used for TDM presents a heightened risk of unlawful access, the protection should only be available if the user “has no reasonable grounds to suspect that the source is infringing”.
- **Rightsholders’ safeguards and public good carve out:** Rightsholders should be provided clear statutory protection against TDM in two complementary ways.
  - For content that is publicly accessible online (freely accessible without paywalls, logins, or other access restrictions), rightsholders should be able to reserve their works from TDM through a machine readable opt out, at the point of availability.
  - For content that is not publicly accessible, rightsholders should be able to reserve their works from TDM through contract or licence terms.
  - **Public good carve-out:** To serve public good objectives, the law should also enable the government to notify specific entities or class of entities and purposes for which these reservations through opt-out restrictions would not apply, for example qualifying research and cultural heritage activity or other public interest uses, provided access is lawful. It is clarified that this public good carve-out would not apply content that is not publicly accessible.
- **Transparency obligations:** India should take note of the EU practice on public summaries of training content and related transparency tools, which aim to improve clarity for rightsholders. A measured path is to monitor early EU implementation and conduct a time bound review that assesses enforcement value, administrative cost for developers and rightsholders, impact on confidentiality and trade secrets, and interoperability with

machine readable opt-out signals. If the review shows clear net benefits, India could consider measured, phased adoption with proportionate safeguards. In the meantime, immediate protection for rightsholders should rest on lawful access, machine readable opt-outs for publicly available online content, and contract based opt-out for non-publicly accessible content, so that rights are meaningfully protected while evidence on transparency measures matures.

Taken together, this approach aims to be balanced and workable.

- Rightsholders would have a clear opt-out for publicly available online content, continued protection of access controls, proportionate security and integrity measures, and sensible limits on onward sharing.
- Developers would have a practical duty of care grounded in lawful access and good faith, clearer rules on retention and sharing, and protection against overbroad contract terms that may seek to restrict TDM.

If implemented well, the framework should support high quality AI training while safeguarding legitimate rights and providing greater operational certainty over time.

### **Industry Practices**

There is a visible set of industry practices that aim to respect site preferences, rightsholders choices and to avoid unlawful sources. Many AI providers already participate in transparency measures in the form of model cards and transparency notes.<sup>1</sup> At the same time, compliance is believed to be uneven in practice. Industry practices include:

#### **Website access files and crawler compliance**

- Website instructions such as robots.txt are commonly used to manage automated access, but they are not a security control and compliance depends on the crawler. Google's documentation states that robots.txt manages crawler traffic and that instructions cannot be enforced against all crawlers, so publishers often combine robots.txt with contractual terms and access controls.
- Common Crawl states that it respects robots.txt, supports crawl delay, and provides a straightforward method to block its crawler, which offers an example of declared practice for a large public crawler.

#### **Developer published crawler controls**

- Major model developers publish pages that describe their crawlers and how site owners can allow or disallow them. OpenAI lists its named crawlers and provides robots.txt examples for permitting or blocking access.
- Anthropic provides equivalent instructions and notes that its crawlers support the crawl delay field in robots.txt and do not attempt to bypass access barriers such as CAPTCHAs.
- Google documents a separate control called Google Extended that lets publishers decide whether their sites are used to improve certain generative features and explains how to reference this setting in robots.txt alongside its general crawler documentation.

#### **Mechanisms beyond robots.txt**

- Some publishers use machine readable notices intended for text and data mining choices. The World Wide Web Consortium community report on the Text and Data Mining Reservation Protocol sets out a method for expressing a reservation and for pointing to licensing policies, including through a well-known policy file. The report explains its relationship to European Union law on commercial text and data mining.
- Creators also use registries. Spawning documents a Do Not Train registry and an application programming interface that dataset builders can query to exclude listed works.

#### **Provenance and transparency tools**

- The C2PA specification for Content Credentials explains how verifiable provenance information can be attached and validated.
- In May 2024 C2PA announced that OpenAI had joined its steering committee, which indicates developer engagement with provenance infrastructure.

#### **Licensed and permissioned sources for training**

- Some providers set out that they use licensed or permissioned sources and public domain materials. Adobe's Firefly frequently asked questions state that current models were trained on licensed content such as Adobe Stock as well as public domain content.



- Shutterstock describes data licensing and a contributor fund that compensates contributors when their works are used for training, and it has disclosed multiyear licensing arrangements with model developers.
- There is also direct licensing between developers and publishers. The Financial Times announced a strategic partnership and licensing agreement with OpenAI in April 2024, and the Associated Press announced a collaboration with OpenAI in July 2023.

#### **Independent reporting on adherence**

- Independent reporting shows that adherence in the field is uneven. Reuters reported in June 2024 that multiple companies were bypassing robots.txt on publisher sites according to a licensing analytics provider.
- WIRED reported that one market participant appeared to access material despite robots.txt exclusions and that the behaviour drew further review by infrastructure providers.
- A factsheet from the Reuters Institute for the Study of Journalism found that by late 2023 a substantial share of leading news sites across several countries had chosen to block AI specific crawlers.

#### **Dataset review and remediation**

- Large open datasets used in research have required safety review and remediation. The Stanford Internet Observatory reported in December 2023 that the LAION five B image link dataset included hundreds of known instances of child sexual abuse material.
- LAION later announced an updated release after a safety review that removed known links to suspected abusive material.

Refer **Annexure I** for source references to current industry practice on website access preferences and source due diligence for model training.

Taken together, the practices noted above show a mixed but maturing landscape. There are documented methods for expressing website preferences and for organising licensing, as well as growing use of provenance tools. There is also risk of uneven adherence and the need for ongoing screening and corrective action at scale.

A proportionate framework can therefore build on practices that are already documented while addressing the gaps that independent reporting and research have identified.

We further recommend the government encourage development of industry practices that meaningfully addresses rightsholders' concerns after due consideration of practical limitations. The government can add value by educating rights holders about their rights, the worth of their works, and options for reserving those rights. Should the government wish to support voluntary collective licensing by aggregating culturally relevant Indian datasets, such efforts could enhance the licensing marketplace and promote the inclusion of Indian context in AI models.

## **II. From Enabling TDM to Calibrating Safeguards: Evolving the Balance of Innovation and Rights**

Copyright laws were crafted in an era, when the "right to reproduction" was synonymous with an author's control over distribution of physical copies. It was not problematic that copyright laws treated mere *access, copying and storage* of a work as infringement. These were viewed as part of a single continuum ending with "*communication to the public*", with no other legitimate purpose. Granting rights at each stage allowed authors to intervene before a work went public. It therefore made sense to give authors exclusive control over every step, as unauthorised access, copying and storage were seen as being inseparable from unlawful distribution.

While this principle still applies to most modern digital uses, such as streaming, downloads, and online sharing, certain digital-age scenarios, particularly the creation of temporary or incidental copies in processes like AI training, raise questions, the original law did not fully anticipate. This is also true for the Indian Copyright Act (Refer **Annexure II – Mapping AI Training Workflows to Potential Copyright Infringements**).

AI training requires accessing and, in some form, copying and storing data, but without showing, distributing or transmitting it to people. These are purely machine-only steps to help models learn patterns. Interpreting the law



to treat each technical step as an infringement would create legal uncertainty or require mandatory licences never intended.<sup>ii</sup>

Protection from unauthorised distribution and publication remains unquestionable, with copyright laws continuing to safeguard creators at the Generative AI output stage. However, to foster AI driven innovation while respecting creators' rights, it is essential to recognise that mere access, copying and storing can serve a new purpose unrelated to communication or transmission and should not be treated as infringement. This recognition is already reflected in the fundamental position adopted by the United States, EU, UK, Israel, Japan and Singapore.

Therefore, the substantive discussion should focus on the merits and forms of additional rightsholders' safeguards and ensuring that TDM is effective. The question, whether TDM,

- constitutes a *real copyright infringement i.e.*, it does lead to unauthorised distribution or
- it serves an *unrelated new purpose or a transformative purpose* and does not lead to unauthorised distribution and thus, it is *(at most) a technical infringement*,

directly informs the substantive discussion by shaping the rationale, scope, and necessity of any additional safeguards or rights that may be considered for creators, including compensation.

It is within this context that we set out our recommendations for enabling TDM, together with the safeguards we believe would be proportionate and appropriate.

### III. The Debate: Is TDM a real copyright infringement or (at most) a technical infringement?

There are approximately 51 lawsuits worldwide touching on copyright and generative AI.<sup>iii</sup> They cluster around two questions.

- First, *copyright grants rightsholders the exclusive right to reproduce their works.<sup>iv</sup> If AI training converts source works into mathematical representations and produces statistically derived outputs with no direct, traceable link to any one work, should that be treated as copyright infringement?*
- Second, *even if training is non expressive, are rightsholders entitled to compensation for the inclusion of their works in training datasets, and if so on what basis and in what amount?*

These questions are unpacked through four concerns:

- *Models "memorise" data which amounts to the act of making copies and storage and thereby reproduction.*
- *Copyright infringement is based on strict liability and there is no place for intentions. So even if copying and adaptation are happening without an intention to reproduce or disseminate, it amounts to infringement.*
- *AI models have used the copyrighted content without the consent of copyright owners and hence, they are entitled to remuneration/compensation for use of their data in the training stage.*
- *Machine learning operates fundamentally differently from human cognition and at exponentially greater speed, make it inappropriate to extend fair-dealing exception to AI models. A further concern is that, once trained on copyrighted material, these models would enter the same or adjacent markets, directly competing with and commercially harming the original creators.*

#### How the training step relates to reproduction

Machine learning systems create transient technical copies to make inputs computable. These copies enable computation and are not made for communication or distribution to the public. Whether models memorise training data remains debated; models are not intended to memorise, although some memorisation may occur under certain conditions. Directionally, the training step is best understood as non-consumptive and intermediary. In that sense, treating temporary, intermediary copies made solely to enable training in the same way as traditional copying whose purpose is communication to the public risks missing this distinction. As courts have observed, the point of training is to learn patterns rather than to replicate or supplant particular works.

#### Lawful access and practical compliance

Copyright is largely strict liability. Nothing in this analysis dilutes the lawful access baseline. Lawful access should apply whether material is publicly available online, accessed under subscription, or obtained by licence or other authorisation. At modern training scale it is not practical to check every item or to identify the rights behind each individual page, image or file. A purely mechanical strict liability posture for every intermediate copy is in tension with the objective of protecting uses that are only for training.

A more effective approach focuses on clear, observable rules.

- For publicly available online content, a machine readable opt out at the point of availability provides a signal that automated systems can detect and respect.
- For content that is not publicly accessible, access and use are governed by contract and licence.

Breaches of these rules should be treated in the usual way. If a valid opt out is ignored or access controls are bypassed, the use is unlawful, and any training protection does not apply. The rightsholder may then seek the standard remedies available under copyright or contract, including injunctions, damages, and termination of access.

### Consent and compensation

Over centuries, the basic purpose of copyright law has been to protect original expression from unfair copying and sharing with the public. Training use is incidental, non-expressive and computational. It is distinct from communication to the public. Against that backdrop, consent and compensation questions are best assessed through scenarios.

- *Scenario 1:* Opt-out is in force. Where a website clearly expresses a machine readable opt out at the point of availability for publicly available content, crawlers should respect it and cease use for training. Given the automated and large-scale nature of collection, seeking case by case consent is not practical at scale. If a valid opt out is ignored, any training that follows should not benefit from training protections and ordinary remedies would follow.
- *Scenario 2:* Contract or licence applies. For content that is not publicly accessible, access and use are governed by contract or licence. Parties may agree terms and price, addressing consent and compensation by agreement.
- *Scenario 3:* No opt out is present. Where content is publicly available online and no opt out has been expressed, training proceeds on that basis. In such cases, no additional statutory consent or compensation would ordinarily arise, without prejudice to voluntary licensing, and always subject to lawful access and observance of access controls.

### Market impact lens and piracy distinction

Broader concerns that AI could reshape or dominate creative markets are appropriately examined under competition law, alongside adjacent regimes such as copyright, consumer protection, data protection and personality rights. To date there is no clear evidence that generative AI has materially undercut creators' revenues at a market wide level. Japan's experience with a broad text and data mining exception is often cited as context, as its content market remains among the largest globally, although correlation does not establish causation.

Training and piracy differ in legal character and market effects:

- Training is an intermediate, non-consumptive process that does not communicate works to the public, whereas
- piracy substitutes for licensed consumption and directly diverts demand.

Potential impacts from training, if any, would arise through downstream use of model outputs. Current methods do not reliably attribute any given output to particular source works or quantify source specific revenue effects.

At the same time, specific risks merit attention. For example, memorisation or reconstruction of protected content could lead to verbatim or near verbatim outputs. This needs to be addressed at the AI output stage and is increasingly, being done. This kind of risk is also not a matter of key policy debate on this subject.

There are other risks that may not be so objectively assessable. For example, would outputs function as close substitutes for a work and displace commissions, licences, or sales? Or a sustained fall in licensing demand for training material affect upstream investment by publishers/ producers? A sustained fall in licensing demand could possibly occur if developers rely on publicly available content without opt-outs being widely used, or if legal uncertainty postpones deals; by contrast, clear lawful access rules with machine readable opt outs for publicly available content and contracts for non-public content could channel demand toward licensed, higher value datasets.

**The key point here** to consider is that **any claim of such output related harm will need to be established by market evidence or decided case by case, with courts looking for evidence** of reproduction of protected expression, the degree of similarity, the availability and use of the original work, and measurable substitution in relevant markets, including lost sales, licences, or commissions.

**Annexure III: Detailed analysis of four concerns on AI training and copyright in India** sets out analysis under the following headings: memorisation and reproduction; strict liability and lawful access; consent and compensation; market impact.

#### **Public good carve-out for text and data mining**

To guard against unintended effects on public interest activity, the law can also consider **allowing the government to notify specific entities and purposes** for which opt-outs do not apply for publicly accessible content, subject to lawful access and other safeguards.

### **/. Order of Policy Options: Where is the place for Statutory Licensing?**

#### **Case for Statutory Licensing:**

- **Opt-outs are unworkable:** Opt-out mechanisms are likely to be ineffective (*technical protocols such as robots.txt are not enforceable, lack of awareness among copyright holders about opt-out measures*)
- **Voluntary licensing is unworkable:** AI developers will have minimal incentive to enter into licensing agreements with numerous small content creators. Voluntary licenses are prevalent only between AI developers and big enterprises because corporations can pursue litigation in the event their opt-outs are not respected. In other words, the deterrence of litigation would be the 'prime' motivation for AI developers to respect opt-out measures. Therefore, voluntary licensing is not a realistic option and hence, statutory licensing is the only solution to ensure fair and just compensation for all types of content creators.
- **Compensation scheme would be fair and implementable:** Compensation under statutory licensing would be based on rates for classes of work being pre-determined by the government (like a compulsory flat-fee structure).
- **Enables TDM:** Statutory licensing would permit AI developers to perform TDM on eligible published works without the rightsholders consent, so long as they meet all legislated safeguards, including royalties, attribution, integrity, and record-keeping requirements.
- **Reduces transaction cost through compulsory collective licensing:** Bilateral licensing in the context of training generative AI is likely to be a prohibitive exercise because it would involve significant number of agreements, resulting in high overheads.

*The above line of inquiry reflects the recommendations in a recent study commissioned by the European Parliament's Policy Department for Justice, Civil Liberties and Institutional Affairs at the request of the Committee on Legal Affairs. The study has recommended that introducing a statutory remuneration scheme is essential to maintain fairness towards creators. However, the study is not the official position of the European Parliament.*

#### **Case against Statutory Licensing:**

- **Lack of Evidence:** Advocating for statutory licensing skips a fundamental question and assumes that *AI training* is a real copyright infringement i.e., it does lead to unauthorised distribution and does not serve an unrelated new purpose. It also assumes that opt-outs are unworkable and the market for voluntary licensing is one-sided in favour of AI training. The flaws in the position that *voluntary licensing is unworkable* is



discussed further in **Annexure IV - Gaps in the Argument that Failure of Voluntary Market Justifies Statutory Licensing**.

- **The Digital Competition Experience:** It is instructive to recall that when the Government introduced the **draft Digital Competition Bill (DCB)** last year like the model followed in the EU, the ensuing debate underscored the need to couple any proposed ex-ante digital-market rules with rigorous, evidence-based market studies to ensure proportionate and targeted interventions. After several months of speculation around the DCB, last month, the government finally indicated that it aims to establish an “evidence-based foundation” through market studies for the proposed DCB. The need for such market studies was again highlighted in the recommendations of the recent Standing Committee on Finance (2025-26) before introducing strong measures like ex-ante.<sup>v</sup>
- **Global Experience:** Unlike the above-mentioned case of ex-ante laws for regulating competition concern in the digital markets, where there are precedents in other jurisdictions, in the case of Generative AI and copyright debate, no country has arrived at a **conclusion** that statutory licensing framework is the way forward. To illustrate:

- **Europe:** Even EU, where its Parliament's Policy Department for Justice has advocated for a statutory remuneration scheme, has not chosen to follow this path.

In December 2024, the Council of the European Union released summary of Member States' contributions to the Presidency's policy questionnaire on the relationship between generative Artificial Intelligence and copyright and related rights.

One of the critical findings clearly showed that there were different views on setting up of guaranteed remuneration schemes for TDM activities in the context of Generative AI. A number of Member States and stakeholders were of the view that the existing legal national and EU norms and the current opt-out regime provide sufficient guarantees to protect the pecuniary interests of the rightsholders and that remuneration schemes, which consist of a limitation of the right holders' exclusive rights, should not be introduced without a proper assessment and evidence of the existence of a market failure.<sup>vi</sup>

- **United States:** The recent 2025 report of the U.S. Copyright Office which favours copyright holders in the context of generative AI and training data, **refrains from recommending immediate statutory licensing or other compulsory interventions**. Rather it advises caution and encourages the establishment of scalable licensing mechanisms, *private or collective*, to ensure rights holders are fairly compensated and maintain meaningful control over the use of their works.<sup>vii</sup>

While the analysis is U.S. focussed, the underlying economic logic about information asymmetry in a government set pricing is globally true. Such regimes require regulators to set licensing terms (*like flat or uniform rates, example: a per page or per token fee*) without full knowledge of stakeholders' valuations, leading to mispriced rates that cause distortions far worse than market-driven negotiations.

- **Spain:** The recent proposal for forced extended collective licensing (ECL) for AI training in Spain was withdrawn primarily because of widespread and vocal opposition from rightsholders, creators, and the cultural sector. It was argued that ECL would erode their exclusive rights, as it would allow collective management organisations to license their works without their direct or express consent, even if they were not members of such bodies. This instance underscores the real risk that such statutory frameworks can backfire if not grounded in robust, evidence led decisions and safeguards.<sup>viii</sup>
- **Unintended Effects:** Statutory licensing would have the unintended effect of closing the doors on private negotiations/voluntary deals. This, combined with the near impossibility associated with creating a compensation scheme that can be fair to different classes of rightsholders, could mean that a statutory licensing would end up being unfair to rightsholders and it would also take away their right to exercise an opt-out. Further, there are questions around – *how exactly will such a scheme be implemented? What machinery will need to be established to manage this? How effective will it be?*

The recent report by the **Office of the Chief Economist at the U.S. Copyright Office** reinforces this caution. It emphasises that the transaction cost reductions attributed to collective intermediation stem from centralised negotiation processes, not from the compulsory nature of licensing. Consequently, the purported benefits of statutory licensing are not unique or sufficient justification for its adoption, as these advantages can be realised through less intrusive, voluntary collective frameworks.<sup>ix</sup> *The challenges with implementing statutory licensing are further discussed in **Annexure V - Challenges with Implementing Statutory Licensing**.*

- **Order of Policy Options:** In the scheme of policy choices, it would not be prudent to impose the most stringent measure without testing other more reasonable measures for their appropriateness. Recommending statutory licensing in such a scenario would be fundamentally flawed and economically unsound, not to mention the significant challenges and unintended outcomes that can be foreseen in implementing it.

We, therefore, submit that **in the absence of any market studies or existing credible evidence**, any attempt to introduce statutory licensing of copyrighted material for use in AI model training could operate, in effect, as a **tax or levy on innovation**. The core rationale for amending copyright law to treat model training as fair dealing is precisely to avoid excessively encumbering technical, non-expressive uses.

## V. Notes on Recommendations

### i. Comparison of proposed recommendations with positions developed in other countries.

Globally, countries with open-ended exceptions, like United States and Israel, have relied on existing laws to clarify/ enable AI training within their copyright framework. On the other hand, countries with narrow/closed list of exceptions like UK, Japan, Singapore, and EU, have amended their copyright laws to specifically include TDM.<sup>x</sup> The scope of existing TDM exceptions in these countries is given below, along with our recommendation for India:

Jurisdiction	Non-commercial TDM	Commercial TDM	Rightsholder opt out (for example, robot.txt)	Lawful access required
UK	Yes <sup>1</sup>	No	No	Yes <sup>2</sup>
Japan	Yes <sup>3</sup>	Yes <sup>3</sup>	No	Limited <sup>4</sup>
EU	Yes <sup>5</sup>	Yes	Yes, for commercial use	Yes <sup>6</sup>
Singapore	Yes <sup>3</sup>	Yes <sup>3</sup>	No	Yes <sup>7</sup>
<b>Recommendation for India</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes, for commercial use with a public good carve out</b>	<b>Yes</b>

Jurisdiction	Recognition of reasonable efforts	Contractual override recognised	Technological protection measures recognised
UK	Yes	No	Yes <sup>8</sup>
Japan	Recognition of reasonable efforts	Yes	Yes
EU	Yes	For commercial TDM for works publicly available online such as a licensed database	Yes <sup>9</sup>
Singapore	Recognition of reasonable efforts	No	Yes
<b>Recommendation for India</b>	<b>Yes</b>	<b>For commercial TDM for works publicly available online such as a licensed database</b>	<b>For commercial TDM</b>

#### Footnotes and explanatory notes

- UK non-commercial research:** Article 29A permits text and data mining for research carried out for a non-commercial purpose. UKIPO guidance explains that the safeguard attaches to the purpose of the mining activity rather than to its outputs. In practice this means the mining itself cannot be done for profit, but the results may be commercialised once research is complete.<sup>xi</sup>
- UK lawful access:** Although not defined in the statute, UKIPO guidance states that lawful access means a legal right to read the work, for example through a subscription or an open licence.<sup>xii</sup>
- Japan and Singapore scope:** These jurisdictions do not distinguish between commercial and non-commercial purposes for TDM. This broader baseline applies to both permissions in the table.<sup>xiii</sup>
- Japan proviso and practical limits:** The proviso to Article 30-4 restricts the exception where the activity would unreasonably prejudice the copyright owner. The Council for Cultural Affairs notes that this covers reproducing fee-based databases designed for analysis without compensation and training that avoids technical measures.<sup>xiv</sup>
- EU non-commercial research institutions:** Article 3 creates a mandatory exception for research organisations and cultural heritage institutions engaged in scientific research on a not-for-profit basis or under a public interest mission. Recital 12 clarifies that public private partnerships are included unless a commercial undertaking exercises decisive influence that could lead to preferential access to results.<sup>xv</sup>
- EU lawful access:** Recital 14 defines lawful access to include subscription or contractual licence, open access policies, and content freely available online.<sup>xvi</sup>
- Singapore lawful access:** Section 244 provides illustrations that lawful access does not cover bypassing paywalls or breaching database terms of use, which offers clearer boundaries for users and rightsholders.<sup>xvii</sup>
- UK technological measures:** UKIPO guidance allows publishers to apply reasonable technical limits to protect stability and security, such as download speed caps or request frequency controls, provided these do not stop or unreasonably restrict TDM.<sup>xviii</sup>



9. **EU technological measures:** Article 3 paragraph 3 of Directive 2019 790 allows rightsholders to apply measures to protect the security and integrity of databases while prohibiting measures that undermine the TDM exception.<sup>xix</sup>

## ii. **Scope of TDM exception: Both Commercial and Non-Commercial Use**

No major economy is proposing a non-commercial-only TDM model. Recent reviews broadly favour, purpose-neutral exceptions, seeing a pure non-commercial carve-out as too restrictive for AI development.

- **UK:** UK allows TDM only for research with a non-commercial purpose and there does not seem to be a stakeholder consensus in the UK on this approach.
- **Europe:** The EU follows a dual-track approach: accredited research and cultural institutions enjoy an unconditional right to undertake text-and-data mining, while all other users may proceed only if rightsholders have not issued a standard machine-readable reservation.
- **Singapore:** Singapore applies a purpose-neutral regime: both commercial and non-commercial actors may conduct text-and-data mining provided they either have lawful access or, acting in good faith, could not reasonably have known the source was infringing, and they use the copies solely for analytic purposes. This notice-and-knowledge test maintains a workable diligence duty across the board, promoting innovation while still shielding rightsholders from clear misuse.
- **United States:** The U.S. allows both commercial and non-commercial TDM, but only as far as each instance clears the fair-use and contract hurdles. There is no blanket TDM comparable to Singapore's notice-and-knowledge test or the EUs opt-out regime.

For India, it would be important to enable TDM for both commercial and non-commercial use. India's research landscape blends public and private funding; imposing separate TDM regimes for 'commercial' and 'non-commercial' use would introduce debilitating legal ambiguity. Mixed-funding projects and evolving end-uses would oblige universities, start-ups and SMEs to second-guess their legal footing, dampening both academic inquiry and industrial R&D. Universities stress that knowledge transfer relies on unrestricted analytic access, and limiting the exception to non-commercial activity would ultimately penalise spin-offs and breakthrough health research without materially enhancing rights-holder protection.

## iii. **Ensure Safeguards for Legitimate Interests of Copyright Owner**

A blanket TDM exception is not desirable as it may result in some unintended consequences. Even, *Japan, which has minimal restrictions, bars use that "unjustly harm" a rights-holder's interest, showing that a broad exception can sit beside appropriate guardrails. The United States Copyright Office's 2025 economic report, cautions that "credibility issues aside, how we treat AI ingestion will have implications for the level of incentives faced by human creators."* Developers share this worry - for maintaining adequate incentives for human creators, given concerns about decline in non-synthetic content available for AI development.<sup>xx</sup>

The exercise of rights through opt outs carries significant implications for the quality and representativeness of AI training datasets. As the Data Provenance Initiative has demonstrated in its study of websites included in three of the most widely used training databases, the categories of data most critical to building high performing models—news outlets, social media, encyclopaedias and academic sources—are also those most likely to restrict access through robots.txt or terms of service (Data Provenance Initiative, 2024, pp. 4–6).<sup>xxi</sup> An opt out under copyright law would leave developers with two difficult options. Either they attempt to secure voluntary licences from an immense and fragmented number of rightsholders, which is practically unworkable, or they exclude precisely those sources that provide the greatest value, thereby reducing both the scale and the quality of AI models.

This issue is not a binary contest between creators and AI developers. It reflects the broader need for a balanced framework that protects the legitimate interests of rightsholders while also preserving the conditions necessary for innovation. Rather than suggesting that opt outs should be removed altogether, the policy challenge is to design them in a way that provides meaningful safeguards for rightsholders without making training unworkable.

Well-designed opt outs, combined with transparency, might provide that right balance where creators retain agency while AI systems continue to improve in ways that ultimately serve creators, innovators and society as a whole.

Therefore, India should adopt incentives and safeguards that preserve a reasonable balance between creating new works and enabling machine learning systems to train on existing ones.

Below we discuss two core safeguards, (1) lawful access requirement and (2) due diligence requirement during TDM activity, because they would ensure that only legitimately obtained content is mined and there is meaningful and appropriate responsibility attached to undertaking TDM. Other safeguards like measures to maintain the security and stability of computer systems and restrictions on redistributing the mined data should also be examined to further protect rightsholders and bolster public trust.

### Lawful Access

The lawful access requirement stipulates that a TDM user must access copyrighted work as per only through means that the copyright holder has authorised and as permitted under law.

For instance, EU Directive 2019/790 (DSM), Recital 14, describes lawful access as: "Access to content based on an open access policy or through contractual arrangements between rightsholders and research organisations...such as subscriptions, or through other lawful means. It further clarifies that lawful access should also cover access to content that is freely available online provided that the copy itself is non-infringing."

The example of Japan shows that countries can choose to forego the lawful access requirement under their TDM exception. Japan allows use of copyrighted works when the purpose is not to enjoy the expression in the work. This cover uses such as testing technology, data analysis, and computer processing that does not involve human perception of the expression, and it is subject to a guardrail that the use must not unreasonably prejudice the interests of the rightsholder.<sup>xxii</sup>

However, we believe that a lawful access requirement is crucial for balancing the seemingly contrasting goals of promoting AI innovation and respecting rightsholders' legitimate interests. This ensures that the TDM exception does not benefit users who are accessing the data through unauthorised means, like bypassing a paywall or hacking. To illustrate, if articles are put up on a website but only accessible by a subscription, users who accessed the system without the authorisation of the publisher (by bypassing the paywall) should not benefit from the TDM exception.

Conversely, once a user lawfully obtains access to the website (like, by payment of subscription fee or accessing an open-access copy), she should remain free to conduct TDM on website's content.<sup>xxiii</sup>

In the EU, copyright law creates two separate TDM rights:

- The first, in Article 3, **applies to TDM carried out by research organisations and cultural heritage institutions**, such as libraries, archives, and museums. These organisations can mine lawfully accessible works for the purposes of scientific research provided the access is obtained legitimately (whether freely available online, via subscription, under licence, or through other authorised means). This right cannot be taken away by a contract or website terms of service under Article 7. The law also allows rightsholders to apply proportionate technical measures to protect the security and integrity of their systems, but these cannot be used to refuse or opt out of the Article 3 research TDM exception.
- The second, referred to as the general exception, in Article 4, **applies to all other users, including those doing TDM for commercial purposes**. Here, right holders may opt out of the exception. If the content is publicly available online (freely accessible without paywalls, logins, or other access restrictions), the opt-out must be made in a machine-readable form, for example, using metadata tags or a "robots.txt" file, so that automated

systems can detect it. If the content is not freely accessible online, contractual restrictions or licence terms are sufficient to reserve rights.

In both cases, works must be accessed lawfully, meaning access must be authorised under copyright law and, where Article 4 applies or access is under licence, must comply with any enforceable contract terms. Any scraping that bypasses access controls or uses pirated material remains outside the scope of the exceptions.

In Singapore, section 244 enables copy, storage and limited communication (*copies made for computational data analysis may be shared to verify results or for collaborative research or study*) for computational data analysis (CDA) exception, Singapore uses CDA instead of TDM, for commercial and non-commercial purposes, subject to the required due diligence. Further, section 187 operates as an anti-override rule that makes it clear that website terms or licence clauses cannot remove permitted uses under Singapore's CDA exception, gives legal certainty so developers with lawful access may rely on the exception even if terms say "no text and data mining," and preserves the balance by keeping lawful access, anti-circumvention, reasonable diligence and limited sharing requirements in force.

### Due Diligence Requirement

Any TDM exception must ensure that developers mine only material to which they have lawful access. Comparative analysis of Hong Kong's consultation draft, Singapore's Copyright Act 2021 and the EU Digital Single Market Directive points to three regulatory approaches:

1. *Hong Kong (consultation draft)*: The TDM exception fails if (i) the copy was obtained without lawful access or (ii) the rightsholder has expressly reserved rights, including by a machine-readable opt-out notice for content that is publicly available online. Other proposed safeguards are that infringing copies cannot be used and the TDM exception does not apply where a relevant licensing scheme for the work is available, and users must keep records and disclose their sources on request. No express "knowledge" defence is provided, so liability may arise even where the developer was unaware of the infringement.
2. *EU (DSM Directive)*: TDM exceptions require lawful access. The general exception also permits an opt out by rightsholders, which for publicly available online content must be expressed in machine readable form at the point of availability. Due diligence requirements:
  - a. *Lawful access*: Reproductions or extractions are permitted only where the material is already lawfully accessible to the user.
  - b. *Retention*: Copies made for TDM may be kept for as long as necessary to conduct the mining (and by correlation – be deleted thereafter).
  - c. *Machine-readable reservation*: The general TDM exception is not available if the rightsholder has expressly reserved their rights "in an appropriate manner"; for content made publicly available online by machine readable means. Even if rights are reserved, including through machine readable reservations for publicly available online content, research organisations and cultural heritage institutions may mine lawfully accessible works for scientific research.
3. *Singapore (Copyright Act 2021)*: Singapore's TDM is permitted subject to these statutory conditions:
  - a. *Purpose & limited sharing*: Every copy is confined to TDM-related use and onward sharing is done only for verification or collaborative research.
  - b. *Lawful access, good faith knowledge safeguard*: The use of TDM exception is valid only where the user either has lawful access or "does not know" that the copy is infringing. If the copy is sourced from a "flagrantly infringing online location",<sup>xxiv</sup> the standard is higher: the user must also have had "no reasonable grounds to know" that the copy was infringing.
  - c. *Guidance on lawful access*: Material obtained by "circumventing paywalls or other access controls" fails the lawful-access test.
  - d. *Anti-Circumvention Safeguard*: Circumventing technological protection measures to obtain access is not permitted.



Hong Kong's proposed lawful access only approach strictly limits mining to authorised sources and permits rights reservations and licensing carve outs, which strongly protects rightsholders but can make web scale TDM impracticable and may be unreasonably strict in practice. The EU allows general TDM unless rights are reserved but requires machine readable reservations at the point of availability and ongoing maintenance across sites, directories and files, which can become compliance heavy and operationally burdensome for both publishers and miners.

Singapore pairs lawful access with a no knowledge and no reasonable grounds safeguard that calls for reasonable diligence, offering a more balanced path, though in the India context, it could be improved by overlaying it with an EU style opt out.<sup>xxv</sup>

Strengthening due diligence will increasingly be about standardising and scaling good industry practices which are suitable to relevant contexts, including public adoption of machine-readable notices, crawler tokens, published logs, provenance workflows, and independent audit of source lists, alongside legal safeguards and enablement.

\*\*\*

**For any queries related to this submission, kindly contact:**

Ashish Aggarwal (asaggarwal@nasscom.in), or Sudipto Banerjee (sudipto@nasscom.in) or Dhananjay Sharma (dhananjay@nasscom.in) with a copy to policy@nasscom.in.

**About Nasscom**

Nasscom is the premier trade body and chamber of commerce of the Tech industry in India and comprises over 3000 member companies including both Indian and multinational organisations that have a presence in India. Established in 1988, nasscom helps the technology products and services industry in India to be trustworthy and innovative across the globe. Our membership spans across the entire spectrum of the industry from start-ups to multinationals and from products to services, Global Service Centres to Engineering firms. Guided by India's vision to become a leading digital economy globally, nasscom focuses on accelerating the pace of transformation of the industry to emerge as the preferred enablers for global digital transformation. For more details, kindly visit [www.nasscom.in](http://www.nasscom.in).

## Annexure I

### References to industry practice on website access preferences and source due diligence for model training

#### Robots dot txt and crawler behaviour<sup>xxvi</sup>

- Google's documentation describes robots dot txt as a file used to guide crawler access and manage server load and explains that it is not a security control and cannot be enforced against all crawlers. This clarifies why many publishers pair robots.txt with contractual terms and access controls.

#### Published controls for named crawlers<sup>xxvii</sup>

- OpenAI publishes a page that lists its crawlers and provides robots dot txt examples to allow or block access. This offers a direct way for site operators to configure access for these crawlers.
- Anthropic provides equivalent instructions and states that its crawlers observe crawl delay and do not attempt to bypass barriers such as CAPTCHAs.
- Google documents a setting called Google Extended that lets publishers decide whether content on their sites may be used to improve certain generative features and explains how to reference this setting in robots dot txt alongside the general crawler documentation.

#### Practice at a major web archive<sup>xxviii</sup>

- Common Crawl states that it respects robots dot txt, supports crawl delay, does not bypass paywalls, and does not log in to private sites. This is relevant because many research datasets are built from its collections and because it sets out a clear approach to lawful access.

#### Notices beyond robots.txt<sup>xxix</sup>

- A World Wide Web Consortium community report sets out the Text and Data Mining Reservation Protocol. It describes a method for declaring a reservation for text and data mining and for linking to licensing policies, including through a well-known policy file. The report explains how this relates to European law on commercial text and data mining.
- Spawning documents a Do Not Train registry and a developer interface that dataset builders can consult to exclude listed works. This provides a central listing that some creators use in addition to website instructions.

#### Provenance and transparency<sup>xxx</sup>

- The C2PA specification for Content Credentials explains how verifiable provenance information about origin and edits can be attached to and checked for digital assets.
- In May 2024 C2PA announced that OpenAI had joined its steering committee. This indicates participation by model developers in the governance of provenance tools.

#### Licensed and permissioned sources for training<sup>xxxi</sup>

- Adobe's Firefly materials state that current models were trained on licensed content such as Adobe Stock and on public domain works and set out related contributor and enterprise information. These disclosures illustrate a traceable sourcing approach.
- Shutterstock describes data licensing and a contributor fund that compensates contributors when their works are used for training and has disclosed multiyear licensing arrangements with model developers.
- Direct licensing between developers and publishers is visible in public announcements, including the Financial Times and OpenAI partnership in April 2024 and the Associated Press collaboration in July 2023.

#### Independent reporting on uneven adherence<sup>xxxii</sup>

- Reuters reported in June 2024 that multiple companies were bypassing robots dot txt on publisher sites, citing analysis from a licensing analytics provider.
- WIRED reported that one market participant appeared to access material despite robots.txt exclusions and that the behaviour drew further review by infrastructure providers.
- A factsheet from the Reuters Institute for the Study of Journalism found that by late 2023 a substantial share of leading news sites across several countries had chosen to block AI specific crawlers.

#### Dataset review and remediation<sup>xxxiii</sup>

- The Stanford Internet Observatory reported in December 2023 that the LAION five B image link dataset included hundreds of known instances of child sexual abuse material.
- LAION later announced an updated release following a safety review that removed known links to suspected abusive material.

\*\*\*

**Annexure II****Mapping AI Training Workflows to Potential Copyright Infringements**

India's "fair dealing" exception under Section 52(1)(a)(i) of the Copyright Act is narrow, covering mainly uses that are "private or personal, including research". These limited carve-outs cannot accommodate the requirements of modern AI development<sup>xxxiv</sup>.

*The table below maps the essential steps that constitute training of models with how they may be construed as violation in the copyright law, as it stands. This is not a section-wise mapping but is based on widely held perceptions and discussions. This table helps us to segway into the specific concerns that may be held by copyright holders.*

Steps Involved	Copyright Considerations
<b>Creation of training datasets, i.e., collection of data from a variety of sources.</b>	If training dataset contains copyrighted works, it could be implied that copies are being made without permission – a possible violation.
<b>Model (pre-) training, i.e., process of transforming inputs like training data, model architecture, training algorithm, etc. into a base model.</b>	In addition to the concern highlighted above, storage can be construed as a possible violation. It is also not clear how much AI models store or "remember" or "memorise" their training data. This can depend on a model's design, size and how it is trained.
<b>Model fine tuning, i.e., modifying a base model for specific use-cases by training on additional data.</b>	Same as above, if copyrighted works are stored and copied during the fine-tuning stage, it can be alleged to be a violation due to "copying".
<b>Generation, i.e., output generation by model in response to input prompts.</b>	If models generate output material that replicates or closely resembles copyrighted works, such outputs may be liable for copyright infringement. This is perhaps the only step that is not a matter of debate or dispute. An output which violates the copyright law as it exists, is a violation.

\*\*\*

**Annexure III****Analysis of Four Concerns on AI Training and Copyright in India****(memorisation and reproduction; strict liability and lawful access; consent and compensation; market impact)****Concern 1: Models “memorise” data which amounts to the act of making copies and storage and thereby reproduction.****Whether models “memorise” data and whether training copies amount to reproduction**

Machine learning requires creating transient technical copies so that computers can read and process inputs. This differs from human learning, where no tangible temporary copy is made. The State of Israel, Ministry of Justice recognised this technical necessity in an official Opinion on the uses of copyrighted materials for machine learning issued 18 December 2022 (an MoJ Office of Legal Counsel paper intended to guide Israeli copyright analysis in the ML context):<sup>xxxv</sup>

*“Indeed, when a human learns, the dataset that enables learning is ‘located’ in her brain, without interfering with copyright law at all. A computer, however, cannot (at present) learn by ‘reading’ content, unless such content is copied first to a dataset that the computer can read. In other words, copyright arises in the context of ML only as a result of a technical-technological limitation of computer learning (that might change as technology advances.”*

These copies enable computation and are not made for distribution or communication to the public. The key point is that developers use them to extract meta-information from a work's expression rather than to consume or disseminate the expression itself.

Whether models “memorise” training data or primarily learn patterns remains debated; models are not intended to memorise, although some degree of memorisation may occur under certain conditions. Directionally, this supports treating training as a non-consumptive, intermediary process.

Current copyright rules were designed around distribution and public communication and map imperfectly to such uses. It is therefore more accurate to analyse transient, intermediary copies made solely to enable machine learning separately from traditional copying made for communication to the public. As a U.S. court recently put it, LLMs are “trained upon works not to race ahead and replicate or supplant them, but to turn a hard corner and create something different”.<sup>xxxvi</sup>

**Concern 2: Copyright infringement is based on strict liability and there is no place for intentions. So even if copying and adaptation are happening without an intention to reproduce or disseminate, it amounts to infringement.****Strict liability, lawful access, and practical enforceability**

Under existing copyright law, infringement can occur regardless of intention. Nothing in this analysis suggests diluting the lawful access baseline. Lawful access remains essential and should apply whether material is publicly available online, accessed under subscription, or obtained through a licence or other authorised means.

Modern training datasets are assembled from multiple channels. These include publicly available web pages that are open to general access, licensed or subscription sources where the provider grants permission under contract, and material that users or partners supply with authorisation. Collection typically happens through automated pipelines that fetch, parse, and transform content into processable formats. At this scale, item-by-item human review and provenance reconstruction for every datum are not feasible in practice, and online provenance signals are often incomplete or absent. The mere presence of content on a website does not by itself prove who owns the rights or whether the site had permission to host it.

The aim, therefore, is an effective lawful access rule that is observable up front and enforceable when breached. For publicly available online content, a clear machine readable opt out at the point of availability provides a bright line that automated systems can detect and respect. For content that is not publicly accessible, contracts and licences govern access and use. This structure creates value for rightsholders because it targets enforcement on



concrete, verifiable events. If access controls are bypassed, access is unlawful. If a valid opt out is present and a crawler proceeds anyway, any training that follows should not benefit from a training exception and ordinary remedies would follow. By contrast, imposing a duty to verify the provenance of every individual item on the open web would be costly to implement, hard to audit, and unlikely to yield proportional benefits for rightsholders.

A workable lawful access framework should be clear ex ante, technically operable at scale, and enforceable ex post. It should channel compliance toward respecting access controls, licences, and opt out signals, rather than toward retrospective item-level audits that are unlikely to improve outcomes. This approach keeps the lawful access requirement intact and makes it more effective in practice.

**Concern 3: AI models have used the copyrighted content without the consent of copyright owners and hence, they are entitled to remuneration/compensation for use of their data in the training stage.**

#### **Consent and compensation for training use**

Over centuries, the basic purpose of copyright law has been to protect original expression from being unfairly copied and shared with the public. Courts have also recognised that intermediate, non-expressive uses that enable search and analysis can be lawful in appropriate circumstances, for example the Second Circuit's decisions in *Authors Guild v. HathiTrust (2014)* and *Authors Guild v. Google (2015)*, which treated large scale copying to support search, accessibility and text analysis as fair use and noted the absence of market substitution when outputs do not communicate protected expression.<sup>xxxvii</sup> While these fair use decisions are in the United States context and arise under a broader and open-ended fair use standard, they are informative on the treatment of intermediate, non-expressive uses and may be read as persuasive.

On that footing, a blanket requirement to obtain prior consent or pay compensation for training on content that is publicly available online is unlikely to be justified generally, while specific legal obligations may still arise under regimes such as the European Union's opt out for publicly available online content.<sup>xxxviii</sup>

To illustrate, we have built three scenarios to analyse the concern of consent and compensation, while keeping the lawful access baseline intact.

- *First scenario* – right to opt-out is in force. Website can exercise opt-out rights through do not crawl restrictions for training models - either via robot.txt files, using any other current/future technology, or by stating this restriction in the website's terms and conditions. For content that is not publicly accessible, reservation can be set through contract or licence terms. In such cases, the AI crawlers must respect this right of reservation opt out and stop using that content for training.

Given this process is automated and happens at an enormous scale, seeking consent from a website that has imposed the restrictions is not practical at scale; therefore, the developer must stop collecting or using that content for training. Any further use should require specific consent, which may include compensation for training.

Therefore, if the crawlers fail to respect a valid opt out, any resulting use for training should not benefit from any training protection. In other words, it should be treated as a breach of a clear rule, and the resulting ordinary consequences and remedies should follow.

- *Second scenario* – one to one contract. AI developer may directly approach the website or the copyright holder and enter into a voluntary arrangement to get access to copyrighted data and agree to a deal at a contractually agreed price that addresses access, permitted uses, and any compensation.

- *Third scenario* - website might not express its right to opt-out. The crawlers can scrape the publicly available content for training, provided access is lawful and no access controls are bypassed. In such a situation, no additional statutory consent is required from the website owner and no statutory compensation would be payable, without prejudice to voluntary licensing.

As highlighted above, it is important to develop a scenario-based understanding, rather than making broad assumptions that consent, and compensation are always required for machine learning. This understanding is missing in the current law.

**Concern 4: Machine learning operates fundamentally differently from human cognition and at exponentially greater speed, make it inappropriate to extend fair-dealing exception to AI models. A further concern is that, once trained on copyrighted material, these models would enter the same or adjacent markets, directly competing with and commercially harming the original creators.**

#### **Human analogy, fair dealing, and potential market harm**

Just as humans learn from experience, observation, and repetition, AI models are trained on vast amounts of data to recognise patterns and generate responses. It could be argued that neither human innovation nor AI output is entirely original; both are re-combinations of existing ideas, knowledge, and influences.

Another similarity is the *attention layers* in machine learning, particularly in deep learning, which is inspired by the human brain's ability where the models focus on relevant information while filtering out distractions.<sup>xxxix</sup> This means not all parts of the input (or encoded input, extracted features, embedding, etc.) have the same importance in generating (decoding) expected output. This is like how human beings attribute their attention.

One can say that there is a clear difference - AI models can "learn" far faster and on a far larger scale than any human. However, this distinction does not change the fact that copyright liability is only tested on these questions: – was protected expression reproduced, adapted, or communicated without permission, and does fair dealing exception apply? Whether there is a single instance of infringement or infringement is at scale, the legal standard is identical. If regurgitating is not happening, the mere capability of machine to learn much faster than human minds does not disentitle it from the status of fair-dealing.

Coming to *second part* of the concern, Section 14 of the Copyright Act, 1957 grants right-holders exclusive economic rights but leaves it to the courts to decide, on a case-by-case basis, whether a new use, such as AI output, unlawfully undercuts those rights within the fair-dealing framework. Courts look for concrete evidence, i.e., declining sales, lost licenses, or direct substitution, rather than hypothetical harm. If an AI output only adds another creative option in the marketplace, it would not undercut the right holders. If the output is treated as a ready substitute i.e., demonstrate losses for a particular work, then it would. So, the kind of evidence copyright holders will bring in to illustrate the negative impact on their markets will play a prominent role in determining the outcome in many cases. However, because this assessment is grounded in factual, case-specific evidence, it cannot be extrapolated to justify broad policy assumptions about overall commercial harm.

To date, there is no clear evidence that generative AI has materially undercut creators' revenues at a market wide level. Japan's experience is often cited as context: a broad TDM exception enacted in 2018 and in force by 2019 coexists with a creative industry that ranks third globally after the United States and China, though this correlation does not establish causation.<sup>xl</sup> Japan's content market in 2021 was reported at ¥12.9 trillion (approximately £66 billion).<sup>xli</sup>

This issue is distinct from digital piracy, which concerns unauthorised reproduction and distribution to the public. In 2024, the Australian Government Productivity Commission observed that, unlike piracy, the training of AI models is not inherently a threat to existing revenue streams of artists and authors in many cases.<sup>xlii</sup> The evidence base is still developing and should be monitored.

\*\*\*

#### Annexure IV

#### Gaps in the Argument that Failure of Voluntary Market Justifies Statutory Licensing

- Statutory licensing can be justified in rare cases, such as excessive market power resulting in hold up problem like, **dominant copyright holders abuse their market power to extort exorbitant licensing fee for such content which is inevitable for AI training (essential input)**. At present, there is no such evidence of hold-up in the Indian market unless we are acting basis a future possibility. Moreover, in the context of AI training, where trillion of data points are used, it is unlikely that certain content creators can wield such influence. Further, we have not come across studies from any jurisdiction which suggests such a possibility.

Such narrow possibility of hold-up may arise only if we are concerned about specialised fine-tuned models that requires data from a particular copyright holder. In such situations, we cannot foreclose the possibility of voluntary licensing being the preferred solution. And in the event, there is an abuse of dominant position by copyright holders, adequate remedies are available under the Competition Act, 2002.

Here, we want to underscore that we have recommended (*refer to Recommendations on pg. 1*) a **carve out for recognised entities to carry out text-and-data mining on publicly available material for non-commercial purposes and they would not be subject to opt-out restrictions, as long as such content is lawfully accessed**. This calibrated carve-out forestalls artificial entry barriers, shielding use cases involving public goods and welfare from excessive fees and onerous bilateral licences, while preserving market incentives to negotiate bespoke deals when commercially worthwhile.

Therefore, creating a statutory licensing framework based on government prescribed licensing fees to address an *assumed hold-up problem* is fundamentally incorrect. On the contrary, such a framework can deprive the parties from the benefits of bilateral negotiations to unlock true value of content or tailor the scope of a licence to their specific circumstances.

Further, AI developers would have to pay **subscription fees** for content behind paywalls and respect other terms and conditions for carrying out TDM activity, irrespective of a voluntary or statutory framework.

- Another major argument in favour of statutory licensing is that **AI crawlers do not respect voluntary TPM like robot.txt** and therefore, no meaningful licensing can happen between small content creators and AI developers. This means a vast set of content creators would be deprived of their rightful compensation.

When AI crawlers ignore robots.txt or other consent signals, whether through negligence or bad faith scraping, the answer is meaningful compliance, not revoking creators' control. We want to emphasise that we are not in favour of blanket TDM exception, which means there would be sufficient **legal safeguards** to protect the rights of copyright holders.

We have recommended that India can adopt a tiered compliance framework that synthesises the strengths of the EU DSM Directive and Singapore's Copyright Act 2021 (*refer to Due Diligence Requirement on pg. 11*). In the event, any TDM activity carried out in the breach of **lawful access** requirements or **opt-out measure** (including machine readable format like Robot.txt or terms and conditions of the website), the safe harbour protection would cease to exist, and consequences of infringement shall follow.

Further, we believe the assumption that voluntary licensing cannot work also stems from trust deficit that opt-outs are never respected. To address this concern, in addition to necessary legal safeguards in the Copyright Act, industry should also develop **voluntary safeguards** which can clearly demonstrate to copyright holders how crawling activities are done, opt-outs are respected and the purpose whether such scraping of copyrighted data falls within the permitted scope. To illustrate, industry can demonstrate best practices like information may be provided like, a general description of the training data source, in a narrative form, information about large data sets used and information about the

crawlers or bots used to collect information. Many AI providers already participate in transparency measures in the form of model cards and transparency notes.<sup>xliii</sup>

Recently, *Cloudflare*, which handles about 20% of all internet traffic, introduced measures to block AI crawlers by default (**a kind of lawful access**). This is creating a pay per crawl marketplace that could allow individual content creator's/website owners charge per-page access fees.<sup>xliiv</sup> Solutions like these illustrates that we should not undermine the wisdom of the market to produce unique measures to address problems around poor enforceability of opt-out.

In addition to the above, the **government can play a constructive role in spreading awareness among rightsholders about technical means on how to enforce their opt-out rights.**

- Moreover, the **market is already moving towards voluntary agreements**, particularly when content has clear commercial value for training AI systems. For instance, AI companies and content providers have signed at least 83 commercial licensing deals, generating ~\$665 million in revenue as of January 2025.<sup>xliv</sup> Another interesting example is **TollBit** which is a content monetisation platform that allows AI bots and data scrapers to pay for content instead of scraping data from **websites**.

Voluntary licenses can also be used to specify how individual creators will receive compensation for AI training. For instance, **contributory fund by Shutterstock** is one such solution, where revenue from all licensing agreements with AI developers is pooled in for periodic distribution among individual creators.<sup>xlvi</sup> Another example comes from **Harper Collins**, where individual authors indicate their consent to Harper Collins for sharing their works with the AI developer. In return, authors receive the direct payment of a flat fee.<sup>xlvii</sup>

While the above-mentioned instances are mostly from the U.S., we cannot pre-empt that similar voluntary arrangements cannot come into existence in India. On the contrary, a statutory licensing framework would kill incentives for participants to devise suitable commercial arrangements.

\*\*\*



## Annexure V

### Challenges with Implementing Statutory Licensing

The recent report by the Office of the Chief Economist at the U.S. Copyright Office examined various policy options in the ongoing debate surrounding the use of copyrighted materials for AI model training. Notably, its observations regarding the proposal of a *blanket statutory licensing regime* are particularly instructive:<sup>xlviii</sup>

*“The reason that compulsory licensing is used sparingly is that such regimes can impose substantial economic inefficiencies, which have the ability to vastly exceed the inefficiencies they are intended to resolve. This is because compulsory licensing regimes require us to supplant the collective wisdom imbedded in market forces with the limited knowledge of some central planners. The central planner who must decide on a licensing rate, but to arrive at the socially optimal rate, the planner must have a comprehensive and explicit understanding of the value of rights to all stakeholders (knowledge that would otherwise be implicitly factored in through market forces). Since such knowledge cannot be possessed in the absence of a market, central planners are forced to set licensing rates with incomplete information. Rates that are too high or rates that are too low will result in an inefficient allocation of resources and deadweight loss. Compulsory licensing is only economically defensible if the inefficiencies that it resolves sufficiently exceed the distinct set of inefficiencies it imposes.”*

There is no evidence to suggest that statutory licensing can deliver fair compensation. Accurately assigning value to each text, image, or sound recording in multibillion-item training sets is infeasible, especially when the influence of any single work on model performance is negligible. To avoid these complications, one can argue that because machine learning does not discriminate between value of different data, a *flat fee structure* would serve as an equitable compensation framework.

Implementing statutory licensing would have the effect of pre-empting the collective wisdom of a voluntary licensing market where parties have the liberty to decide the value of copyrighted works used for training. **Rather, a flat fee structure would misprice content, while aggregate costs could rise to levels that make AI development economically untenable. Equally damaging will be the impact on content creators because mispriced content would disincentivise them to produce original content. Therefore, such arrangement will adversely impact both sides of the market.**

The report by the Office of the Chief Economist at the U.S. Copyright Office has highlighted another critical challenge with statutory licensing is that it takes away the flexibility, which is particularly valuable in licensing copyright protected content for training. In generative AI training, on the one hand, copyright holders are vastly different from each other, and, on the other hand, AI developers are likely to differ from each other too. The report states:

*“Three key contingencies are particularly significant: (1) heterogeneities in the quantity, diversity, and access regimes for training data; (2) heterogeneities in developer type (e.g., corporate versus startup or closed source versus open source); and (3) heterogeneities in user needs and content types. The remainder of this part discusses each of these contingencies in turn.”*

*First*, copyright holders differ from each other. *Second*, the AI products that are offered based on the trained generative AI models differ. *Third*, AI developers can differ in terms of the business models they pursue. Many of these differences have implications for the optimal remuneration of copyright protected content. In particular, it may affect how royalties are structured: lump sum, ad valorem, per unit, or even in the form of equity stakes.<sup>xlix</sup>

**However, statutory licensing framework risks side stepping all the above critical heterogeneities and may treat them at par, through a government set flat fee structure. Any decision to impose statutory framework should factor in the future implications like impact on incentives of stakeholders and associated costs.** For statutory licensing to be considered as a credible option, the case for it would need to be first evidenced through **robust evidence**.

## Endnotes

<sup>i</sup> What's documented in AI? Systematic Analysis of 32K AI Model Cards. (2024, February 7). arXiv preprint.

<https://arxiv.org/html/2402.05160v1>

<sup>ii</sup> As a recent United States judgment explains: "...LLMs have not reproduced to the public a given work's creative elements, nor even one author's identifiable expressive style...[it] has outputted grammar, composition, and style that the underlying LLM distilled from thousands of works. But if someone were to read all the modern-day classics because of their exceptional expression, memorise them, and then emulate a blend of their best writing, would that violate the Copyright Act? Of course, not."

Bartz, A., Graeber, C., & Johnson, K. W. (2025, June 23). Order on fair use, Bartz v. Anthropic, United States District Court, Northern District of California (p. 13) [Legal case]. <https://copyrightalliance.org/wp-content/uploads/2025/06/Bartz-v-Anthropic-Order.pdf>

<sup>iii</sup> Ethical Tech Initiative. (n.d.). DAIL – the Database of AI Litigation. <https://blogs.gwu.edu/law-eti/ai-litigation-database/>

<sup>iv</sup> Government of India. (1957). The Copyright Act, 1957 (Act 14 of 1957), Section 14. India Code (Legislative Department).

<sup>v</sup> Standing Committee on Finance (2025-26), Twenty-Fifth Report, Evolving Role of Competition Commission of India in the Economy, Particularly in the Digital Landscape, [https://sansad.in/getFile/lsscommittee/Finance/18\\_Finance\\_25.pdf?source=loksabhadocs](https://sansad.in/getFile/lsscommittee/Finance/18_Finance_25.pdf?source=loksabhadocs)

<sup>vi</sup> Council of the European Union. (2024). Policy questionnaire on the relationship between generative artificial intelligence and copyright and related rights – revised presidency summary of the Member States' contribution (p. 23).

<https://data.consilium.europa.eu/doc/document/ST-16710-2024-REV-1/en/pdf>

<sup>vii</sup> United States Copyright Office. (2024). Identifying the economic implications of artificial intelligence for copyright policy (p. 48).

<https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>

<sup>viii</sup> International Publishers Association. (2025, February). Madrid withdraws the Royal Decree on AI licenses.

<https://internationalpublishers.org/madrid-withdraws-the-royal-decree-on-ai-licenses/>

<sup>ix</sup> United States Copyright Office. (2024). Identifying the economic implications of artificial intelligence for copyright policy (p. 48).

<https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>

<sup>x</sup> Government of the Hong Kong Special Administrative Region, Intellectual Property Department. (2024). Public consultation on generative AI and copyright (p. 32 discusses Situations in Other Jurisdictions). <https://www.ipd.gov.hk/filemanager/ipd/en/share/consultation-papers/Eng-Copyright-and-AI-Consultation-Paper-20240708.pdf>

<sup>xi</sup> UK non-commercial research purpose and treatment of outputs. United Kingdom. (1988). Copyright, Designs and Patents Act 1988, c. 48, s. 29A. Legislation.gov.uk. <https://www.legislation.gov.uk/ukpga/1988/48/section/29A>; UK Intellectual Property Office. (2014). Exceptions to copyright: Research (pp. 6–11). GOV.UK.

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/375954/Research.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf)

<sup>xii</sup> UK lawful access definition. UK Intellectual Property Office. (n.d.). Exceptions to copyright. GOV.UK.

<https://www.gov.uk/guidance/exceptions-to-copyright>

<sup>xiii</sup> Japan and Singapore do not distinguish commercial and non-commercial TDM. Japan. (1970). Copyright Act (Act No. 48 of 1970), art. 30-4. Japan Law Translation. <https://www.japaneselawtranslation.go.jp/en/laws/view/3379/en>

<sup>xiv</sup> Japan proviso on unreasonable prejudice and official interpretative guidance. Japan. (1970). Copyright Act (Act No. 48 of 1970), art. 30-4 proviso. Japan Law Translation. <https://www.japaneselawtranslation.go.jp/en/laws/view/3379/en>; Agency for Cultural Affairs, Government of Japan. (2024). General understanding on AI and copyright in Japan (overview note).

[https://www.bunka.go.jp/english/policy/copyright/pdf/94055801\\_01.pdf](https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf)

<sup>xv</sup> EU Article 3 scope for research organisations and CHIs, with Recital 12. European Parliament and Council of the European Union. (2019). Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (OJ L 130, 17 May 2019, pp. 92–125). Articles 3 and Recital 12. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790>

<sup>xvi</sup> EU lawful access concept in Recital 14. European Parliament and Council of the European Union. (2019). Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (OJ L 130, 17 May 2019). Recital 14. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790>

<sup>xvii</sup> Singapore lawful access illustrations under s. 244. Singapore. Law Revision Commission. (2022). Copyright Act 2021 2020 Revised Edition, consolidated to 1 November 2022, s. 244, pp. 162–165, illustrations on p. 163. [https://wipolex-resources-eu-central-1-358922420655.s3.amazonaws.com/edocs/lexdocs/laws/en/sg/sg179en\\_1.pdf](https://wipolex-resources-eu-central-1-358922420655.s3.amazonaws.com/edocs/lexdocs/laws/en/sg/sg179en_1.pdf)

<sup>xviii</sup> UK technological protection measures and contract terms in guidance. UK Intellectual Property Office. (2014). Exceptions to copyright: Research (pp. 6–9). GOV.UK.

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/375954/Research.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf)

<sup>xix</sup> EU Article 3 paragraph 3 on protective measures that must not undermine the exception. European Parliament and Council of the European Union. (2019). Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (OJ L 130, 17 May 2019). Article 3(3). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790>

<sup>xx</sup> The Data That Powers A.I. Is Disappearing Fast, The New York Times, <https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html>

<sup>xxi</sup> Data Provenance Initiative. (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons (pp. 4–6). Data Provenance Initiative. [https://www.dataprovenance.org/Consent\\_in\\_Crisis.pdf](https://www.dataprovenance.org/Consent_in_Crisis.pdf)

<sup>xxii</sup> Government of Japan. (2018). Copyright Act (Act No. 48 of 1970, as amended), Article 30-4: Exploitation without the purpose of enjoying the thoughts or sentiments expressed in a work. Japanese Law Translation Database, <https://www.japaneselawtranslation.go.jp/en/laws/view/3379>

<sup>xxiii</sup> Publications Office of the European Union, Study on the Legal Framework of Text and Data Mining (TDM),

<https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en>

<sup>xxiv</sup> Republic of Singapore. (2021). Copyright Act 2021 (No. 22 of 2021), s. 99 "What is a flagrantly infringing online location." Singapore Statutes Online, <https://sso.agc.gov.sg/Act/CA2021>

<sup>xxv</sup> Section 244(e) of the Singapore Copyright Act, 2021, <https://sso.agc.gov.sg/Act/CA2021?Provides=P15-#pr244->

<sup>xxvi</sup> Robots.txt and crawler behaviour. Google. (2025, February 4). Introduction to robots.txt. Google Search Central.

<https://developers.google.com/search/docs/crawling-indexing/robots/intro>



<sup>xxxii</sup> Published controls for named crawlers:

- OpenAI. (n.d.). Overview of OpenAI crawlers. <https://platform.openai.com/docs/bots>
- Anthropic. (n.d.). Does Anthropic crawl data from the web, and how can site owners block the crawler? Anthropic Help Center. <https://support.anthropic.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler>
- Google. (2023, September 28). An update on web publisher controls. The Keyword. <https://blog.google/technology/ai/an-update-on-web-publisher-controls/>
- Google. (2025, June 19). AI features and your website. Google Search Central. <https://developers.google.com/search/docs/appearance/ai-features>
- Google. (2025, March 6). List of Google's common crawlers. Google Search Central. <https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers>

<sup>xxxiii</sup> Practice at a major web archive:

- Common Crawl Foundation. (n.d.). FAQ. <https://commoncrawl.org/faq>
- Common Crawl Foundation. (n.d.). About CCBot. <https://commoncrawl.org/connectivity>

<sup>xxxix</sup> Notices beyond robots.txt:

- W3C TDM Reservation Protocol Community Group. (2024). Text and Data Mining Reservation Protocol. W3C Community Group Final Report. <https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep/>
- Spawning. (n.d.). Spawning API documentation. <https://spawning.ai/api>

<sup>xxx</sup> Provenance and transparency:

- Coalition for Content Provenance and Authenticity. (2024). C2PA explainer. <https://c2pa.org/specifications/explainer/>
- Coalition for Content Provenance and Authenticity. (2024, May 7). OpenAI joins C2PA steering committee. <https://c2pa.org/openai-joins-c2pa-steering-committee/>

<sup>xxxii</sup> Licensed and permissioned sources for training:

- Adobe. (n.d.). Adobe Firefly FAQ: What is Firefly trained on? <https://www.adobe.com/products/firefly/faq.html>
- Shutterstock, Inc. (n.d.). Data licensing and the contributor fund. Shutterstock Help Center. <https://support.submit.shutterstock.com/s/article/Data-Licensing-and-the-Contributor-Fund>
- Shutterstock, Inc. (2023, July 11). Shutterstock expands partnership with OpenAI, signs new six-year agreement to provide high quality training data. Investor Relations. <https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year>

## Direct licensing between developers and publishers:

- Financial Times. (2024, April 29). Financial Times announces strategic partnership with OpenAI. FT Press Office. [https://aboutus.ft.com/press\\_release/openai](https://aboutus.ft.com/press_release/openai)
- The Associated Press. (2023, July 13). AP and OpenAI agree to share select news content and technology in new collaboration. AP Press Releases. <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/>

<sup>xxxiii</sup> Independent reporting on uneven adherence:

- Schoenick, C., & Dastin, J. (2024, June 18). Exclusive: Some AI bots evade web rules to scrape websites. Reuters. <https://www.reuters.com/technology/ai-some-ai-bots-evade-web-rules-scrape-websites-2024-06-18/>
- Keck, C., & Cox, J. (2024, June 20). Perplexity is ignoring robots.txt and scraping websites without permission. WIRED. <https://www.wired.com/story/perplexity-ai-scraping-publishers-robots-txt/>
- Fletcher, R. (2024, February 22). How many news websites block AI crawlers? Reuters Institute for the Study of Journalism. <https://reutersinstitute.politics.ox.ac.uk/how-many-news-websites-block-ai-crawlers>

<sup>xxxiii</sup> Dataset review and remediation:

- Thiel, D., Brown, I., Carbajal, J., & Hurd, R. (2023, December 23). Identifying and eliminating CSAM in generative ML training data and models [Report]. Stanford Internet Observatory. <https://purl.stanford.edu/kh752sm9123>
- LAION e.V. (2024, February 29). Re-LAION-5B: Update on safety review and removal workflow [Blog post]. <https://laion.ai/blog/relaion-5b/>

<sup>xxxiv</sup> Report on AI Governance Guidelines Development <https://indiaai.s3.ap-south-1.amazonaws.com/docs/subcommittee-report-dec26.pdf>, pg. 10.<sup>xxxv</sup> State of Israel, Ministry of Justice, Opinion: uses of Copyrighted Materials for Learning (2022),<https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf><sup>xxxvi</sup> Bartz v. Anthropic PBC, N.D. Cal., Order on Fair Use, June 23, 2025, pp. 12–13<sup>xxxvii</sup> U.S. Copyright Office. (n.d.). Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) [Fair Use Index summary]. U.S. Copyright Office, <https://www.copyright.gov/fair-use/summaries/authorsguild-hathitrust-2dcir2014.pdf><sup>xxxviii</sup> European Parliament and Council of the European Union. (2019). Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Official Journal of the European Union, L 130, 92–125. EUR-Lex. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790><sup>xxxix</sup> Attention in the Human Brain and Its Applications in ML, <https://thegradient.pub/attention-in-human-brain-and-its-applications-in-ml/><sup>xl</sup> Government of Japan. (2018). Copyright Act [Act No. 48 of 1970, as amended by Act No. 30 of 2018], art. 30-4. Japanese Law Translation Database. <https://www.japaneselawtranslation.go.jp/en/laws/view/3379><sup>xli</sup> Cabinet Secretariat, Headquarters for the Realization of New Capitalism. (2024, April 17). Basic materials — “Size of the global content market.” [https://www.cas.go.jp/jp/seisaku/atarashii\\_sihonsyugi/kaigi/dai26/shiryoku1.pdf](https://www.cas.go.jp/jp/seisaku/atarashii_sihonsyugi/kaigi/dai26/shiryoku1.pdf)<sup>xlii</sup> Productivity Commission. (2024, January 11). Making the most of the AI opportunity: Research paper 3 Australian Government.<https://www.pc.gov.au/research/completed/making-the-most-of-the-ai-opportunity/ai-paper3-data.pdf><sup>xliii</sup> What's documented in AI? Systematic Analysis of 32K AI Model Cards, 07 Feb 2024, <https://arxiv.org/html/2402.05160v1><sup>xliv</sup> Radsch, C. (2025, July 11). Cloudflare Wades into the Battle Over AI Consent and Compensation. Tech Policy Press.<https://www.techpolicy.press/cloudflare-wades-into-the-battle-over-ai-consent-and-compensation/><sup>xlv</sup> CREATE Centre, Copyright and AI: Response by the CREATE Centre to the UK Government's Consultation (2025), <https://zenodo.org/records/14931964>



---

<sup>xvi</sup> Shutterstock Data Licensing and the Contributor Fund, [https://support.submit.shutterstock.com/s/article/Shutterstock-Data-Licensing-and-the-Contributor-Fund?language=en\\_US](https://support.submit.shutterstock.com/s/article/Shutterstock-Data-Licensing-and-the-Contributor-Fund?language=en_US)

<sup>xvii</sup> The Authors Guild, HarperCollins AI Licensing Deal, <https://authorsguild.org/news/harpercollins-ai-licensing-deal/>

<sup>xviii</sup> Identifying the Economic Implications of Artificial Intelligence for Copyright Policy, Context and Direction for Economic Research, Edited by Brent Lutes, *Chief Economist, United States Copyright Office*, Chapter 7, Controlling the Use of Copyrighted Materials in Training), pg. 49, February 2025, <https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>

<sup>xix</sup> Jorge Padilla and Kadambari Prasad, Generative AI Models at the Gate: Licensing Frameworks for the Effective and Efficient Protection of Copyright Protected Content in an AI World (2025), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=5263547](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5263547)

\*\*\*