The Alan Turing Institute

Creative Grey Zones: Copyright in the Age of Hybridity

About The Alan Turing Institute

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence. The Institute is named in honour of Alan Turing, whose pioneering work in theoretical and applied mathematics, engineering, and computing is considered to have laid the foundations for modern-day data science and artificial intelligence. The Institute's goals are to undertake world-class research in data science and artificial intelligence, apply its research to real-world problems driving economic impact and societal good, lead the training of a new generation of scientists, and shape the public conversation around data and algorithms.

Authors

Uzma Chaudhry, Ann Borda, Stephanie Forbes, Joe Jones, Antonella Perini, Mattias Rättzén, Mary Stewart-David

Acknowledgements

We extend our gratitude to Mattias Rättzén (IP Attorney and Founder, Midvinter) for his authorial contribution to *Section 2* of this report, and to Joe Jones (Director Research & Insights, IAPP) and Stephanie Forbes (Summer Privacy Fellow, IAPP) for their valuable authorial contributions to *Section 3*. We are also grateful to Mary Stewart-David of York University for contributing use case (1) in *Section 6*, and to Chris Morrison of the Bodleian Libraries at University of Oxford and Rowan Wilson of University of Oxford for contributing use cases (2) and (3) also featured in *Section 6* of this report.

We would also like to thank Don McCombie (Partner, Clifford Chance) and John Collomosse (Sr. Principal Scientist, Adobe Research) for their expert insights and members of the AI Governance & Regulatory Innovation team at the Alan Turing Institute, namely Dr Florian Ostmann (Director), Christopher Thomas (Senior Research Associate) and Dr Christopher Nathan (Policy Fellow), for their expert insights, peer review, and timely feedback, which greatly refined this document.

We are also deeply thankful to Emma Rowlands, who led the design of this report.

Citation

Cite this work as:
Chaudhry, U., Borda, A., Forbes, S., Jones, J., Perini, A., Rättzén, M., Stewart-David, M. (2025). Creative Grey Zones: Copyright in the Age of Hybridity. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.17750379

Table of contents

How to read this report

This report aims to be accessible and informative and has therefore been written for a broad audience, from those encountering copyright & AI for the first time to experienced professionals in law, policy, technology, creative industries, and technical standards.

Based on your background, you may find certain sections of the report particularly relevant. We recommend the following reading strategies:

For **all readers**: *Executive Summary*; *Section 1*, *Introduction*; and *Section 7, Conclusion.*

If you are a **newcomer**: *Section 2, Rationale for copyright law and the impact of AI*; *Section 3, Copyright, Designs, and Patents Act 1988*; and *Section 5, The UK's Copyright and AI: Consultation 2024-2025.*

If you are a **policymaker or regulator**: *Section 2, Rationale for copyright law and the impact of AI*; *Section 3, Copyright, Designs, and Patents Act 1988*; *Section 4, Standardising copyright* alongside *Annexure I: Emerging technical standards*; *Section 5, The UK's Copyright and AI: Consultation 2024-2025.*

If you are a **lawyer or legal scholar**: *Section 3, Copyright, Designs, and Patents Act 1988*; *Section 4, Standardising copyright* alongside *Annexure I: Emerging technical standards*; *Section 5, The UK's Copyright and AI: Consultation 2024-2025*; and *Section 6, Hybrid use cases.*

If you are **involved with or interested in technical standards**: *Section 3, Copyright, Designs, and Patents Act 1988*; *Section 4, Standardising copyright* alongside *Annexure I: Emerging technical standards*; *Section 5, The UK's Copyright and AI: Consultation 2024-2025.*

If you are a **technologist or creative practitioner**, *Section 3, Copyright, Designs, and Patents Act 1988; Section 4, Standardising copyright* alongside *Annexure I: Emerging technical standards*; *Section 5, The UK's Copyright and AI: Consultation 2024-2025; and Section 6, Hybrid use cases.*

If you are **part of the wider responsible AI community**, *Section 2, Rationale for copyright law and the impact of AI; Section 4, Standardising copyright*; *Section 5, The UK's Copyright and AI: Consultation 2024-2025*; and *Section 6, Hybrid use cases.*

List of acronyms

ACCCT	Access, Control, Consent, Compensation and Transparency
CC	Creative Commons
CDPA	Copyright, Designs and Patents Act
CJEU	Court of Justice of the European Union
CLA	Copyright Licensing Agency
C2PA	Coalition for Content Provenance and Authenticity
DECaDE	Centre for the Decentralized Digital Economy
EU	European Union
GPAI	General-purpose artificial intelligence
IAB	Internet Architecture Board
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
ISCC	International Standard Content Code
ISO	International Organization for Standardization
LDMA	Literary, dramatic, musical or artistic
LLM	Large language models
REP	Robots Exclusion Protocol
SAG-AFTRA	Screen Actors Guild-American Federation of Television and Radio Artists
SDOs	Standards Development Organisations
TDM	Text and data mining
UK	United Kingdom
US	United States

Glossary

Terminology	Meaning assigned in the report
Artificial Intelligence (AI)	AI is referred to in the broadest sense and includes techniques such as machine learning and deep learning. It also includes generative AI models and systems, and where specified, agents.
AI training	A process by which AI models learn to recognise statistical patterns in data collected through techniques such as text and data mining. In this process, models analyse large collections of data to identify patterns, relationships, and structures therein.
Copyright	***This report assigns the same meaning as has been established under Section 1 of the Copyright, Designs and Patents Act 1988.** A property right that subsists in the following descriptions of work: (a) Original literary, dramatic, musical or artistic works, (b) Sound recordings, film [or broadcasts], and (c) The typographical arrangement of published works
Generative AI	A subfield of deep learning that analyses and learns underlying patterns and structures to generate text, images, audio, video or other synthetic media in response to user prompts.
Hybrid stakeholders	An individual or entity operating across multiple domains (such as technology, creativity, art, design, research, or production), who contributes to the development, support, or delivery of hybrid creative outcomes. Hybridity exists on a spectrum and can cover creative practitioners and technologists who blend digital and / or AI tools to produce new forms of creative and artistic works, as well as involve facilitators, data stewards, or service providers, for example. Hybrid stakeholders reflect the fluid and interdisciplinary nature of contemporary creative and technological ecosystems in a technology driven world.

Large Language Model (LLM)	Models capable of generating responses to natural language prompts. LLMs are pretrained on large text datasets and rely on deep learning and machine learning techniques to analyse and learn text patterns and relationships between words to generate dialogue like responses to user prompts.
Licence	Permission obtained from a rightsholder to do something that would otherwise be an infringement of copyright.
Machine-readable rights reservation	Refers to the use of standardised, structured data formats (such as metadata tags) to express and communicate the rights status or usage restrictions of a work in a way that can be automatically recognized and processed by web crawlers or bots.
Rightsholders	Original creators that lawfully own their copyright currently. Licensees (holders of a copyright licence) are not captured under this definition.
Technical standards	Formal documents that provide uniform rules, guidelines, or specifications for repeatable tasks. They are developed through expert consensus. Technical standards include de jure standards, i.e. standards developed by formal Standards Development Organisations (SDOs). They also include de facto standards, i.e. they have fallen into customary and widespread use even though they are not officially recognised as standards by formal SDOs.
Text and Data Mining (TDM)	***The report assigns the same meaning as has been established under Article 2(2) of Directive (EU) 2019/790 on Copyright in the Digital Single Market.** 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.
Technologist	A professional who focuses on the practical application of technologies to solve problems, design, build, and maintain AI systems and software applications.

Executive summary

The rise of artificial intelligence (AI), especially since the generative AI boom, has catalysed a global re-examination of copyright law. As AI models increasingly ingest content and generate outputs based on the ingested content, the efficacy of legal frameworks that have traditionally governed authorship is being tested.

This tension has stemmed from the large-scale use of copyrighted material, through techniques like text and data mining, for training AI systems without explicit authorisation from the rightsholders. This has raised questions about whether such use, without express authorisation, constitutes copyright infringement. This growing friction is reflected in a wave of ongoing litigation across multiple jurisdictions, as well as in proactive policy reforms, such as in the United Kingdom (UK) where the government initiated a Copyright and AI: Consultation late 2024.

This report has been written against the backdrop of the UK's recent consultation, a process that exemplifies the re-evaluation of the UK's copyright regime in response to copyright and AI. Framing the discussion within the live policy process ensures that findings are timely and relevant for the ongoing decision-making.

The report explores four critical areas: ethics, law, technical standards, and hybrid stakeholders. Through in-depth analysis, and focused case studies, we find that the rapid evolution of AI models is pushing each of these domains into grey zones and applying the existing regime to balance the moral and legal interests of both technologists and rightsholders has become increasingly challenging. This difficulty is compounded in the modern day where hybrid stakeholders are blending digital and/or AI tools to produce new forms of creative and artistic works. In various parts of the report, impacts on the hybrid stakeholders are analysed specifically, and impacts on technologists and creatives are analysed more generally.

However, at the outset, it is important to acknowledge the challenge of providing a fully balanced view across all four categories. This is because certain issues and risks affect some stakeholders more than others. Another reason is the transitional state of the current copyright and standards landscape, where established frameworks are

being questioned for their adequacy, and there is disagreement over what reforms ought to look like. Moreover, interdependence of the four grey zones means that reform in one area cannot succeed without progress in others. As such, while legal and policy reforms may need to address each of these areas individually, maintaining alignment across them may also require attention to their interconnections to support greater coherence and reduce uncertainty within the broader framework.

Research findings

This project has been initiated to better understand the current copyright landscape, and the impacts of the government's proposal to introduce a text and data mining exception underpinned by a rights-reservation model (i.e., a text and data mining exception be allowed unless an opt-out has been declared by a rightsholder) and transparency mechanisms. To do so, we brought together a multidisciplinary group comprising AI ethicists, legal and policy experts to develop this report. We also developed case studies to explore the interests of stakeholders that do not neatly fit within the "tech vs. creatives" binary, i.e. the hybrid stakeholders, who are blending artistic and computational methods for their creative works. Our key findings across the four identified grey zones include:

Ethics

- **Copyright law is grounded in strong moral and jurisprudential rationales. It not only safeguards the market value but also affirms the dignity of creative work.**

  Copyright recognises the inherent rights of creators to control and be credited for their creative works, grounded in valuing personal autonomy, self-expression, and the moral integrity of creative labour. Copyright protection is not only about the economic incentive, i.e., remunerating individual creators, but also about serving a broader public interest of encouraging long-term investment in creativity and fostering cultural heritage. Overprotection or underprotection could potentially disrupt the moral, jurisprudential, and economic rationales that have long served to delicately balance private and public interests.

- **The comparison between AI training and human learning is contested, particularly in relation to the scale, purpose, and impact of generative AI systems**

  The debate centres on the ingestion and use of copyrighted data during the training phase being analogous to human learning and inspiration-seeking on the one hand. However, on the other hand, critics argue that AI training is different from human cognition as AI systems not only learn but replicate works at an industrial scale and for commercial purposes. The challenge is compounded where generative AI outputs compete with the original creators.

- **Model collapse and cultural entropy represent emerging concerns where diminished human input and overreliance on synthetic content may lead to a decline in originality**

  A shift in the balance between human and AI-generated content could alter the dynamics of creative production. As a result, models increasingly trained on synthetic rather than original human-made works, could increase the risk that future outputs may become homogenised, less diverse, and less innovative. This feedback loop may erode the cultural richness that human creators contribute.

Law

- **Whether, or the extent to which, works are "memorised" by AI models remains an unresolved question both legally and technically**

  It is still unclear if, how, or the extent to which, AI models retain copyrighted material from the training dataset. Even if it can be said that these systems do not store text, audio, image, or videos files in the same way as traditional computer files, they can sometimes reproduce data verbatim from the training dataset. This also varies significantly depending on the type of model used. This raises questions about the distinction between statistical pattern learning and reproduction of protected works.

- **Geographical location of AI training can significantly affect legal liability**
  Although international copyright is harmonised to an extent through the Berne Convention, jurisdictions retain discretion in how they apply exceptions. As a result, AI training activities could be lawful in one jurisdiction but infringing in another, making the jurisdiction of AI training a critical factor in determining copyright infringement liability.

- **The nature of AI training sits uneasily with the procedural requirements of copyright law**
  Building generative AI models involves harvesting data from diverse online sources through automated scraping methods and / or by using pre-scraped data collected by others. This large-scale and opaque process creates practical challenges in determining whether copyright works are included in a training dataset and for tracing and contacting individual rightsholders for obtaining permissions retrospectively. As a result, the technical realities of AI training do not neatly align with procedural requirements established by copyright law.

Technical standards

- **Emerging technical standards are both promising and burdensome**
  Machine-readable protocols, particularly unit-based mechanisms, offer promising tools to rightsholders to enable clearer copyright protection, and for technologists to avoid unintended infringement. They offer the granularity that site-based identifiers such as robots.txt lack. However, at the same time, if a text and data mining exception based on declaring opt-outs is introduced, it would shift the responsibility on rightsholders to take additional steps to manage their rights. At present, they benefit from automatic protection of copyright under the law with no additional formalities required to safeguard rights.

- **Digital content is rarely static. It is constantly resized, reformatted, remixed, and reshared across platforms.**

  This dynamic nature makes it technically complex, if not impossible, to track and manage rights across every instance and use case. While emerging technical standards aim to encode metadata, and track provenance, these systems often struggle to keep pace with the fluidity of digital circulation. This can be useful for technologists in avoiding unintended infringement but simultaneously create challenges for rightsholders. Effective rights management in this context requires not only interoperable standards but also presupposes a high level of technical literacy on the part of rightsholders.

- **Technical standardisation offers a potential route to certainty for evolving legal and industry needs**

  Although standards are not substitutes for legal certainty, they can perform a quasi-legal role, especially where the law lags, and community driven, interoperable standards can fill the gaps effectively. Standards create best practices which can offer certainty in the absence of binding rules. However, to avoid fragmentation, it is important for standards to have legitimacy through industry consensus and regulatory uptake. Success will also depend on stakeholder inclusion. As much of the burden of rights management under a potential TDM exception for commercial uses would befall the rightsholders, their participation in standards development processes can ensure their interests are entrenched and catered for in the best practices that may eventually emerge.

Hybrid stakeholders

- **The rise of hybrid stakeholder roles is reshaping traditional understanding of creativity and ownership in the digital era**

  In today's digital environment, the longstanding binary between "creatives" and "technologists" while relevant, can also be limiting, as these roles are increasingly and deeply entangled. Hybrid stakeholders work across disciplinary boundaries; curating datasets, fine-tuning models, and embedding

creative direction into technical systems. These blurred roles challenge legal frameworks that depend on fixed definitions.

- **There is a growing recognition among hybrid stakeholders of the need to adapt copyright frameworks to the evolving digital landscape**
  The case studies in this report illustrate ongoing efforts in the UK and internationally to modernise copyright law, from proposals to refine fair dealing and introduce hybrid models that balance AI innovation with creators' rights, to expanding legal provisions that better support libraries, archives, and cultural heritage organisations in digital access, preservation and archiving. Collectively, these developments point toward an emerging consensus on updating copyright frameworks through new exceptions, rules or standards that are better suited to the realities of the evolving digital age.

- **There is stakeholder support to operationalise copyright frameworks through practical, collaborative, and standards-based approaches**
  The case studies highlight support for initiatives aimed at translating legal principles into workable systems through industry standards and collective licencing. Emerging practices such as C2PA are promoting greater transparency, responsibility and interoperability in AI contexts. At the same time, collective licencing schemes led by organisations such as Copyright Licensing Agency, Publishers' Licensing Services, Authors' Licensing and Collecting Society, the Copyright Clearance Center, etc., are exploring ways to create harmonised frameworks and codes of practice.

Section 1. Introduction

The rapid advancement of artificial intelligence has prompted a re-evaluation of how copyright law ought to operate in the digital age. As AI tools become more integrated, they raise complex questions about various aspects of AI, ranging from the use of copyrighted data to form part of training datasets to the jurisdiction where they were trained to the ownership of outputs. Copyright challenges have gained centre stage in recent litigation and are also being examined by governments, whether through consultations[1] or legislative action, such as the General-Purpose AI Code of Practice[2] under the European Union (EU) AI Act.[3]

In the UK, the issue was taken up by the government through the Copyright and AI: Consultation in December 2024 (the consultation) which focused on the disputed nature of the application of current UK law to AI training. On the one hand, the consultation acknowledged that rightsholders are finding it difficult to control the use of their works in AI training and get compensated for such use. On the other hand, it highlighted that legal uncertainty could undermine investment in and adoption of AI in the UK. As such, the consultation centred on how to balance the development of world-leading AI models in the UK with rightsholders interests in a manner that also promotes transparency. The consultation proposed to replace the UK's current text and data mining (TDM) exception, which is permitted for non-commercial purposes, with a TDM exception for commercial purposes supported by a rights reservation model underpinned by transparency mechanisms. This proposed exception overlaps

[1] Intellectual Property Office, Department for Science, Innovation & Technology, and Department for Culture, Media & Sport (2024, December 17*). Copyright and AI: Consultation.* GOV.UK Available at https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence

[2] European Commission (n.d.). *The General-Purpose AI Code of Practice*. Available at https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[3] Regulation (EU) 2024/1689 (Artificial Intelligence Act), *Official Journal of the European Union,* L 2024/1689 (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

with the exception provided under Article 4 of the EU Directive on Copyright in the Digital Single Market.[4]

Moreover, the consultation was framed in terms of a binary conflict between the technology sector and the creative industry. However, this framing, while important, does not fully capture the complexities of the evolving digital landscape. Between these two groups exists a hybrid category of stakeholders who do not operate solely as either creatives or technologists. Instead, they occupy a fluid position where creative expression is enhanced, augmented or transformed through AI systems. Their work exemplifies the blurred boundaries between human authorship and machine assistance.

Against this backdrop, this report explores the UK's current copyright framework in the context of tensions caused by generative AI. It does so by first exploring the traditional and current rationales behind the existence of copyright law and then mapping the existing landscape of law and technical standards. The first three sections thus provide the foundation for a structured analysis of the consultation. This is followed by case studies that explore and examine various contexts of hybridity. The purpose of the case studies is to illustrate that the range of impacted stakeholders extends beyond the traditional binary of "technology" and "creative" sectors, highlighting the emergence of a third category of stakeholders who operate within a legal and technical grey zone. Rather than analysing specific legal or practical impacts, the case studies aim to broaden understanding of the evolving role of these actors within the digital ecosystem and to encourage dialogue and policy reforms that recognise the diversity of stakeholder experiences within the UK's copyright framework.

Throughout the report, implications for technologists, creatives, and hybrid stakeholders are analysed, particularly through the dimensions of ethics, law, and technical standards–each of which is currently operating within its own grey zone.

---

[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 and amending Directives 96/9/EC and 2001/29/EC. *Official Journal of the European Union, L130, 92-125.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790

Scope and methodology

The focus of the report is the ingestion of copyrighted content for training AI systems and not ownership of AI generated outputs. Other types of intellectual property (IP), such as patents, trade marks, and designs, are also not within the scope of this report.

The report offers insights on:

- Ethical and jurisprudential rationales that have shaped copyright law
- The existing copyright regime in the UK and exceptions thereunder
- The current landscape of technical standards
- Analysis of the proposed TDM exception for commercial purposes with rights reservation supported by transparency mechanisms
- Case studies featuring hybrid use cases

Each section covering the aforementioned has benefited from expert input and by drawing on interdisciplinary research and analysis of current norms and practice. The report also includes case studies carried out through primary and desk research to spotlight some of the hybrid use cases and demonstrate that hybrid contexts or use-cases are not a one-size-fits-all.

While the report focuses on practical impacts for hybrid stakeholders wherever relevant, it also simultaneously offers a general analysis of the current landscape. This is to:

(a) make the report accessible to a wide audience;
(b) situate the hybrid stakeholder category within the wider legal and technical contexts; and
(c) illustrate the challenge of balancing the impacts of any solution between the diverse and sometimes competing interests of different stakeholder groups

Observations across the four grey zones have been compiled in a conclusion towards the end of the report.

Objective

This report is exploratory in nature. The objective is to inform, rather than prescribe, by bringing together multistakeholder insights across the four grey zones analysed in this report. This has been done by mapping the current landscape, identifying emerging tensions and grey areas, and considering the potential benefits and drawbacks of proposed approaches. A key objective has been to lay the groundwork for informed, cross-sectoral dialogue for the future.

Audience

The report has been written with a broad audience in mind, including those new to the field, with the goal of making the discussion accessible, informative, and grounded in multidisciplinary perspectives. However, it may especially be helpful for the following: policymakers, lawyers, stakeholders interested in or involved with technical standards, technologists, creatives, hybrid stakeholders and the wider responsible AI community.

Section 2. Rationale for copyright law and the impact of AI

Traditional and current rationales for copyright law

Copyright has long been justified as a means of fostering the creation and dissemination of original works. In doing so, it aims to strike a balance between the rights of authors and the interests of the public, particularly by ensuring that the public ultimately benefits from access to those works. Underlying this balancing act are philosophical, economic, and policy considerations that have evolved over centuries.

One of the earliest examples of copyright enforcement dates back to the sixth century and involved a dispute between two missionaries, St. Columba and St. Finian, over a copied book manuscript. The dispute was eventually resolved by King Diarmait of Ireland, who famously ruled: "To every cow belongs her calf; therefore, to every book belongs its copy".[5] As such, the theoretical foundations of copyright can be traced to centuries-old theories individual rights and property. John Locke's property theory supposes that individuals have a natural entitlement to the fruits of their labour,[6] which, when applied to creative works as "property", suggests that authors should be rewarded for their intellectual efforts. Similarly, utilitarian perspectives have shaped the idea that society benefits when authors are granted exclusive economic rights over their works.[7] The rationale is that, because exclusive rights grant authors a legal monopoly, they provide a financial incentive to create and disseminate. That, in turn, is meant to help to correct potential market failures that would otherwise result in underproduction from free-riding on others' efforts. Another foundational idea, influenced by Hegelian philosophy, views creative works as extensions of the author's

[5] See (n.d.). The Earliest Surviving Manuscript Written in Ireland, the Oldest Surviving Irish Manuscript of the Psalter, and the Earliest Recorded Historical Case-Law on the Right to Copy. Jeremy Norman's History of Information. https://www.historyofinformation.com/detail.php?id=1396

[6] Menell, P. S. (2000). Intellectual property: General theories. In B. Bouckaert & G. De Geest (Eds.), *Encyclopedia of law and economics* (Vol. 2, pp. 130-131).

[7] Goold, P. R., & Simon, D. A. (2024). On Copyright Utilitarianism. *Indiana Law Journal, 99*(3).

own personality and identity.[8] From this perspective, copyright protection serves not merely an economic purpose but also a moral one: to respect the autonomy of authors, protect their personal connection with their work, and preserve the integrity of their creations.

Copyright law attempts to balance private and public interests by, on the one hand, affording authors with exclusive rights—both economic and moral—over their creations and, on the other hand, carving out certain exceptions and limitations from those rights for the benefit of society. As such, copyright serves to propel the dissemination of original works, preserve cultural heritage, and enable expressions of diversity in copyrighted materials. In many jurisdictions, including the UK,[9] authors are granted exclusive rights to reproduce, distribute, communicate, and adapt their works for a limited period of time. In the UK, that period currently extends to 70 years after the author's death,[10] after which the work enters the public domain and becomes freely available for public use. By granting this control, copyright ensures, first, that the integrity of the original piece can be preserved and that creators are held accountable for their productions and prevent unauthorised alterations or distortions.[11] Second, it ensures that after the established time period of copyright, creations become part of a shared cultural heritage that is accessible to all. Copyright infringement occurs when someone does any of the acts restricted by the copyright in relation to the whole or a substantial part of a protected work without permission.[12] It undermines the author's ability to control how their work is used and deprives them of the opportunity to benefit from their creation.

Copyright is not absolute, and exceptions and limitations play a vital role in its contours. In the UK, as in many other jurisdictions, these provisions have evolved in a piecemeal fashion and define when the public are allowed to carry out acts which would otherwise require permission (such as a licence) from authors as rightsholders. Although the scope and detail of these exceptions vary significantly between

---

[8] Kanu, P. (2008). Intellectual Property and Hegelian Justification. *National University of Juridical Sciences Law Review, 1*, p. 360-361.

[9] S.16(1) Copyright, Designs and Patents Act 1988 (hereinafter, "CDPA 1988").

[10] *Id.* at s.12(1).

[11] Joffrain, T. (2001). Deriving a (moral) right for creators. *Tex. Int'l LJ*, 36, 735.

[12] S.16(2)-(3) CDPA 1988.

jurisdictions, they often reflect broader policy aims, such as promoting education, facilitating research, and encouraging transformative uses that benefit public discourse or innovation. All exceptions and limitations must comply with international legal norms. The so-called "three-step test" in international copyright treaties requires that exceptions and limitations:

i) apply only in certain special cases,
ii) do not conflict with the normal exploitation of the work, and
iii) do not unreasonably prejudice the legitimate interests of the rightsholder.[13]

In practical terms, this means that exceptions and limitations cannot be so broad or general as to undermine the commercial viability of the original work or disproportionately harm authors' legitimate claims to control and profit from their own creations.

Another important feature of copyright, unlike other IP rights, is the automatic nature of its protection. Copyright arises the moment a qualifying work is created and fixed in a tangible medium; no registration or formalities are required for protection to apply. [14] Moreover, copyright is territorial in nature, [15] meaning that authors simultaneously enjoy a patchwork of national copyrights, each governed by domestic law, across different jurisdictions.

Policy considerations in the digital economy have influenced and reshaped the contours of copyright law over the years. The legal framework now has to account for the vast volumes of content being created, shared and monetised through digital platforms, but also with new venues of potential infringement. Online intermediaries and content-sharing services increasingly function as both gatekeepers and distributors of copyrighted material, which has led to debates over their responsibilities and the mechanisms through which rightsholders are remunerated.

---

[13] Berne Convention for the Protection of Literary and Artistic Works, art. 9. World Intellectual Property Organization (1971) (hereinafter, "Berne Convention"); Agreement on Trade-Related Aspects of Intellectual Property Rights, art. 13. World Trade Organization (1994).
[14] Berne Convention, art. 5(2).
[15] Rättzén, M. (2024). Location Is All You Need: Copyright Extraterritoriality and Where To Train Your AI. *Columbia Science & Technology Law Review, 26*, p. 228-229.

At the same time, the ease of digital reproduction and global distribution channels have created new opportunities for rightsholders to disseminate and commercialise their works at scale. Together, these developments have increasingly connected modern copyright policy to the digital ecosystem and its market players. In this light, although copyright's traditional rationales remain relevant, they require recalibration to reflect the structural and economic realities of the digital age.

The challenges posed by AI and their impact on copyright

The rapid advancement of AI developments, particularly in generative applications, has introduced new challenges to copyright law and policy. AI systems frequently rely on access to vast, sometimes web-scraped, datasets for training, which may comprise copyright works. This has raised questions as to whether AI developers need authorisation from rightsholders to train their models on protected content, and if so, what compensation should be paid to those rightsholders.

The nature of AI training

AI training fundamentally operates through TDM, a process in which machine learning models analyse datasets to identify statistical patterns and make new predictions.[16] The success of generative AI models in making such predictions hinges on the quality, quantity, and representativeness of the data used during training, as well as the algorithms' capability to process and "act upon" the data.

Training generative AI models, particularly large language models (LLMs) and models used to generate visual or audio content, begins with the collection of vast quantities of raw data.[17] Models with different use cases will need different types of data, such as text, images, and videos. Because large-scale and web-scraped datasets may lack curation of quality and source integrity, such materials will often contain copyright

[16] Palmer, A., Jiménez, R., & Gervilla, E. (2011). Data mining: Machine learning and statistical techniques. In K. Funatsu (Ed.), *Knowledge-oriented applications in data mining* (pp. 373-374).
[17] Liu, Y. & He, H. et al. (2024). Understanding LLMs: A Comprehensive Overview from Training to Inference. *Neurocomputing, 620*.

works. The data contained in training datasets, once pre-processed, is broken down into smaller units called tokens. These tokens are then transformed into numerical representations that the model can process, a step known as encoding.[18] How encoded tokens are used more precisely will differ depending on the model architecture and design. Transformer-based models, for example, use attention heads to assign weights to input tokens, and encode the position of each token sequence through a process called positional embedding.[19] The attention mechanism in the transformer model then uses these embeddings to weigh the relevance of each token in the context of the entire sequence.[20] This allows the model to relate a single token with other tokens and differentiate between different tokens based on their positions in the sequence. It is from this so-called pre-training process that the model "learns" statistical relationships between tokens.

AI models are typically trained to identify probabilistic associations (e.g. word co-occurrences or stylistic features) rather than "storing" full works. In practice, however, some degree of reproduction can occur at different stages. For example, a "copy" of a work or parts of a work can be made when training datasets are downloaded and processed, or later when an output is generated that closely mirrors training data, a phenomenon sometimes called "memorisation" or data regurgitation. Research suggests that modern LLMs and image-generation models seldom output entire copyright works verbatim but can reproduce substantial fragments on occasion.[21] Developers typically try to minimise such "memorisation", but the risk of unintended replication remains a technical concern.

Impact on creative industries

Generative AI models have rapidly advanced and are already being deployed in many creative fields. Text-based models can be used to create articles, poetry, screenplays,

---

[18] *Id.* at p. 3.

[19] *Id.* at p. 2.

[20] *Id. See also* Ashish Vaswani et al. *Attention Is All You Need*, Arxiv (Aug. 2, 2023), https://doi.org/10.48550/arXiv.1706.03762.

[21] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, b., Ippolito, D. & Wallace, E. (2023). Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (pp. 5253-5270).

and software code. Image-based models can generate pictures from text prompts, while audio-based models can generate songs, melodies, voiceovers, dubbing, and audiobooks. Video generators can be used to produce anything from video frames to movie scenes. The list of possible use cases and applications is expected to continue to grow as new developments make these models more sophisticated and widespread.

Users range from professional creators to more novice users. For experienced creators, generative AI has expanded the range of technical and creative tools, supporting applications such as de-aging and pre-visualisation of scenes to plan shots and sequences in the film industry, and rapid prototyping in visual design and animation. The availability of AI "assistants" such as chatbots means that essentially anyone with a prompt can rapidly produce content that mimics human creativity. This broadening access to tools substantially lowers the entry barriers to content production, enabling individual creators and smaller organisations to produce high-quality text, images, audio or video without necessarily needing large teams. Recent studies suggest that current generative AI could automate as much as one-quarter of tasks in arts, design, media and entertainment sectors in the US.[22] Generative AI increasingly alters the creative process itself, and several large creative and media companies are already integrating AI into their workforces and product offerings. For example, The Washington Post uses AI to convert text articles into natural-sounding audio in multiple languages,[23] and The Economist uses AI in research tasks and video editing.[24] In the UK, BBC Sounds uses generative AI tools to generate subtitles for some of their programmes.[25] In this way, creators themselves are benefitting from AI,

---

[22] Briggs, J., & Kodnani, D. (2023, March 26). *The potentially large effects of artificial intelligence on economic growth* (Global Economics Analyst). Goldman Sachs. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html.

[23] The Washington Post. (2024, May 20). *The Washington Post expands accessibility to AI-read newsletters, enhancing user experience through audio*. https://www.washingtonpost.com/pr/2024/05/20/washington-post-expands-accessibility-ai-read-newsletters-enhancing-user-experience-through-audio/.

[24] The Economist. (2024, August 6). *How we handle AI-generated content*. https://myaccount.economist.com/s/article/How-we-handle-AI-generated-content.

[25] Dimir, A. (2024, August 29). *Using Gen AI to add subtitles on BBC Sounds.* BBC. https://www.bbc.com/mediacentre/2024/using-gen-ai-to-add-subtitles-on-bbc-sounds

by allowing them to produce more content and sometimes new types of content. However, the ultimate impact on the quantity, quality, and variety of creative output remains an open question. Over time, generative AI could also increase productivity demands across industry, putting pressure on all creators to use AI to remain competitive.

There are risks that generative AI poses to creative jobs and livelihoods. Many professional artists, writers, and performers fear displacement. A recent study from Queen Mary University of London in collaboration with The Alan Turing Institute and the Institute for the Future of Work showed that AI had "diminished the skill and agency of creative workers who are now often asked to review AI-generated work rather than creating their own original work, leading to a reduction in the financial value being attributed to creative work."[26] Other expressed worries are that AI will be able to produce substitute content or that creators' unique styles could be mimicked without permission, both affecting creators' economic rights and potentially leading to reputational damage. For example, the Screen Actors Guild–American Federation of Television and Radio Artists (SAG-AFTRA) had its members go on strike in 2024 against video game companies, requiring that actors such as performers and voice actors have the right to give, or deny, consent for the automated use of their faces, voices and bodies, and are provided fair compensation.[27] Similarly, Hollywood writers went on strike in 2023 over concerns of the impact of AI on writers' and actors' jobs.[28] These concerns are heightened by the fact that many creators operate as freelancers or small businesses and are sometimes subject to deceptive contractual agreements with vague or overly broad clauses, which are now granting unlimited rights to the use of their image, voice, or likeness for AI training. This is exacerbated by already precarious working conditions, where creators are pressured into signing contracts

[26] Queen Mary University of London. (2025, January 23). *Creative industry workers feel job worth and security under threat from AI*, https://www.qmul.ac.uk/media/news/2025/queen-mary-news/pr/creative-industry-workers-feel-job-worth-and-security-under-threat-from-ai-.html.
[27] SAG-AFTRA. (2024, July 26). *SAG-AFTRA strikes video games over AI*, https://www.sagaftra.org/sag-aftra-strikes-video-games-over-ai.
[28] Writers Guild of America. (2023, May 2). *Writers Guild of America calls strike effective Tuesday, May 2*, https://www.wga.org/news-events/news/press/writers-guild-of-america-calls-strike-effective-tuesday-may-2.

that offer poor compensation.[29] Some creators and production companies are trying to turn this dynamic around by licensing their works for AI training. In the UK, the Copyright Licensing Agency (CLA), representing British authors, publishers and visual artists, announced earlier this year that they will offer collective licences for generative AI training.[30] Given the novelty of these licensing mechanisms, it is currently unclear to what extent this could become a sufficient revenue stream for rightsholders.

Looking ahead, AI agents—autonomous programs that, to some degree, can plan and execute multi-step tasks—are also beginning to transform creative workflows in ways that push beyond current generative AI capabilities. These systems differ from single-purpose tools by operating with greater independence, integrating external tools, and interacting with other agents and services. This has the potential to manage entire production pipelines from conceptualisation to final output. Further on the horizon, artificial general intelligence, i.e. AI with broad human-like capabilities, represents one of the long-term ambitions in the field of AI development. AI with such capabilities could profoundly transform the creative industries, like many other sectors of society. Unlike today's AI that excels at defined or specific tasks, AI that, in theory, could operate as autonomous creative entities could further challenge our understanding of who "creates" art and what we consider as an artistic work, as AI becomes not just a tool, but both the beginning and the end of the creative process.

How AI is challenging copyright

The rise of generative AI has sparked fundamental questions about whether the principles and institutions of copyright remain viable in their current form. Traditionally, copyright has been concerned with the creation and commercialisation of individual original works as a result of authors using their skill, knowledge and labour. Generative AI technologies challenge that practice by democratising the creation process to an unprecedented scale and speed, while significantly reducing

[29] Institute for the Future of Work (2025). *Creative Industries and GenAI: Good Work impacts on a sector in transition.* London: Institute for the Future of Work.
[30] Copyright Licensing Agency. (2025, April 23). *Development of CLA generative AI licence*, https://www.cla.co.uk/development-of-cla-generative-ai-licence.

marginal costs. This enables both professional authors and newcomers, and particularly those who may lack technical or artistic training to produce polished, expert-level content. Generative AI turns its training data, which includes creative works, into a valuable commodity. What governments are left to grapple with is who should reap the economic benefits from that value—and under what conditions.

What is central to this debate is whether AI's ingestion of copyright works constitutes permissible "learning" or infringing reproduction.[31] Proponents of the former argue that statistical pattern recognition, the core of machine learning, is analogous to human learning and to inspiration-seeking, which constitutes an established practice in the field, and should fall under applicable exceptions and limitations such as fair use,[32] research, or temporary copying. They contend that treating AI training as infringement, which triggers the need to obtain a licence, would stifle innovation, as no developer could feasibly license every piece of text, image, or audio included in a large-scale dataset. Critics, however, point out that AI training involves the systematic and often commercial copying of works, even if only in tokenised or compressed form. They argue that this process is fundamentally different from human cognition: machines do not merely learn, but replicate and remix works at an industrial scale, and often without attribution. Moreover, it is asserted that the outputs of generative AI could in some cases compete with original creators, depending on how users formulate the prompts.[33] For these reasons, many participants from creative industries believe such use must be licensed and authors appropriately remunerated.

---

[31] Another related question is whether generative AI models "communicate" works to the public through their outputs, which becomes relevant when models have memorised parts of the training data. A recent referral to the European Court of Justice touches upon these and related issues, *see* C-250/25, *Like Company v Google*.

[32] In a recent decision in *Bartz v. Anthropic PBC*, Case 3:24-cv-05417-WHA (N.D.Ca.), Judge Alsup found that Anthropic's use of copyrighted materials in AI training, without permission from certain rightsholders, constituted fair use in the United States., because it was considered highly transformative and there had been no evidence from the plaintiffs (who were individual book authors) that Anthropic's Claude would usurp the market for their works.

[33] See, e.g., Complaint at p. 5, Andersen v. Stability AI Ltd., No. 23-cv-00201-WHO (N.D. Cal. Filed Oct. 30, 2023) ("These resulting derived images compete in the marketplace with the original images. Until now, when a purchaser seeks a new image 'in the style' of a given artist, they must pay to commission or license an original image from that artist. Now, those purchasers can use the artist's works contained in Stable Diffusion along with the artist's name to generate new works in the artist's

Though licensing generative AI training data presents unique challenges compared to traditional licensing frameworks, it is believed that much of the data contained in datasets used by generative AI models has been scraped from publicly available online sources,[34] where authors may be unknown, uncredited, or difficult to identify. Some of these datasets may also contain pirated content.[35] The number of potential rightsholders with legitimate claims may be very vast as a result. Although individual claims from individual rightsholders may be small, collectively the total potential liability could be substantial. The UK Government's recent public consultation on copyright and AI placed emphasis on licensing as a possible solution (See *Section 5: The UK's Copyright and AI: Consultation 2024-2025*),[36] but left the door open as to how that could be implemented, and also as to possible copyright exceptions and limitations (in which case no licence would be needed), especially for non-commercial TDM practices.

In this regard, policymakers face a risk of either overprotection, which could hinder AI innovation or its accessibility, or underprotection, which could prejudice human creators. Generative AI also raises new questions about the appropriation of an artist's personal expression or style. Copyright law protects original, specific

---

style without compensating the artist at all."); Complaint at p. 2, Authors Guild v. OpenAI Inc., No. 1:23-cv-08292 (S.D.N.Y. filed Sept. 19, 2023) ("Defendants' LLMs endanger fiction writers' ability to make a living, in that the LLMs allow anyone to generate—automatically and freely (or very cheaply)—texts that they would otherwise pay writers to create. Moreover, Defendants' LLMs can spit out derivative works: material that is based on, mimics, summarizes, or paraphrases Plaintiffs' works, and harms the market for them").

[34] For example, OpenAI has stated that it has used publicly accessible datasets such as Common Crawl, WebText2, and Wikipedia for training ChatGPT-3. Tom B. Brown et al., *Language Models are Few-Shot Learners*, Arxiv (July 22, 2020), pp. 8-9, https://arxiv.org/abs/2005.14165. See also the UK Government's public consultation on copyright and AI, at para. 42 (stating that, "[b]ut in many cases, AI models are trained using works made available to the public on the Internet. These are often not expressly licensed for AI model training, and the creators of those works are not compensated for their use.").

[35] Reisner, A. (2025, March 20). *The Unbelievable Scale of AI's Pirated-Books Problem.* The Atlantic. https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ and RettighedsAlliancen (March 2025). *Report on pirated content used in the training of generative AI.* https://rettighedsalliancen.dk/wp-content/uploads/2025/03/Report-on-pirated-content-used-in-training-of-AI.pdf

[36] See paras. 13 and 94-98 of the consultation.

expressions, not artistic style or vibe,[37] which leave creators vulnerable when AI mimics their content or technique, or actors' appearances or voices. This so-called "creation gap" has prompted calls for extending copyright protection or establishing new exclusive rights such as publicity or personality rights, which would let creators control how their likeness and personal identifiable features are commercially used.[38]

Central to advancing discussions on copyright challenges are efforts to address key accountability and transparency concerns. Developers tend not to provide detailed documentation on the sources and content of the datasets used to train their models. As a result, creators currently struggle to know whether their works have been used as part of training datasets. This complicates the enforcement and licensing of such works. It also places the burden on creators to identify infringements. At the scale and speed at which generative AI models are built and used, this entails an immense effort. Policymakers are exploring requirements for AI developers to publish details about their training data (so-called "transparency reports"). The EU AI Act,[39] for example, will obligate AI developers to document the data used.[40] Similar measures are under discussion in the UK as transparency was highlighted in the UK Government's public consultation as "fundamental" to build trust and to enable rightsholders to enforce their rights.[41]

Another problem arises if AI-generated content, as opposed to human-produced content, becomes the dominant form of content creation in the future. If AI-generated content then becomes a substantial part of the training corpus, there is a risk that AI, being trained on itself, could degrade in quality and diversity over time. Models trained primarily on AI-generated material—rather than on human-made works—may learn from increasingly homogenous, repetitive inputs. When that happens, a phenomenon,

---

[37] Rättzén, M. (2024). Location Is All You Need: Copyright Extraterritoriality and Where To Train Your AI. *Columbia Science & Technology Law Review, 26*, p. 201-202.

[38] See the UK Government's public consultation on copyright and AI, at paras. 169-179.

[39] Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (hereinafter, the "EU AI Act").

[40] *Id.* at Article 53.1(a), referring to the information set out in Annex XI (which includes "information on the data used for training"), and Article 53.1(d) (referring to a "sufficiently detailed summary about the content used for training").

[41] See the UK Government's public consultation on copyright and AI, at paras. 11 and 103-107

sometimes referred to as “model collapse”, could lead to a feedback loop of declining diversity in outputs and deteriorating performance.[42]

The future of AI in creative fields will depend on how policies reconcile all these factors. Ultimately, generative AI both enriches and destabilises the creative ecosystem, and the copyright framework must adapt to this new paradigm. The question is how. It is a matter for policymakers to address these difficult questions in a manner which strikes a balance between the various stakeholders and diverse interests at stake. Hybrid stakeholders illustrate the nuance that is needed. These creators may want both open access to data when using AI tools themselves, and compensation when others use their original works. Policymakers will need to consider such complex roles.

[42] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 629*(8012), 1122-1128.

Section 3. Copyright, Designs, and Patents Act 1988

Copyright in the UK is governed by the Copyright, Designs, and Patents Act (CDPA) as amended by statutory instruments, incorporation of EU directives and post-Brexit adjustments. The current copyright regime in the UK, in comparison to the US and EU regimes, offers very narrow exceptions, making the interaction with AI particularly challenging.

At the outset, it is important to note that EU copyright legislation and precedents dating from before the UK's exit from the EU also continue to apply in the UK. Although the EU has never attempted any comprehensive harmonisation of copyright law, unlike for trade marks and design rights, it has passed legislation addressing specific issues or discrete subject matter.[43]

The EU courts have expanded the application of these Directives over time, most notably by the Court of Justice of the European Union (CJEU) in the landmark case of *Infopaq International A/S v Danske Dagblades Forening.*[44] The court in *Infopaq* combined the "*author's own intellectual creation*" standard for originality and subsistence from the Software, Database and Term Directives with rights of copyright holders defined in the InfoSoc Directive, stating that a "*harmonised legal framework for copyright*" had been established.

[43] Such as the Directive 91/250/EEC on the legal protection of computer programs (since codified under Directive 2009/24/EC, together the Software Directive), the Directive 96/9/EC on the legal protection of databases (the Database Directive), Directive 2001/29/EC on copyright and related rights in the information society (the InfoSoc Directive), and Directive 2006/116/EC (the Term Directive).

[44] Infopaq International A/S v Danske Dagblades Forening, (2009). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62008CJ0005

The scope and substance of “copyright” in the UK

Copyright is a type of intellectual property that provides a human author (known as the rightsholder) a legal right to protect their creative works. Generally, copyright protections under the CDPA apply to two main categories of works:

- **Authorial works**: include literary, dramatic, musical or artistic works (LDMA) and must be “original”.[45]

  To be original under the UK case law, authorial works needed to exhibit a degree of labour, skill or personal judgement in their creation, which was a low threshold to satisfy. However, the UK was still part of the EU during the *Infopaq* judgment. That case arguably expanded the application of the "authors' own intellectual creation" test for originality from the discrete categories of works covered by the Software Directive and InfoSoc Directive to cover all categories of copyright works, and this approach has recently been confirmed by the English Court of Appeal.[46] The European standard requires that the work be a result of “free and creative choices” that reflect “the author’s personal touch.”[47]

- **Entrepreneurial works**: include sound recordings, films and broadcasts and do not need to satisfy the same originality standard in order to benefit from

[45] Copyright, Designs and Patents Act (CDPA) 1988 s.1(1)(a). Available at: https://www.legislation.gov.uk/ukpga/1988/48/contents; Cordell, N., & Potts, B. (2024, August 20). *Ownership of AI-generated content in the UK*. A&O Shearman. https://www.aoshearman.com/en/insights/ownership-of-ai-generated-content-in-the-uk; Appleton, S. (2025, April 11). *Ownership Issues in AI-Generated Content: Who Owns the Copyright?*. Moore Law. https://www.moore-law.co.uk/ownership-issues-in-ai-generated-content-who-owns-the-copyright/.

[46] *THJ Systems Ltd. v. Sheridan* [2023] EWCA Civ 1354 (THJ Systems) https://caselaw.nationalarchives.gov.uk/ewca/civ/2023/1354?query=THJ+Systems+Ltd+Daniel+Sheridan

[47] Cordell, N., & Potts, B.; THJ Systems Ltd. v. Sheridan [2023] EWCA Civ 1354 (THJ Systems) at [23-24]; Boston, M. (2024, May 14). Originality in the UK means “the author’s own intellectual creation” including for software generated works – the debate is over (… for now). Fieldfisher. Available at https://www.fieldfisher.com/en/services/intellectual-property/intellectual-property-blog/originality-in-the-uk-means-the-author-s-own-intellectual-creation-including-for-software-generated-works-%20the-debate-is-over-for-now.

copyright protections, but they cannot solely consist of copies of another's copyright works.[48]

According to section 9 (1),[49] copyright protections of an original authorial work belong to the author, meaning the person who creates the LDMA work. Section 9 (2) of the CDPA maps the designated author in each instance of the recognised entrepreneurial works.[50]

As per Section 16 (1) CDPA, the substance of the legal right conferred upon the rightsholder is the exclusive use of their copyright. That is, only the rightsholder can copy the work, issue copies of the work to the public, rent or lend the work to public, perform, show or play the work in public, communicate the work to the public, make an adaptation of the work.[51] The nature of this bundle of rights is both economic[52] and moral.[53]

As such, if a person or entity wishes to carry out one of the acts restricted by copyright in respect someone else's copyrighted work, under UK law, unless that act benefits from one of the exceptions from infringement liability they are required to gain express permission from the rightsholder for any such use to avoid copyright infringement. Permission can be granted through a license or assignment of copyright.

Copyright is a legally enforceable right, i.e., the rightsholder can seek a range of judicial remedies against an infringer, such as injunctive relief, damages, destruction of copied works, among others.

---

[48] Cordell, N., & Potts, B.

[49] CDPA s.9(1).

[50] CDPA s.9(2).

[51] See also: s.17 – s.21 CDPA 1988

[52] That is, the rightsholder can gain remuneration through commercial exploitation of their creative works through sales, licensing or royalties.

[53] The UK recognises four non-transferrable moral rights under Chapter IV of the CDPA: the right to attribution, the right to object to derogatory treatment of a work, the right to object to false attribution, and the right to privacy of certain photographs and films. However, these moral rights do not apply to computer-generated works.

Computer-generated works

The UK was among the few jurisdictions in the world to recognise the copyrightability of a computer-generated work and was one of the first to do so back in 1988 at the outset of the CDPA.[54] A "computer-generated" work is defined as "*generated by computer in circumstances such that there is no human author of the work*."[55] For computer-generated works protected by copyright as LDMA, section 9 (3) of the CDPA recognises the person who makes the "arrangements necessary for the creation of the work" as the proper author.

However, if the EU test for originality is now to be applied in the UK, the future applicability of section 9 (3) is in doubt. Case law of the CJEU, which remains the law in the UK, has linked the "authors' own intellectual creation" standard for originality to "the author's personality".[56] Whilst the CJEU has not explicitly stated that this necessarily requires the presence of a human author, the Advocate General's opinion in the *Painer* case has stated that only human authors can qualify for copyright protection. If this is now the position under UK law, computer-generated works without a human author may no longer qualify for copyright protection.

If a human author is not pre-requisite for subsistence of copyright under UK law, there are nevertheless differing interpretations of what these necessary arrangements entail and complex technological and organisational contexts in which such arrangement are made, such as in hybrid use cases (see *Section 6: Hybrid use cases)*, often make it unclear who exactly the author of a computer-generated work is.[57]

---

[54] CDPA s.9(3); Cordell, N., & Potts, B.; Lee, J. (2021). Computer-generated Works under the CDPA 1988. *The Chinese University of Hong Kong Faculty of Law*, 2021(65), 177, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3956911. Other jurisdictions include New Zealand (Copyright Act 1994 s.5(2)(a)), Ireland (Copyright and Related Rights Act 2000 21(f)), India (Copyright Act 1957 s.2(d)(vi)), Hong Kong (Copyright Ordinance s.11(3)).

[55] CDPA s.178. Under this definition, however, it is unclear where the line is for what constitutes "computer-generated" as compared to "computer-assisted" work. Erickson, K. (2024, January 31). *Copyright protection in AI-generated works*. Creative Industries Policy and Evidence Centre. https://pec.ac.uk/blog_entries/copyright-protection-in-ai-generated-works/.

[56] *Eva-Maria Painer v Standard VerlagsGmbH and Others*. [2011] Case C-145/10 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62010CJ0145_SUM

[57] See Appleton, S. (2025, April 11). *Ownership Issues in AI-Generated Content: Who Owns the Copyright?.* Moore Law. https://www.moore-law.co.uk/ownership-issues-in-ai-generated-content-

The English courts have only explicitly addressed this issue once in *Nova Productions Ltd. v. Mazooma Games Ltd.* back in 2007.[58] There, the Court of Appeal held that images generated during gameplay were protectable computer-generated works, and held that the programmer of the game, rather than the player of the game itself, was the person who had made the necessary arrangements for the creation of the images.[59] The court considered the fact that the programmer had written the code for the program and designed the rules and logic to the game to point in favour of his authorship. Scholars have proposed other factors that may be relevant in understanding what arrangements are necessary to constitute authorship, including: the initiative or intention to create the work, the proximity to the act of final creation, the extent to which the arrangements are responsible for the creation of the work, the extent to which the arrangements are responsible for shaping the work, and broadly, investment.[60] By analogy, there may be disputes as to whether the provider of a managed AI tool or the user of the tool is the "person undertaking the arrangements necessary".

One other difference of note between computer-generated works and traditional works is the duration of copyright protections. According to section 12 (7), computer-generated works are only protected for fifty years from the year in which the work was made, whereas other works are generally protected for seventy years from the year in which the author dies.[61]

---

who-owns-the-copyright/; Lee, J. (2021). Computer-generated Works under the CDPA 1988. *The Chinese University of Hong Kong Faculty of Law*, 2021(65), page 193, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3956911

[58] Syn Ong, Authors Alliance, May 19 2025*, The UK's Curious Case of Copyright for AI-Generated Works: What Section 9(3) Means Today*, https://www.authorsalliance.org/2025/05/19/the-uks-curious-case-of-copyright-for-ai-generated-works-what-section-93-means-today/.

[59] Nova Productions Ltd. V. Mazooma Games Ltd. & Ors. [2007] EWCA Civ 219 (Nova).

[60] Lee, J. (2021). Computer-generated Works under the CDPA 1988. *The Chinese University of Hong Kong Faculty of Law*, 2021(65), page 188, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3956911.

[61] CDPA s.12(1)-(3), s.12(7).

Copyright infringement liability

Section 16 CDPA provides for a "closed list" of specific acts which, if carried out in the UK in relation to a copyright work with the permission of the copyright owner, would constitute copyright infringement.

EU copyright law takes a different approach. It does not provide for a fixed list of different types of infringing act. Instead, the legal test for copyright infringement established by *Infopaq*[62] centres on the concept of the "*appropriation of the author's own intellectual creation*" and assesses whether the use of a work involves the unauthorised appropriation of the original intellectual effort and creativity of the author. The ruling, which has been confirmed in subsequent case law,[63] emphasises that even small parts of a work, such as a few words from a newspaper article, can be protected if they reflect the author's intellectual creation, meaning that appropriation of even such minor parts of a work can constitute copyright infringement.

Exceptions from copyright infringement liability

UK copyright law also provides a "closed list" of exceptions from copyright infringement liability that apply in specific, defined circumstances; if an otherwise infringing act falls within the scope of an exception, the person carrying out that act will have no liability for copyright infringement. This is in contrast to the approach taken in territories such as the US, where there is a general concept of "fair use", discussed later in this section, that can be, and has been, adapted to accommodate technological developments.

The CDPA, as originally enacted, provided a long list of exceptions for specific circumstances, but most of the current exceptions that are applicable to software and the internet have been added through amendments to the CDPA in 2003 and more recently, in 2014, to allow for an expansion of modern exceptions that otherwise could leave users liable for incidental infringements in a changing digital age.[64]

---

[62] ECJ, Case C-5/08

[63] Cofemel v G-Star Raw Case C-683/17

[64] The Copyright and Related Rights Regulations 2003. Available at: https://www.legislation.gov.uk/uksi/2003/2498/contents; The Copyright and Rights in Performances

Closed list exceptions

First, exceptions exist to allow for the use and copying of lawfully obtained copyright work if is being used in a way that helps blind or otherwise disabled people access it.[65] To apply this exception, notification to the copyright owner is required.[66] Similarly, there are exceptions for educational, illustrative, non-commercial research, and private study purposes if the use of copyright works conforms with the principle of fair dealing and the original owner is sufficiently acknowledged.[67] Exceptions can also be applied, subject to fair dealing, for criticism, review, quotations, and reporting of current events.[68] Quotations were specifically added in the recent amendment in 2014 to give people greater freedom to quote the works of others as long as the use is reasonable and fair.[69] Generally, long extracts and photographs are not included in this.

TDM involves accessing and analysing large datasets to identify patterns and trends to train AI. Obtaining permission for this typically involves acquiring a licence or relying on an exception. As the law currently stands, Section 29A (1) CDPA only allows for a TDM exception for non-commercial research.[70] Although the CDPA does not explicitly define "non-commercial use", it generally means something that is not primarily intended for monetary compensation by either an individual or an organisation.[71] Additionally, Recital 42 of the InfoSoq Directive, which underpins Article 29A, also provides guidance on the meaning of non-commercial use. It states that whether an activity is non-commercial should be judged based on the nature of

---

(Personal Copies for Private Use) Regulations 2014. Available at: www.legislation.gov.uk/uksi/2014/2361/contents/made;
Intellectual Property Office. (2014). *Exceptions to copyright: Guidance for creators and copyright owners*. GOV.UK. July 2025. 1.

[65] CDPA s.31A; Intellectual Property Office. (2014) at 3; *see also* Intellectual Property Office. (2019). *The Marrakesh Treaty*. GOV.UK. July 2025. https://www.gov.uk/guidance/the-marrakesh-treaty.

[66] CDPA s.31A; Intellectual Property Office. (2021).

[67] CDPA s.32-s.36; Intellectual Property Office. (2021); Intellectual Property Office. (2014) at 4.

[68] CDPA s.30.

[69] *Id.*; Intellectual Property Office. (2014) at 7.

[70] CDPA s.29A.

[71] Intellectual Property Office (2014, October). *Exceptions to copyright: Guidance for creators and copyright owners*. GOV.UK. Available at https://assets.publishing.service.gov.uk/media/5a7f4cf640f0b62305b864e6/Exceptions_to_copyright_-_Guidance_for_creators_and_copyright_owners.pdf

the activity itself rather than the organisation's structure or how it is funded. Academic commentary suggests that commercial entities could then, in principle, engage in non-commercial research if it is carried out for a purpose that does not generate revenue.[72]

Therefore, the TDM exception in the UK allows for non-commercial researchers to make temporary copies of protected works such that automated techniques can be used to analyse large amounts of information which the UK has recognised can have valuable research purposes.[73] Non-commercial researchers must already have lawful access to the copyright work they are analysing to apply this exception.[74] Examples for lawful access include paying for subscription to a journal or database or through open licences.[75] Relatedly, the CDPA allows for the temporary copy of protected works where it is necessary to do so incidentally to transmit the work or for any other lawful purpose.[76]

Whilst some of these exceptions are subject to the concept of "fair dealing", this should not be confused with "fair use" under copyright law of the US; the two are conceptually very different. Most importantly, at the outset, it is necessary to note that unlike fair use in the US, fair dealing in the UK is not an independent exception.

---

[72] Rättzén, M. (2024). Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI. *Colum. Sci. & Tech. L. Rev., 26*, 175, page 215.

[73] Intellectual Property Office (2014b, October). *Exceptions to copyright: Research*. GOV.UK. Available at https://assets.publishing.service.gov.uk/media/5a7d678ee5274a02dcdf4502/Research.pdf; Simmons + Simmons. (2025, January 13). *How can AI and copyright co-exist? – UK Government re-consults*. https://wwnw.simmons-simmons.com/en/publications/cm5v8hen106d8tr0k3psasqxi/uk-ai-copyright-policy.

[74] The UK allows database and platform owners to provide security measures on their platforms, generally, however they must not unreasonably restrict researcher's ability to access the material. Similarly, contracts to limit researcher's ability to take advantage of this exception are unenforceable. Intellectual Property Office. (2014). *Exceptions to copyright: Research*. GOV.UK. August 2025. 6. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

[75] Intellectual Property Office (2014b, October). *Exceptions to copyright: Research*. GOV.UK. Available at https://assets.publishing.service.gov.uk/media/5a7d678ee5274a02dcdf4502/Research.pdf

[76] CDPA s.28A.

Fair Dealing vs Fair Use

Fair Dealing	Fair Use
A legal exception for copyright infringement under UK law used to establish whether a use of copyright material is lawful and, therefore, whether it infringes on copyright.[77] Fair dealing allows for the use of a limited amount of someone else's protected work without a license.[78] The CDPA does not prescribe a set standard defined in UK legislation for what constitutes fair dealing. Instead, whether a dealing is fair is always a question of fact, degree, and impression in each case and ultimately requires analysis as to whether a fair-minded and honest person would have dealt with the work in a similar manner.[79] Some of the factors taken into consideration can include the amount of work taken and how significant it was, whether the owner is likely to suffer economically from the use of the	A legal exception for copyright infringement in the US, specifically 17 U.S.C. § 107.[81] Statutorily, there are four factors involved in a judicial analysis for determining fair use: • the purpose and character of such use, including whether such use is of a commercial nature or is for nonprofit educational purposes; • the nature of the copyright 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 value of the copyrighted work Fair use is broad in scope and allows for uses that fall outside of specific

[77] Intellectual Property Office. (2014) at 8.; Intellectual Property Office. (2021).

[78] Intellectual Property Office. (2014) at 8.; Businesses may have a harder time proving that their use is conformant with fair dealing than individuals, but ultimately, the analysis is circumstantial and depends on the nature of the use and not the identity of the person doing so. Bainbridge, J. (2025, March 31). *A guide to copyright and fair dealing in the UK*. Harper James. https://harperjames.co.uk/article/fair-dealing-copyright/.

[79] Intellectual Property Office. (2014) at 8.; Intellectual Property Office. (2021).

[81] U.S. Copyright Act of 1976, 17 U.S.C. § 107 (2025). https://www.law.cornell.edu/uscode/text/17/107

work, as well as how the artistic work has been used and if it had been previously published or not. As such, fair dealing is a narrow exception and only applies to a limited list of uses including non-commercial research and private study, quotations, criticism, current events reporting, and works of parody, but not including text and data mining or temporary copies, for example.[80] Current TDM activities to train AI amount to commercial activities. TDM is allowed in the UK only for non-commercial activities. As such, fair dealing remains a factor restricted to assessing whether an exception applies only in those closed list non-commercial scenarios, such as non-commercial research and private study.	categories so long as the use is considered fair within the aforementioned factors. Fair use also places strong emphasis on transformative use, i.e. whether something new has been added with a further purpose or different character that does not substitute for the original.[82]

Legal challenges of AI training

AI training processes can take many forms. The legal significance of the steps in the training process primarily concerns the following: copyright gives a copyright owner the right to prevent others from carrying out acts restricted by copyright, the most important "restricted act" being the making "copies" of a work.

[80] Intellectual Property Office. (2014) at 8.; Design and Artists Copyright Society (DACS). (n.d.) *Permitted uses and exceptions*. July 2025. https://www.dacs.org.uk/advice/articles/copyright-infringement/permitted-uses.

[82] Owen, R. (2024, April 22). *Understanding fair dealing and leveraging licensing*. CLA. Availabel at https://cla.co.uk/understanding-fair-dealing-and-leveraging-licensing/

The first key question in assessing whether an AI training process[83] infringes copyright is therefore whether *copies* are made during the training process at all. This is a factual question that must be answered by reference to the technical details of a given AI training process.

The details of training processes are often kept secret by the AI developer. However, one way of training an LLM involves repeatedly loading the same text into computer memory, masking words in a sentence and requiring the model to predict what the missing words should be, and iterating through the sentence in successive rounds of training, masking and un-masking different words each time and adjusting the model parameters through each iteration. This may allow a model to calculate the probabilities of words within a sentence, and also to determine which words are most relevant, and most determinative of the meaning of the sentence. Such a process involves making large numbers of copies of the works in the training dataset. An alternative approach for image generation models is known as "diffusion" training, where the model iterates over numerous copies of an image with successively greater amounts of "noise" added to the image, making it progressively more difficult to identify the content and features of the image *(see Section 2: Rationale for copyright law and the impact of AI).*

Most of these copies will be temporary and transient (which is relevant to one of the potential copyright exceptions), but permanent copies may also be made at various points in the process of crawling, downloading and preparing the training dataset prior to commencing the training process.

An unresolved question, both technically and legally, is whether (or the extent to which) works are "memorised" by AI models, which (unless input and output filters are applied) appear able to generate or re-constitute verbatim copies of texts from the training dataset, even if a copy of that text is not "stored" in the same way as a traditional computer text file. Some research shows that generative AI models do not

[83] It is important to note that the AI training process involves multiple steps, and what the exact steps are will vary depending on what type of AI technology is concerned.

frequently output entire copyright works verbatim but can occasionally reproduce substantial fragments of the same.[84]

The geographical location where the training process takes place can have a significant impact on the question of legal liability.[85] Copyright exceptions and limitations are harmonised to a limited extent through international treaties such as the Berne Convention, which permits signatory countries some flexibility on how they implement exceptions from copyright infringement liability. The same set of facts underlying a dispute regarding AI training could lead very different outcomes depending on which country's law applies. A full discussion of the conflict of laws principles applicable to copyright is beyond the scope of this chapter, but in general the applicable law will be that of the physical location of the servers on which the training process is carried out and (if different) the location of the individuals who are operating the servers.[86]

Getty Images v Stability AI[87]

On 4 November 2025, the High Court of England and Wales handed down judgment in *Getty Images (US) Inc. and others v Stability AI Ltd*, the UK's first and most important ruling on some of the applicable intellectual property issues concerning generative AI training cases and their international context.

The claimants initially alleged that the training process of the defendants' "Stable Diffusion" image generation model infringed their copyright in large numbers of photographs. In late 2023, the defendants sought to "strike out" the claims relating

[84] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, b., Ippolito, D. & Wallace, E. (2023). Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (pp. 5253-5270).

[85] Rättzén, M. (2024). Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI. Colum. Sci. & Tech. L. Rev., 26, 175, pages 235-241.

[86] For more discussion, see Rättzén, M. (2024). Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI. Colum. Sci. & Tech. L. Rev., 26, 175, pages 235-239.

[87] High Court of Justice (England & Wales). (2025). *Getty Images v. Stability AI*. Retrieved from https://www.judiciary.uk/wp-content/uploads/2025/11/Getty-Images-v-Stability-AI.pdf ; Popple, L. (2025, July 9). Getty Images v Stability AI: where are we after the trial copyright? *TaylorWessing.* https://www.taylorwessing.com/en/insights-and-events/insights/2025/07/getty-v-stability

to infringement of UK copyright, on the basis that no aspect of the training process was carried out in the UK. Whilst the strike out application failed because there was some doubt, factually, as to whether some training had taken place in the UK, by the time the case reached trial in July 2025, the claims relating to AI training had been dropped entirely, rendering them beyond the scope of the judgment of 4 November 2025 which did not address the question of whether AI training carried out in the UK infringes UK copyright. A parallel claim between essentially the same parties was commenced in the US courts in January 2025 but has yet to reach trial.

The claimants had also alleged that certain outputs of the Stable Diffusion model infringed the copyright in certain images within the training dataset. Whilst this part of the case was still live at the beginning of the trial, by the end of the trial the claimants had dropped that aspect of the claim, so those claims will not be addressed in the judgment.

At the outset of the claim, Getty Images had argued that the pre-trained Stable Diffusion model itself constituted a copy of the copyright works within the underlying training dataset. This would have made importation and use of the model in the UK an act of primary copyright infringement. However, this claim had also been dropped by the time of trial.

The High Court's judgment did, however, offer guidance on the meaning of "article" and "infringing copy" for the purposes of secondary copyright infringement.

The CDPA provides that the importation, possession in the course of business, sale or offer for sale, or distribution in the course of business of an "article" that a person knows or has reason to believe in an infringing copy of a work is an act of secondary infringement (CDPA ss 22-23). An "article" is an infringing copy if its making constituted an infringement of a work (CDPA s.27(2)) or if it (a) has been imported from outside the UK and (b) its making in the UK would have constituted an infringement of the copyright in the work in question (CDPA s.27(3)).

Getty claimed that Stability AI was liable for secondary copyright infringement under sections 22 and 23 of the Copyright, Designs and Patent Act 1988 by virtue of the importation of Stable Diffusion (through its downloading in the UK) and its distribution in the course of business via the Hugging Face platform.

Siding with Getty, the judgment held that the term "article" was not limited in its definition to tangible things and could include electronic copies stored in an

intangible form and software, such as was the case for the pre-trained Stable Diffusion model. By making this claim, Getty sought to prevent the importation and use of the model in the UK, even if the model had been trained outside of the UK.

However, and in siding with Stability, the judgment rejected Getty's claim that Stable Diffusion was an "infringing copy" merely because its making via the repeated exposure of the model weights of Stable Diffusion to training data comprising Getty's copyright works would have constituted infringement if done in the UK. The judge observed that while the model weights are altered during the process of training by exposure to copyright works in the Getty Assets, the model weights are not themselves an infringing copy, nor do they store or reproduce an infringing copy. Moreover, by the end of the training process, the model does not store any copyright works in the Getty Assets, nor has it ever done.

This whittled down in scope ruling leaves it open, potentially for some time, for the questions relating to primary copyright infringement to be resolved in the UK. There are other cases in the UK that have been threatened, but that means it will still be some years before they are decided or settled without judicial determination. As for Getty and Stability, they are locked in a parallel ongoing case in California where there might be more clarity on the primary infringement issues, albeit under applicable US law.

Section 4. Standardising copyright

Under current UK copyright law, developers of AI models are, in principle, required to obtain permission from rightsholders before using their copyrighted works for training AI systems that serve a commercial purpose. Under EU copyright law, TDM for commercial purposes is permitted unless the copyright owner has reserved their rights. A similar opt-out regime is one of the options that has been proposed for the UK and is discussed in detail in *Section 5*: *The UK's Copyright and AI: Consultation 2024-2025.*

To construct a training dataset for a generative AI model, a crawler is sent on vast portions of the internet to scrape data. Despite due diligence, it cannot fully be known in advance which restricted data the crawler may collect. AI developers also commonly use pre-scraped data that has already been collated by others. Once the data has been collected, it is a significant practical challenge:

a) to know if restricted data formed part of the training dataset given the size and opacity; and
b) if it did, then securing permissions retroactively may be extremely difficult as it would require identifying the countless rightsholders and contacting each one of them.

In the absence of legal certainty about whether TDM is fair game[88] unless the rightsholder declares a machine-readable opt out, and amid ongoing litigation, awareness and use of technical standards have arguably gained prominence. This is particularly so with the Robots Exclusion Protocol (REP) which was originally designed to instruct search engine crawlers on which pages to index or ignore, and not as a mechanism to facilitate copyrights management. However, notwithstanding its shortcomings, it is now being repurposed for copyright by rightsholders to communicate an opt-out, and by developers to recognise the opt-out to avoid unintended infringement.

[88] Currently, there is uncertainty about both the effect of the opt-outs and about whether an opt-out has been declared in respect of a particular work.

Although the current landscape of technical standards falls short of meeting the changing needs of both technologists and rightsholders, they still hold strong potential to support the evolving AI ecosystem in a rights respecting way. To facilitate web crawling that respects rights reservation, there are two key parts of the puzzle that will require attention:

1. **Machine-readable rights reservation mechanisms**
   Mechanisms which clearly signal to crawlers which content they are not allowed to use.

2. **Machine-readable expression of intent behind the opt-out**
   Refers to standardised vocabulary, understood by crawlers, to effectively communicate the opt-out intention (e.g., are rightsholders only opting out of AI training or also search engine indexing?)

Achieving this level of granular control requires a combination of interoperable standards that allow for nuanced permissions that can support a copyright regime that balances interests of rightsholders, technologists, and hybrid stakeholders.

This section maps the existing landscape of technical standards. This mapping explores fragmentation in existing standards. It also demonstrates the potential of technical standards to fill gaps in the ecosystem by examining existing standards and highlighting emerging developments.

What are technical standards?

Technical standards are agreed-upon guidelines and processes that aim to promote assurance, quality, efficiency and trust. Technical standards are of increasing interest in the AI context as mechanisms that promote innovation and trust in AI systems.[89]

Technical standards can be de jure. That is, they are standards which are official, approved and developed by recognised Standards Development Organisations

[89] *Standards at a glance*. AI Standards Hub. https://aistandardshub.org/resource/main-training-page-example/1-what-are-standards/

(SDOs) such as International Organization for Standardization (ISO) or Institute of Electrical and Electronics Engineers (IEEE).

Technical standards can also be de facto, which means they are not officially approved by any SDOs but have fallen into customary use.[90] The REP is an example of a de facto standard.

Standards development is a collaborative and consensus-driven process that often involves a wide range of stakeholders, including experts, academics, civil society, governments, and international organisations. Participation in standards development is voluntary.

For emerging technical standards in this area, please also refer to *Annexure* I.

Current landscape

1 - Machine-readable rights reservation mechanisms

Currently, there are two dominant approaches for machine readable rights-reservation: site-based and unit-based.

- *Site-based* [91] [92]

Site-based or location-based mechanisms allow website operators to set broad, domain-level permissions or restrictions on how content hosted on their site can be used. Instead of addressing individual works, these strategies apply across the entire website or specific webpages of a site.

[90] Bartleson, K. (2012, November 14). What's The Difference Between De Jure And De Facto Standards? *Electronic Design.* Retrieved 22 October 2025 from https://www.electronicdesign.com/technologies/embedded/article/21796209/whats-the-difference-between-de-jure-and-de-facto-standards

[91] Balan, K., Gilbert, A., & Collomosse, J. (2025). Content ARCs: decentralized content rights in the age of generative AI. In *IET Conference Proceedings CP940, 2025*(22), 42-48. doi.org/10.1049/icp.2025.2955

[92] Keller, P. (2024, May 16). Considerations for opt-out compliance policies by AI model developers. *Open Future.* Retrieved 22 October, 2025, from https://www.ietf.org/slides/slides-aicontrolws-considerations-for-opt-out-compliance-policies-by-ai-model-developers-00.pdf

The most established example in this space is the robots.txt file (RFC 9309),[93] which has long been used to control access by web crawlers for search indexing. More recently, reliance on robots.txt has been repurposed as a practical mechanism to signal opt-outs from TDM. As different bots can be targeted by name and specific paths can be included or excluded, these mechanisms offer a relatively simple way for website operators hosting their copyright content to express rights reservation at scale. However, the interests of a website operator and the rightsholders of content hosted on their website may not always be aligned.

Challenges

Although site-based mechanisms such as the de facto robots.txt, are widely adopted, and therefore offer some degree of certainty, they remain an imperfect solution for balancing copyright protection with AI development.

As these mechanisms operate at the site-level, they do not offer granular control, which is what rightsholders might prefer. The website operator and/or the rightsholders may want the website as a whole to be accessible but would wish to retain control over the use of copyrighted content hosted on the website.

Another challenge of reliance on site-based mechanisms to communicate opt-out in the copyright context is that only entities controlling the website or domain can block web crawlers, and the interests of a website operator and the owners of the content hosted on the website may not always be aligned. For example, a website operator's main goal may be to maximise traffic, whereas the rightsholder will have a greater focus on protection of their copyright.

As such, if the rightsholder does not exercise control over the website and the site operator does not opt-out, crawlers will likely scrape data that may be protected by copyright. It is more likely that rightsholders may block bots from their own websites. However, if their content is linked or embedded elsewhere on the internet on a third party site that is not subject to an opt-out, it will probably form part of training datasets

[93] Koster et al. (2022, September). *Robots Exclusion Protocol.* Internet Engineering Task Force. Retrieved 22 October, 2025, from https://datatracker.ietf.org/doc/html/rfc9309

without the crawler or downstream users of the crawl having any way of knowing about the rightsholder's opt-out on its own site.

In other words, effective communication of opt-outs using site-based mechanisms is tied to rightsholder control of website infrastructure.

As the UK's new proposal draws inspiration from the EU's copyright frameworks, a few points warrant attention. The EU opt-out regime does not expressly require that every single copy of a work appearing online must be opted-out in order to suspend the protection given to AI developers by the TDM exception:

1. If a single expression of the opt-out is legally sufficient to suspend the TDM exception in respect of every other copy of the copyrighted work, this makes it even harder for AI developers to ensure compliance. In the example given above, the crawler would need to check all sources of a given copyrighted work for potential opt-outs before a developer can have confidence that it is legally entitled to copy that work for the purposes of AI training; and
2. Alternatively, if every single copy of a work does need to be accompanied by an opt-out, this shifts the onus to the rightsholder to ensure that every copy of their work that appears online is accompanied by an opt-out. For their own sites, or sites that can be controlled via licence or other contractual terms, subject to having sufficient commercial bargaining power, it is at least theoretically possible for the rightsholder to ensure opt-outs are expressed in respect of all authorised copies of their work. However, they have no way of controlling pirated or other unauthorised copies of their work, and crawlers have no way of knowing whether a particular instance of a work is authorised or not.

Robots.txt

Robots Exclusion Protocol

The REP is a de facto standard. That is, it has become a widely accepted custom for controlling web crawler access to websites. The robots.txt file, a foundational element of web protocol, originated in the early days of the internet when it was small enough to keep a list of all bots and address concerns about server overload. The intention was to set polite rules for automated access, and not security or legal enforcement.

As content scraping and AI training have become widespread, website owners have begun using robots.txt to signal disallowance of such practices. However, REP was designed for web crawling and search indexing by search engines, and not to manage data usage rights. Importantly, REP does not allow for nuanced preferences. A site cannot simultaneously remain visible whilst also opting out of AI training. As such, the opt-out made via REP is binary, and does not fully capture a rightsholders expectations.

Technical workings[94]

Service owners implement the REP through the robots.txt file, placing it at the root of the website, which they use to establish rules that they expect crawlers to follow when accessing their content via URIs. For example, [*https://example.com/robots.txt*] provides instructions to web crawlers about which parts of a site may be accessed.

Technically, when a crawler visits a site, it first looks for robots.txt at the root domain. If present, the file sets out directives based on the crawler's user-agent string, for example, *Googlebot* or *Bingbot* or * for all crawlers. Within these sections, site owners may use rules such as Disallow to block crawling of specific paths or Allow to grant exceptions. Some crawlers also recognise optional directives like Crawl-delay to request a pause between fetches or Sitemap to point to an XML sitemap.[95]

The robots.txt file is a publicly available file that can be seen by anyone to learn what parts of a server a website owner does not want bots to use. It is important to note that robots.txt does not prevent all access. It merely requests that compliant

[94] *About/robots.txt*. Robotstxr.org. Available at https://www.robotstxt.org/robotstxt.html

[95] An XML sitemap is a structured file that lists the pages of a website to help search engines discover and crawl content more efficiently. It provides metadata such as update frequency and last modified dates but does not control access to content.

crawlers refrain from visiting listed paths. Malicious bots can and often do ignore it. Even the good or compliant bots may still index a disallowed URL.

Limitations

Not all robots or crawlers will respect the protocol, and while REP shapes crawler behaviour, it cannot offer full access control, fully guarantee against indexing, and lacks legal effect. The technical effectiveness of the protocol depends on crawler adherence. Some crawlers adhere to the instructions, but others, such as malware bots, do not. Additionally, different crawlers may interpret the file's syntax and directives in various ways. This can lead to inconsistencies in how crawlers follow the instructions. Understanding and using the syntax supported by different crawlers can help with this limitation, but it is a burdensome task. These limitations may result in many crawlers to ignore, misinterpret, or bypass the protocol.[96]

The REP does not guarantee that crawlers will not index or partially expose disallowed pages. Even if a page is disallowed from crawling, it can still be indexed. If other sites link to it, its URL and metadata may appear in search results.[97] Preventing indexation, such as in Google search, requires extra measures like protecting the files on the website server with a password, or using a noindex meta tag or response header for the URL.[98] These limitations affect the capacity of the protocol to ensure control over content visibility.

In the copyright context, the biggest limitation of robots.txt is that it cannot differentiate between an opt-out for search indexing and an opt-out for AI training because it was built to manage web crawler access, not rights management or reuse permissions. The simple vocabulary, Allow and Disallow, only instructs crawlers whether they may fetch a page, without containing additional expressions or preferences about AI training opt-out. A rightsholder may apply robots.txt with the intention to block a web crawler from using content for AI training, but its effect will be to prevent compliant web crawlers from accessing or indexing the content for search engine purposes, while non-compliant crawlers may ignore it entirely.

Internet Architecture Board (IAB) AI-CONTROL Workshop[99]

---

[96] Google for Developers (n.d.) *Introduction to robotx.txt.* Retrieved 22 October, 2025, from https://developers.google.com/search/docs/crawling-indexing/robots/intro

[97] Google for Developers (n.d.)

[98] Google for Developers (n.d.)

[99] *IAB Workshop on AI-CONTROL (aicontrolws)*. Internet Engineering Task Force Datatracker. https://datatracker.ietf.org/group/aicontrolws/about/

Recognising the growing tensions between technologists and rightsholders over publicly accessible content, the IAB, a committee of the Internet Engineering Task Force (IETF), convened the AI-CONTROL Workshop in 2024 to take stock of the technical status quo. The workshop brought together researchers, standards experts, and industry stakeholders to examine if current standards, specifically robots.txt, were adequate for expressing opt-outs from AI training. The consensus was that while robots.txt remains a foundational standard for guiding crawler behaviour, it was inadequate to handle complexities of the modern web.[100] It lacks the ability to distinguish between different purposes of access, such as indexing for search versus harvesting data for model training and relies on identifying crawlers by name rather than by function or declared use.

- *Unit-based* [101] [102]

Unit-based protocols offer a more granular approach by allowing creators and rightsholders to attach permissions or restrictions directly to individual pieces of content. Unlike site-based strategies, which operate at the domain level, unit-based approaches embed machine-readable metadata within specific media files, regardless of where they are hosted. This enables persistent control over content, even when it is separated from its original context. Emerging standards designed for the AI training context include embedded metadata systems such as Coalition for Content Provenance and Authenticity (C2PA).[103]

Challenges

A difficulty with unit-based protocols is the assumption that content is static and file-based, when in reality, especially in the virtual world, content is often dynamic and constantly reencodes, resizes, recompresses and also constantly updates such as on news websites, social media, live-streams, or even platforms like GitHub. As a result, unit-based standards may fail to recognise the change in “units”, missing important

[100] Nottingham, M, & Krishnan, S. (2025). *IAB AI-CONTROL Workshop Report*. Internet Engineering Task Force. Retrieved 22 October, 2025, from https://datatracker.ietf.org/doc/draft-iab-ai-control-report

[101] Balan, K., Gilbert, A., & Collomosse, J. (2025)

[102] Keller, P. (2024)

[103] Coalition for Content Provenance and Authenticity. https://c2pa.org/

updates which could lead to difficulties with establishing provenance, misattribute copyright or lead to non-compliance with licensing terms because the metadata that carries the right information is missing.[104] In theory, unit-based opt-out signals, such as "do not train", can travel with the content wherever it goes, across websites or platforms, allowing rightsholders to assert control. However, effectiveness of such opt-outs can be significantly undermined through metadata stripping, for e.g., through a change in format or size of a file. When metadata is removed, the associated opt-out signal is lost, making it difficult to enforce creator preferences.[105]

Unit-based mechanisms can be a double-edged sword. While the granularity offers the nuance that allows rightsholders to express multiple preferences, this same granularity can lead to fragmentation. Without wide adoption of standardised approaches, the effectiveness becomes uneven. Moreover, the burden placed on rightsholders can be significant. At present, the burden on rightsholders is very limited as their copyright is protected automatically and does not necessarily require registration. But the evolving tensions between AI training and balancing rightsholders' interests means more active rights management is expected from the rightsholders. The burden is arguably more significant than if they had to register their copyright. Rightsholders may lack the technical knowledge, resources, and infrastructure to consistently apply various unit-based mechanisms across their content. Rightsholders may also have insufficient control over licensees or distributors of their content, as existing licensing arrangements will be silent as to opt-out requirements, and even sophisticated licensors may have inadequate bargaining power to secure opt-outs when negotiating new agreements with powerful counterparties.

As such, the precision and flexibility offered by unit-based mechanisms risks being undercut by practical and commercial barriers to implementation and interoperability.

[104] Keller (2024)

[105] Keller (2024)

Unit-based mechanisms[106]			
Standard	**Purpose**	**Rights management**	**Limitations**
Coalition for Content Provenance and Authenticity (C2PA)[107]	- Establishes provenance and authenticity through cryptographic binding[108] that travels with the digital asset	- Records who created the work, when, and how it was edited - C2PA is tamper-evident[109]	- Possible to strip metadata[110] - Using non-C2PA tools can impact completeness of provenance data[111] - Does not record use or access. Only relevant for preserving provenance[112]
International Standard Content Code (ISCC)[113]	- Open-source ISO standard that provides a unique content-derived identifier (like a digital fingerprint)[114] - Designed for cross-sector applicability and to identify	- Accounts for the evolving and dynamic nature of digital content,[116] and identifies, tracks, and manages various forms of digital content, including text, images, audio, and video[117] - ISCC generates from the digital content itself	- As ISCC is generated from the content itself, it does not carry proof of ownership of data - Does not manage or verify authoritative metadata, relying instead on external systems for context and

[106] Keller (2024)

[107] Coalition for Content Provenance and Authenticity

[108] Coalition for Content Provenance and Authenticity (2023). C2PA Explainer. Available at https://spec.c2pa.org/specifications/specifications/1.3/explainer/_attachments/Explainer.pdf

[109] Coalition for Content Provenance and Authenticity

[110] *C2PA in ChatGPT Images.* OpenAI. https://help.openai.com/en/articles/8912793-c2pa-in-chatgpt-images

[111] Coalition for Content Provenance and Authenticity (2023)

[112] Coalition for Content Provenance and Authenticity

[113] ISCC – Content Codes. https://iscc.codes/

[114] ISCC – Content Codes

[116] ISCC (n.d.) *About the ISCC*. https://iscc.foundation/iscc/#specification

[117] Pan, T. (2024, May 27*). ISO Publishes New Standard Enabling Content Transparency – ISO 24138.* ISCC. Available at iso-publishes-new-standard-enabling-content-transparency-iso-24138

	content in decentralised and networked environments [115] - Tracks digital assets and verifies data integrity	- Algorithmically bound to digital content[118] - Enables content matching by clustering media files that are similar to each other. Thus, helping with managing copyrights, AI-generated content, and discovering related digital materials.[119]	rights information[120]
Digital Watermarking	- Embeds hidden markers in the content itself, making it an integral part of the digital asset [121]	- It travels with the asset, ensuring traceability[122] - Security features, such as encryption, make tampering detectable, thus protecting integrity[123] - When combined with other unit-based mechanisms, such as C2PA, they improve durability of content credentials, and provenance remains	- It is difficult, but not impossible, to strip or degrade watermarks[125] - AI text detectors can lead to false positives, i.e., they can incorrectly identify human-generated content as AI-generated content,

[115] ISCC – Content Codes

[118] *ISSC – Concept*. ISCC. https://iscc.codes/concept/

[119] Pan, T. (2024, May 27*). ISO Publishes New Standard Enabling Content Transparency – ISO 24138*. ISCC https://iscc.io/iso-publishes-new-standard-enabling-content-transparency-iso-24138/

[120] International Organization for Standardization (2024). *ISO 24138:2024(en) Information and documentation – International Standard Content Code (ISCC)* https://www.iso.org/obp/ui/en/#iso:std:iso:24138:ed-1:v1:en

[121] Digimarc (2023, October 30). *About Digital Watermarks.* Retrieved 22 October, 2025, from https://www.digimarc.com/blog/about-digital-watermarks

[122] Digimarc (2023)

[123] Digimarc (2023)

[125]Madiega, T. (2023, December). Generative AI and watermarking. *European Parliament Research Service.* Available at https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf

			discoverable and verifiable.[124]		impacting reliability[126]
Opt-out registers (e.g., *Do Not Train Registry*)[127]	-	Allows rightsholders to manage permissions at the level of individual content	- Useful for model trainers to know what to exclude from training datasets [128] - Empowers rightsholders to assert their rights	-	Success depends on rightsholders registering their opt-out and model developers actually checking the registries and respecting the opt-out

**For a more comprehensive overview, please refer to Annexure I: Emerging technical standards*

B - Machine-readable expression of intent behind the opt-out

As has been highlighted by Open Future, for a workable model of standards that supports copyright interests, it is necessary that we have "*machine-readable vocabulary to describe the intent of the opt-out*".[129] For instance, a protocol that enables an opt-out from TDM in its entirety would be wide in scope and would also apply to other forms of computational analysis that may otherwise be desirable for the rightsholders. Consequently, Open Future recommends that developing standardised, machine-readable vocabulary that is more granular than binary (i.e. opting out of all TDM or declaring no opt-out) is desirable. For instance, along with a no-tdm option, having a no-generative-ai option could allow rightsholders to specifically opt-out of AI training but still permit TDM for other uses such as AI supported search engine.[130]

Currently, there are very limited ways available to rightsholders to communicate their opt-out which can be all-or-nothing in nature. As we have seen, robots.txt, is very binary. It allows rightsholders to either permit or prohibit TDM in general, without the ability to specify which types of uses they object to. It was also highlighted in the IAB

[124] Collomosse, J. & Guinard, D. (2025, June 2). Digital watermarking for interoperable and durable Content Credentials. Retrieved 22 October 2025 from https://contentauthenticity.org/blog/digital-watermarking-interoperable-durable-content-credentials

[126] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected?. *arXiv preprint arXiv:2303.11156.*

[127] Do Not Train. https://spawning.ai

[128] Keller, P. (2024)

[129] Keller, P. (2024)

[130] Keller, P. (2024)

AI-CONTROL Workshop Report that there are many conflicting and ambiguous vocabularies already in use.[131] This non-standardised approach highlights the convoluted reality of how content is used on the web and how current approaches fall short.

Emerging Developments

- **AI Preferences (AIPREF) Working Group, IETF[132]**

Building on the insights and gaps identified during IAB AI-CONTROL Workshop, the IETF formally chartered a new working group, AIPREF, in 2025. This group is now tasked with developing a standardised, machine-readable vocabulary that allows rightsholders to express granular preferences around how their content may be used by AI systems.[133] Crucially, AIPREF is not just about language, it also seeks to define how these preferences should be attached to content, whether via HTTP headers, well-known URIs, or embedded metadata. The group is also considering how to handle cases where multiple, potentially conflicting, preference signals coexist. While AIPREF does not address enforcement or authentication, as those are separate policy, legal and infrastructure challenges, but its work marks an important shift toward purpose-aware, interoperable signalling in the AI ecosystem.

- **Copyright Framework of Access, Control, Consent, Compensation and Transparency (ACCCT)[134]**

Developed by CoSTAR National Lab, the framework proposes a comprehensive, rights-based infrastructure to support equitable and scalable use of content in the development of generative AI. The focus is the full lifecycle, from provenance and authenticity, and rights expressions through to compensation. It is designed as a set of building blocks and best practices that can guide the development of interoperable

[131] Nottingham & Krishnan (2025)

[132] *AI Preference*. IETF Datatracker. https://datatracker.ietf.org/wg/aipref/about/

[133] Krishnan, S. & Nottingham, M. (2025, February 27). *IETF setting standards for AI preferences*. Internet Engineering Task Force https://www.ietf.org/blog/aipref-wg/

[134] Bennett, J., Collomosse, J., Gregory-Clarke, R., Jones, J., Lycett, M., & Saunders, W. (2025). Time to ACCCT: Providing Creative Industries and AI Developers with a Copyright Framework of Access, Control, Consent, Compensation and Transparency.

tools, legal standards, and business models. It identifies the following functional components necessary for a fair ecosystem:

(1) Asserting unit-level provenance of digital assets using open, patent-free standards that are interoperable across platforms;
(2) Ensuring content identification via durable metadata, watermarking, or fingerprinting;
(3) Identifying ownership and rights holder, potentially via decentralised ledgers to mitigate scalability and governance challenges;
(4) Creating machine-readable licences as a scalable way to issue usage rights to another individual, and allowing granular permissions for generative AI uses;
(5) Issuing machine-readable contracts after a licence is issued, outlining, e.g. terms of use and remuneration;
(6) Ensuring attribution in copyright law, through technologies that trace which data contributed to AI outputs, enabling identification and compensation for rights holders; and
(7) Creating value to compensate the rightsholder for use of the asset based on the license and the contract.

- **Content ARCs (Authenticity, Rights, Compensation)**[135]

Developed by the Centre for the Decentralized Digital Economy (DECaDE), the Content ARCs framework proposes a scalable, interoperable approach for managing rights and compensating creators when their content is used in generative AI training. It combines "*open standards for provenance and dynamic licensing with data attribution, and decentralized technologies*" to enable verifiable asset identification, rights management, and fair compensation. Unlike Digital Rights Management, it is not suggested to be used as a control access mechanism but rather as a legal and technical infrastructure (licenses for content use and value creation). The framework is structured around three interconnected phases that are supported by open standards and decentralised technologies:

(1) **Authenticity:** which establishes the identity and verifies the authenticity of a digital asset in a way that can easily be verified.

[135] Balan, K., Gilbert, A., & Collomosse, J. (2025)

(2) **Rights:** which enables rights holders to claim the ownership of the digital asset they created, as well as to “create digital contracts to license that asset for use by others”.

(3) **Compensation:** through which rightsholders can extract value from their digital assets when a licensee exercises a granted license (e.g. automated ways to determine the level of payment, such as through smart contracts or data attribution technologies).

Section 5. The UK's copyright and AI: consultation 2024-2025

Towards the end of 2024, the UK Government initiated a consultation seeking feedback on its proposal for amendments to the copyright regime in the UK to balance the interests of the creative industry with technological innovation.[136]

In light of the growing challenges that AI has introduced to copyright law, the objectives of the consultation were threefold:[137]

1. Provide control to rightsholders over the use of their content by AI models through licensing and remuneration
2. Provide legal access to high quality data to AI developers
3. Improve transparency of the copyright framework

To achieve these objectives, the government put forth four policy proposals, including a preferred option:[138]

Policy Proposals	
Option 0	Law remains as it is with no changes
Option 1	Require licensing in all areas
Option 2	Introduce a broad data mining exception which will allow training without consent
Option 3	Introduce a TDM exception operationalised through a rights reservation model underpinned by transparency measures

[136] Intellectual Property Office, Department for Science, Innovation & Technology, & Department for Culture, Media & Sport (2024, December 17). *Copyright and AI: Consultation*. Gov.UK. Retrieved July 3, 2025, from https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence

[137] See footnote 1

[138] See footnote 1

The government's preference was Option 3. However, it is to be noted that following the feedback on the consultation from the wider community, priorities of the government might have shifted.[139] After input from the creative industry, recent reports suggest that Option 3 may no longer be the preferred solution and the government may widen the options to determine an appropriate amendment to the copyright regime.[140] It remains to be seen precisely what those options would entail.

Data mining options in commercial and international contexts

In practice, AI developers have generally been able to gain practical access to sufficient training data. The primary question concerns the legal status of using that data. If:

a. an AI training process involved the making of copies, and
b. the content was not licensed for the purpose of AI training, and
c. there was no applicable statutory or common law exception permitting the use of that content for the AI training process in the applicable territory in which the training process took place, then
d. there will be liability for copyright infringement.

For training processes that have already taken place, this leads to the commercial question that is motivating large volumes of litigation; do the AI developers need to pay all of the applicable rightsholders on whose works their models have been trained?

The legal status of the training process is only one part of the copyright infringement analysis. Another important question, which receives limited attention in the UK Government consultation, is the legal status of the pre-trained models themselves. Specifically, should pre-trained AI models be regarded as "copies" of the works on

---

[139] Bambridge, J. (2025, May 22). *UK minister admits 'regret' over AI and copyright row*. Politico. Retrieved July 3, 2025, from https://www.politico.eu/article/uk-minister-admits-regret-over-ai-and-copyright-row/

[140] Courea, E., & Milmo, D. (2025, May 4). Ministers reconsider changes to UK copyright law ahead of vote. *The Guardian*. Retrieved July 3, 2025, from https://www.theguardian.com/technology/2025/may/04/ministers-uk-copyright-artificial-intelligence-parliament-vote

which they have been trained and/or do pre-trained models store or contain copies of those works? If the answer to those questions is "yes", making copies of the pre-trained models, issuing copies of the model to the public, or communicating the model to the public would all be acts requiring a licence from the rightsholder to avoid infringement.

The consequences of this for policymaking are as follows:

- If pre-trained models do not themselves constitute "copies", or are not regarded as "containing copies" of the works in the dataset, a user or vendor of the pre-trained models in the UK would potentially have no obligation to pay the underlying rightsholder and that rightsholder would likely have no legal basis to oppose the commercial exploitation of the pre-trained model in the UK.
- If the UK has a more restrictive regime for AI training than other countries but does not regard the pre-trained model itself as an "infringing copy", AI developers will still have access to the UK market without having to account to any copyright holders for any revenues they generate in the UK. However, some other countries already offer broad exceptions to copyright infringement liability for AI training. If a given training process can be carried out in such a training-favourable jurisdiction without attracting any liability to pay the underlying rightsholder, and then the pre-trained model can be offered in the UK without attracting any liability, no rightsholder is compensated for the use of their work in connection with the AI model.
- Therefore, it is important to consider liability for AI training more broadly rather than in isolation. For example, if UK copyright law is more restrictive than other countries for AI training, but entirely permissive with respect to the commercial exploitation of pre-trained models, such a policy may serve to incentivise developers to carry out AI training overseas, whilst still being able to take advantage of the UK market, which is still the sixth largest economy in the world.[141] In other words, such a policy may harm the domestic AI industry, without any concomitant benefit to rightsholders.

[141] (2025, October 28). Top 20 largest economies in the world in 2025: GDP rankings and key insights. *Forbes India*. https://www.forbesindia.com/article/explainers/top-10-largest-economies-in-the-world/86159/1

- The EU AI Act identified this issue. Whilst the specific way the EU AI Act addresses this issue is very complicated,[142] and it does not actually modify substantive copyright law, the EU AI Act does attempt to provide a "level playing field" for AI model providers who want to access the EU market, without putting EU-based AI developers at a comparative disadvantage.

The analysis of options 0 and 3 (given the current state and the proposed direction of the government proposal) set out below should be understood in the light of the above.

Practical impacts of Option 0

There has been no decision by any UK court as to whether pre-trained AI models themselves do or do not constitute or contain copies of underlying works from their training data sets. Whilst this may ultimately depend on the technical details of each individual pre-trained model, the claimants in the recent *Getty Images v Stability AI* proceedings in the English High Court did not advance any argument at trial that the pre-trained "Stable Diffusion" image generation model itself constituted or contained a copy of any individual works in the training set. This may indicate that the claimants likely did not view the claim as having sufficient merit to pursue it at trial.

The current UK TDM exception is contained in section 29A (1) (a) of the Copyright, Designs and Patents Act 1988. This states that a copy will not be infringing if:

> "*the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose*" (emphasis added)

No UK court has yet issued a judgment on the scope of this exception, and no such cases are known to be currently pending. Even without judicial guidance, it is clear that the UK's TDM exception is narrower in scope than the equivalent exception in

[142] 'Providers' of 'general purpose AI models' in the EU are required to implement a policy to respect the opt-outs from the EU's text and data mining exception, regardless of whether the model was trained in the EU – see Article 53 of the EU AI Act.

many other countries, including the EU. The interpretation of the words “research” and “for a non-commercial purpose” will be important, but this is likely to exclude the development of most commercially available LLMs, regardless of whether they are made available on an open source or quasi-open-source basis without requiring the payment of an up-front fee.

Subject to any judicial decisions to the contrary, Option 0 would potentially imply that:

- The users or vendors of many LLMs and tools incorporating pre-trained LLMs in the UK, if they did not train the model in the UK, or did not train the model at all, may have no obligation to account to any owners of copyright in respect of acts taking place in the UK. Similarly, hybrid stakeholders, who may be using pre-trained models in numerous different configurations, or as part of custom tools built on top of LLMs, may not need to account to the owners of copyright works in an underlying dataset where the technologist has not themselves carried out any AI model training.
- Hybrid stakeholders may be more likely to carry out fine-tuning exercises on existing pre-trained LLMs, which will involve the use of a smaller, more targeted dataset. If the content within that fine-tuning dataset is not licensed for the purposes of fine-tuning, there is a risk that this could give rise to infringement liability.
- Companies wishing to train their own models for commercial purposes would need the permission of the owners of copyright works within the dataset to avoid liability for UK copyright infringement. To do so, they would need to seek licences from individual rightsholders or collective licensing representatives, which would likely involve a commercial negotiation. If they decided to carry out those training processes in a country with a broad TDM exception instead of the UK, they might be able to do so without obtaining licences from the applicable copyright owners yet still be able to exploit their resulting AI models commercially in the UK.

Option 3: TDM exception through a rights reservation model underpinned by transparency

As per the consultation, the proposed TDM exception would permit data mining for all purposes, including commercial ones. If this option becomes law, it would apply only where the user has lawful access, such as through contractually available data so that rightsholders can be remunerated at point of access. The success of this package depends on two key components; rights reservation and transparency.[143]

Rights reservation

A rights-reservation mechanism would entail a standardised machine-readable protocol, such as the REP *(see Section 4: Standardising copyright),* which was the subject of the consultation, that can block web crawlers from harvesting data of rightsholders. As such, under the proposed TDM exception, a license will need to be procured where a rightsholder has specifically opted-out using a standardised protocol to reserve their rights. The legal effect of this would be that:

(a) data can be mined without permission unless the rightsholder has specifically opted-out, i.e. placing the burden of protecting the rights on the rightsholder; and

(b) it would amount to copyright infringement if the developer includes such data in the AI training despite the explicit opt-out.

On the one hand, this approach supports the government's objectives of control and access. By using clear, standardised protocols, rightsholders can retain control over their content. For AI developers, this is beneficial as it allows them to mine any data, unless explicitly reserved, reducing legal risk in accessing high quality data. In theory, this ought to result in a licensing market for high-value content balancing access with creator compensation.

On the other hand, given the current technical infrastructure of machine-readable protocols, as discussed in *Section 4: Standardising copyright*, this approach creates practical challenges in the fulfilment of the government's objectives. Ensuring consistent and widespread adoption of protocols is extremely challenging and this

[143] See footnote 1

could lead to inadvertent use of reserved data if opt-out signals are missed, improperly implemented or not uniformly recognised. As a result, unintentional infringement can complicate compliance, and lack of uniform adoption can create enforcement challenges.

For the rights reservation model to be workable, rightsholders need clarity as to how they should communicate their opt-out. Similarly, AI developers need clarity and certainty as to where they need to check for the existence of an opt-out.

The fact that copyright is an unregistered right in the UK means there is no central register on which a rightsholder could register their opt-out. There are also questions as to whether the requirement for rightsholders to opt-*out* to prevent their works being used in AI training without compensation, rather than having to opt-*in* to permit their works to be used in AI training, could be viewed as a "formality" that rightsholders need to comply with in order to obtain full copyright protection. Article 5 (2) of the Berne Convention for the Protection of Artistic and Literary Works, to which the UK is a signatory, states that the "*enjoyment and exercise of* [copyright] *shall not be subject to any formality*".[144] No court has yet been asked to assess whether the opt-out requirement under EU law is compatible with the Berne Convention, but legal challenges remain a possibility.

Aside from the technical considerations addressed at *Section 4: Standardising copyright*, the diverse range of ways that copyright works are distributed make it difficult to implement a consistent approach to communicating and identifying opt-outs. For example, a photographer may license their photograph to an agency, which then in turn licenses that image to newspapers and magazines, who then post that image online on their own public-facing websites.

In that case, the rightsholder (the photographer) does not control the websites that host the applicable copyrighted work (the photograph). However, an AI developer who scrapes content from the newspaper website may be unaware of the identity of the photographer, or that the operator of the newspaper website does not own the copyright at all or only some of the images and other copyright content hosted on their

[144] Berne Convention for the Protection of Literary and Artistic Works (1886/1887) 828 UNTS 221, art 5 https://www.wipo.int/wipolex/en/text/283698#P109_16834

website. By checking the robots.txt file or terms and conditions of the newspaper website, the AI developer would have no way of knowing whether the photographer or the photography agency had communicated an opt-out on their own respective websites.

Furthermore, the commercial interests of the photographer and the newspaper website may not be aligned. Once the newspaper has paid for a licence to use the image, the photographer typically does not obtain any additional web traffic-based compensation, regardless of whether the website gets 10 hits or 10 million hits. However, the newspaper's advertising revenue typically will be linked to page views. Any “opt-out” protocol which means the newspaper website does not get indexed by the main search engines could lead to a significant loss of web traffic and revenue, which may discourage adoption of such a protocol by newspaper publishers.

In those circumstances, a diligent AI developer who has implemented checks for opt-outs within its web crawling workflow, even very sophisticated checks, would still be faced with uncertainty that an opt-out had been expressed by the rightsholder in a different forum.

Expanding the scope of the TDM exception in Option 3 to protect developers who can demonstrate that they have exercised appropriate diligence in checking for rights reservations would be one way of providing certainty to AI developers, while potentially impacting the interests of rights holders.

Transparency

The TDM exception and rights reservation package also emphasises the need for greater transparency regarding the sources of training material to ensure legal compliance and foster trust between rightsholders and developers. Currently, it is extremely difficult to determine if copyright works are used in AI training due to limited or no disclosure from AI developers. In the consultation, the government supported the argument that provenance of training data is key to enforcing rights, understanding legal risks, and encouraging AI adoption. However, at the same time, it recognised the practical challenges associated with enabling transparency in extremely large datasets, especially for small enterprises and new entrants. As such, the consultation’s text suggested that this area is still under exploration, and any

transparency measures would need to have a balanced approach and be both proportionate and justified.

Some of the earliest court proceedings that were commenced against AI developers related to models that were known to be trained on open-source databases, which allowed rightsholders to confirm whether their works had been included in the training dataset. More recently, AI developers have tended to keep the details of their training datasets confidential. This is partly motivated by the desire to maintain their competitive edge, with details of training processes potentially being valuable trade secrets, but also partly informed by the potential for copyright infringement claims from the owners of copyright works within the training corpus. There have been several proposed AI transparency laws internationally. The EU's transparency regime, in particular, is addressed below.

Overlaps with the EU's copyright regime

The UK's proposed approach is largely inspired by the EU's copyright framework. As such, the government's proposal has various overlaps with the EU's approach, which is also evolving.

(a) Article 4 of the Digital Single Market Copyright Directive (Directive (EU) 2019/790)[145]

This provision introduces a TDM exception in the EU and has become very relevant for AI.

Article 4
Digital Single Market Copyright Directive

1.	**Scope of the TDM**	reproductions and extractions of lawfully accessible works are allowed
2.	**Retention of works**	for as long as it is necessary for the purposes of TDM

[145] (2019, April 17). *DIRECTIVE (EU) 2019/790*. Official Journal of the European Union. Retrieved July 3, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790

3.	**Rights reservation**	the TDM exception will not apply where a rightsholder has explicitly opted out, such as by using a machine-readable protocol
4.	**Uses**	where no explicit rights reservation has been made, the exception will also cover commercial uses

As is evident, the UK government's proposal is very much in line with the EU's Copyright Directive. However, as also recognised in the consultation, under the EU regime, it is not entirely clear what constitutes a valid rights-reservation model. Although the UK government's aim is to standardise such protocols as much as possible, as discussed in *Section 4: Standardising copyright*, there will be many practical challenges associated with doing so.

(b) Article 53 (1) (d) of the EU AI Act[146]

As mentioned earlier, the success of the UK's proposal will depend on effective transparency measures. This component of the proposal is also inspired by the EU AI Act.

Under Article 53 (1) (d) of the EU AI Act, providers[147] of general-purpose AI models (GPAI) are under an obligation to report the sources of their training data in a publicly available and sufficiently detailed summary. In advance of the obligations for GPAI models, which took effect under the EU AI Act 2 August 2025, the European Commission released an Explanatory Note and Template for the Public Summary of Training Content for GPAI models on 24 July 2025.[148]

---

[146] (2024, June 13). *REGULATION (EU) 2024/1689*. Official Journal of the European Union. Retrieved July 3, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

[147] Defined as "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge"

[148] European Commission (2025, July 24). Exploratory Notice and Template for the Public Summary of Training Content for general-purpose AI models. *European Commission*. Retrieved July 3, 2025, from https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

Article 53 (1) (d), read with Recital 107, reveals that the purpose for the summary is to promote transparency on training content for GPAI models that is protected by law and to enable parties with legitimate interests, including rightsholders, to enforce their rights under Union law. Although the Summary is meant to be comprehensive, however, there is no requirement for it to be technically detailed.

Whilst the full details of the template are beyond the scope of this chapter, we note that the template requires the "providers" of general-purpose AI models to make the following information publicly available:

1. **General information**: Basic information to identify the model and the "modalities" (i.e. text, image, audio and video) present in the training data.
2. **List of data sources**: This requires disclosure of the main datasets used to train the model, including:
    a. any large publicly available databases, defined as those "made available publicly for free" (for example Common Crawl);
    b. any private databases (datasets for which licences have been concluded with the rightsholders and datasets obtained from data intermediaries, where the rightsholders have not granted a licence),
    c. a "comprehensive narrative description" of the data scraped online, including a general description of the crawler(s) used, the date range over which data was collected, and a summary of the most relevant domain names scraped. SMEs need to disclose the lower of the top 5% of all domain names scraped and the top 1,000 domains, whilst larger companies must disclosure the top 10% of all domain names scraped, determined by reference to the size of the content scraped;
    d. information about user data used to train the model; and
    e. details of any synthetic data used to train the model.
3. **Data processing aspects**: The provider needs to describe the measures taken to identify and "respect" rightsholder opt-outs and to remove illegal content (such as child sexual abuse and terrorist content).

The template does not require model providers to list individual copyright works item-by-item, although rightsholders may be able to ascertain that information indirectly from the producers of any crawled datasets used.

(c) General Purpose AI Code of Practice[149]

Compliance with the EU AI Act will, by and large, depend on voluntary compliance with harmonised standards currently being developed by Cen-Cenelec.[150] Before harmonised standards are published, according to Article 53 (4), compliance with Article 53 (1) (d) can be demonstrated by complying with codes of practice in the interim. To this end, a final version of the GPAI Code of Practice was drafted and published in July 2025. It is now a standard, but voluntary way for providers of GPAI models to comply with their legal obligations under the EU AI Act.[151]

The GPAI Code of Practice reinforces the requirements of Article 4 of the EU's Copyright Directive. At the same time, while encouraging widespread adoption of robots.txt, it is still not clear what will be a valid rights-reservation model as Measure I.2.3, section 1 (b) also requires that best efforts be made to identify other relevant and appropriate protocols that may block web crawling. Moreover, commitments under Measure I.2.3 do not affect the rightsholders' ability to reserve their content by any other appropriate means and emphasises the importance of collaboration in the development of technical standards. While this approach aims to strike a balance between the two stakeholders, it may also contribute to increased fragmentation in the rights-reservation landscape.

Practical impacts of Option 3

Given the unregistered nature of UK copyright, and the limitations imposed for the same under international treaties, implementing Option 3 in a manner that gives

[149] European Commission (2025, July 10). *The General-Purpose AI Code of Practice*. Retrieved July 25, 2025, from https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[150] (n.d.). *Harmonised Standards*. European Commission. Retrieved July 3, 2025, from https://single-market-economy.ec.europa.eu/single-market/goods/european-standards/harmonised-standards_en

[151] Leone de Castris, A. (2025). AI Governance around the world: European Union. *The Alan Turing Institute.* https://doi.org/10.5281/zenodo.17054419

certainty both to AI developers and rightsholders may present many practical challenges.

As previously noted, the absence of any central registry for rights-reservations may create uncertainty for AI developers. Even where they exercise diligence in checking for rights reservations online, checking a combination of the robots.txt file and even analysing the wording of website terms and conditions for any reservations implemented in natural language, they may still face residual risk if rights reservation has been made through channels other than the website(s) hosting training content.

Copyright compliance will be only one of several factors that AI developers consider when deciding where to carry out their AI training. Smaller developers may only have staff and computing resource in the UK, and they may not be in a position to pick and choose where to invest in new personnel or computing resource.

For AI developers with the flexibility to choose between jurisdictions to conduct model training, the residual uncertainty that remains even after looking for online rights-reservations may make the UK a comparatively less attractive jurisdiction than those offering broad, unconditional TDM exceptions. If the models trained in a more permissive jurisdiction can still be exploited in the UK without needing to have complied with a UK opt-out framework during training (i.e. without implementing EU-style “level playing field” obligations), the intended balance between supporting AI development and protecting rightsholders' interests may be difficult to achieve.

Section 6. Hybrid use cases

This section focuses on *hybrid* use cases that are real-world examples involving hybrid stakeholders who contribute to the development, support, or delivery of hybrid creative outcomes, for instance, blending digital and/or AI tools to produce new forms of creative and artistic works.

The use cases have been included to demonstrate the evolving role of actors within the AI and digital ecosystem, who often wear multiple hats, with the aim to encourage dialogue and policy reforms that recognise the diversity of stakeholder experiences within the UK's copyright framework.

We recommend reading this chapter alongside Section 2: *Rationale for copyright law and the impact of AI;* Section 3: *Copyright, Designs, and Patents Act 1988; and* Section 4: *Standardising copyright.*

Who is a hybrid stakeholder?

"An individual or entity operating across multiple domains (such as technology, creativity, art, design, research, or production), who contributes to the development, support, or delivery of hybrid creative outcomes. Hybridity exists on a spectrum and can cover creative practitioners and technologists who blend digital and / or AI tools to produce new forms of creative and artistic works, as well as involve facilitators, data stewards, or service providers, for example. Hybrid stakeholders reflect the fluid and interdisciplinary nature of contemporary creative and technological ecosystems in a technology driven world."

Hybrid stakeholders can cover one or more roles, such as, but not exclusively:

Creative practitioner. A practitioner who applies creative skills and interdisciplinary approaches in their professions leading to creative expression, outputs and/or outcomes, such as artists, writers, musicians, performers, producers, educators.

Data Steward. An individual or entity with oversight and data governance role responsible for managing data assets on behalf of an organisation or others to ensure appropriate access, quality, security, and compliance.

Facilitator. A professional or entity that facilitates across stakeholders the provision of expert advice, consultancy, or forms of stakeholder engagement.

Researcher. A person or group who conduct research in a subject field, especially applying methods or processes in order to create and/or discover new knowledge or to reach a new understanding of a research problem, question, or phenomenon.

Technologist. A professional who focuses on the practical application of technologies to solve problems, design, build, and maintain AI systems and software applications.

Service provider. An individual vendor or business that provides solutions and/or services to end users and organisations, such as media hosting, technology and data services.

Real world examples of hybrid roles are highlighted in the uses cases provided in this section.

Setting the Scene

According to Band and Gerafi[152], there are more than 40 countries which have fair dealing or fair use provisions in their copyright laws. In this mix, there are already jurisdictional examples of hybrid fair dealing which combines the open-ended framework of fair dealing and the closed-ended approach of enumeration, an approach which is used in some jurisdictions to balance copyright protection with user rights, such as "research" and "educational" use, where it allows for specific exceptions[153].

[152] Band, J. and Gerafi, J. (2024). *The Fair Use/Fair Dealing Handbook.* Joint PIJIP/TLS Research Paper Series. Available at: https://digitalcommons.wcl.american.edu/research/141

[153] Band, J. and Gerafi, J. (2024); Rosati, E. (2025). Copyright Exceptions and Fair Use Defences for AI Training Done for "Research" and "Learning," or the Inescapable Licensing Horizon. *European Journal of Risk Regulation*: 1-24. doi:10.1017/err.2025.10035; Zhang, S. (2024). Comparing Fair Dealing with Fair Use: Why Fair Dealing Can Better Balance Copyright Interests? *Proceedings of the 3rd International Conference on Business and Policy Studies*. doi.org/10.54254/2754-1169/75/20241717

In the general application of AI, especially in terms of copyright, issues have revolved around the processes that lead up to the generation of AI outputs, namely TDM[154]. Central to the TDM processes, particularly unauthorised and non-consent TDM, there are implications for content creation and/or re-use and the potential rights of rightsholders which can be in direct tension[155] (see *Section 3: Copyright, Designs, and Patents Act 1988*).

This is especially critical for those who have hybrid or dual professional roles, e.g. researcher and creative technologist. Such roles rely on pursuing research and gathering and re-using data for both research and creative outputs with autonomy[156].

The use of generative AI has been particularly divisive within hybrid artistic and creative communities, focusing on AI models being trained on artists' copyrighted work without permission[157] and concerns about AI displacing creative stakeholders, such as highlighted in the SAG-AFTRA strike [158] but also AI providing creative possibilities within culture and the wider arts, for instance[159].

Legal action[160] against technology companies has further drawn attention to hybrid creatives' copyrights and how their online artwork is collected, reused, and remixed. Copyright-protected works are considered desirable as training data by larger AI

---

[154] OECD (2025). Intellectual property issues in artificial intelligence trained on scraped data. *OECD Artificial Intelligence Papers*, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/d5241a23-en; Tyagi, K. (2024). Copyright, text & data mining and the innovation dimension of generative AI. *Journal of Intellectual Property Law & Practice, 19*(7), 557-570. doi.org/10.1093/jiplp/jpae028

[155] Hutson, J. (2024). The Evolving Role of Copyright Law in the Age of AI-Generated Works. *Journal of Digital Technologies and Law, 2*(4), 886-914; Tyagi, K. (2024).

[156] Rosati, E. (2025)

[157] Appel, G., Neelbauer, J. Schweidel, D. A. (2023) Generative AI Has an Intellectual Property Problem. *Harvard Business Review*. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

[158] Feingold, S. (2024) AI and Hollywood: 5 questions for SAG-AFTRA's chief negotiator. *World Economic Forum*. https://www.weforum.org/stories/2024/03/ai-hollywood-strike-sag-aftra-technology/

[159] Miltner, K., Highfield, T. (2024). The Possibilities of "Good" Generative AI in the Cultural and Creative Industries. *The British Academy*. doi.org/10.5871/digital- society/9780856726989.001

[160] Chayke, K. (2023). Is A.I. Art Stealing from Artists? *New Yorker.* https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists

technology companies, such as OpenAI, as these data ensure more optimal development of AI models[161].

The innovation paradigm in hybrid scenarios equally highlights that copyright can be seen as a particular barrier[162]; yet the purpose of copyright to help catalyse creative expression and innovation is key to hybrid considerations[163]. In an AI context, additional considerations need to be made in relation to the "hybrid" ecosystem and supply chain in which various research and creative communities and/or individuals may rely, including cross-jurisdictional digital infrastructures, tools and applications which may be used to scrape, capture, share, store, or archive digital media, among other processes[164]. Some of these AI related processes are outlined in *Section 2: Rationale for copyright law and the impact of AI* and *Section 3: Copyright, Designs, and Patents Act 1988* of the report in relation to existing copyright implications.

Hybrid Use Cases

The hybrid use case examples below are not intended to be inclusive, but to provide a spectrum of real-world scenarios across UK and international jurisdictions, highlighting dependencies, current and potential impacts, and suggested ways to achieve an equitable balance in protecting human creative agency through copyright and related safeguards.

Each case study involves different stakeholder roles and hybrid settings as outlined in the following table:

[161] OpenAI—written evidence (LLM0113) House of Lords Communications and Digital Select Committee inquiry: Large language models.
https://committees.parliament.uk/writtenevidence/126981/pdf/

[162] Reid, A. (2019). Copyright Policy as Catalyst and Barrier to Innovation and Free Expression. *Cath. U. L. Rev., 68,* 33.
Rosati, E. (2019). Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity. *Asia Pacific Law Review 27*(2), 198-217.

[163] Hutson, J. (2024); Tyagi, K. (2024).

[164] OECD (2025).

List of Case Studies

Use case identifier	Short description	Hybrid stakeholder (roles)
Case Study 1: Mary Stewart-David, York University	A creative practitioner/ researcher creating and re-using immersive and AI generated media for creative outcomes	Creative practitioner Researcher Technologist
Case Study 2: Oxford University Business Studies	Researcher's use of commercial data for research and education purposes creating new models and synthetic data which could be seen as substitutional for the commercial product	Data Steward Researcher Technologist
Case Study 3: Oxford University Visual Geometry Group	AI research group generating and scraping copyrighted and open data for research purposes	Data Steward Researcher Technologist
Case Study 4: Refik Anadol, Echoes of the Earth: Living Archive	Generative AI art exhibition and large language model created using large-scale multi-jurisdictional public datasets	Creative Practitioner Technologist
Case Study 5: Uppbeat (uppbeat.io)	Music licensing platform and generative AI playlist for musicians and online content creators	Creative Practitioner Data Steward Service Provider
Case Study 6: GLAM-E Lab TDM Report (glamelab.org)	Monitoring and resource management of large-scale web scraping of licensed digital cultural collections for AI training	Data Steward Facilitator Service Provider

Note on use case compilation: Use cases 1 to 3 have been provided directly by participating co-authors MSD (1), CM and RW (2-3), and use cases 4 to 6 have been compiled through desktop research and public sources by co-author AB.

Case Study 1: Mary Stewart-David, York University (Creative practitioner and researcher in immersive media)

1) Context

As an immersive theatre maker and creative technologist, I practice deception on a daily basis. Working in Virtual Reality (VR), I make musicals in the metaverse, performed by digital actors on a computer-generated replica stage. Our cast of colourful avatars, digital twins of their human counterparts, are crafted by capturing the face, body, movement, and voice of real actors through a process that, in another context, might be considered deep fakery. These digital doppelgängers are clones of human actors; avatars who sing, dance and act in our interactive and immersive music theatre productions, performed live in VR.

Who is involved?

Our shows are written and produced by a team of virtual theatre-makers based in the UK and the US, collaborating and performing online in real time. Audiences can join the show from anywhere in the world with internet access.

Jurisdiction(s)

While our production company is UK-based, our place of business is the metaverse. In the US, the SAG-AFTRA strikes led to agreements on the use of actors' likenesses in film and games [165]. In the UK, entertainment unions have been slower to act. Independent producers, who make up most of our sector, are not required to work under union rules. This raises questions about working standards, and ethical treatment of actors and creatives, and the concern that the imposition of greater regulation in the UK might drive virtual production companies such as ours, to relocate to more favourable domains.

How does this use case reflect hybridity?

The use case focuses on a creative practitioner who has a blended role as a technologist and researcher, working across self-authored creative outputs, copyrighted media and technology platforms in the UK and the US.

[165] Feingold, S. (2024)

Why is this working example important for hybrid practitioners?

Like many creatives in the UK, I am concerned by the government's proposal to remove copyright restrictions on TDM for LLMs. My work has long been vulnerable to copyright theft, especially in jurisdictions where enforcement is weak. Large organisations can defend their IP; small independent makers often cannot afford the time or money to do so.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

As artists, our power to control our output is limited in the age of AI. But there are steps we can take to protect our brand identities. Adding content credentials to our work with tools like the open source C2PA protocol[166] can establish data provenance and makes tracking and licensing our work easier and potentially more effective (see *Section 4: Standardising copyright).* We can also ensure that we only use ethical AI models that similarly watermark their output and transparently license their training data. Indemnity insurance, and collective licensing agreements[167], could be of use here, and may well become the norm, while pay-to-scrape models could offer a new source of income.

[166] Coalition for Content Provenance and Authenticity (C2PA) https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html

[167] The Copyright Licensing Agency (CLA), a UK not-for-profit, is developing a Generative AI (GenAI) Training Licence, and provides professional guidance on GenAI and copyright contexts. See: https://cla.co.uk/ai-and-copyright/

Case Study 2: Oxford University Business Studies (Research use of commercial data)

1) Context

This case study focuses on model and synthetic data derived from proprietary commercial datasets. It arises from a research project being undertaken by an Oxford based researcher who asked for support from the Bodleian Libraries' copyright support service.

Who is involved?

Oxford researcher working with supply chain data provided by commercial 'firm level data' provider.

Jurisdiction(s)

Research is taking place in the UK. The contract for the commercial data is governed by the laws of England and Wales. However, the researcher would like to share the outputs of the research globally and openly.

How does this use case reflect hybridity?

This use case intersects multiple domains, including academic research and data science. Individuals are hybrid practitioners that simultaneously have roles of researchers, technologists, and data stewards.

Why is this working example important for hybrid practitioners?

It highlights the challenges in working with proprietary datasets where new insights drawn from academic analysis lead to the creation of models and synthetic data which could be seen as substitutional for the commercial product.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

These types of scenarios require a risk assessment by the researcher who in turn needs support on the legal and ethical implications of completing their research and releasing their findings, data and models. The most challenging aspect is that contractual agreements explicitly restrict activities that may be permitted under law. In addition, there is ambiguity in what type of intellectual property protection applies to the dataset under UK law which flows through to lack of clarity on what legal defences are available in the event of challenge.

2) Other challenges and opportunities

Socio-technical

There are clear legal restrictions on the creation of a tool that could substitute use of a commercial product. But is the purpose of academic research to protect existing business models, or is market disruption an inevitable outcome of technological developments that may have significant societal benefits if made available to a broader number of stakeholders? To what extent should this be a decision for individual researchers or is there a case for codifying research practices through co-production with various stakeholders to provide some level of assurance and protection?

Also, there are more questions than answers - to what extent does the law (particularly copyright and contract law) support innovative research? Does it need to be amended to enable potentially disruptive outputs and if it was, what would the implications be for commercial information providers? Do we need to rethink the traditional idea/expression distinction in copyright law and are current legal developments providing protection for databases and data more broadly going to meet the needs of researchers and society as whole? Development of codes of fair practice in cutting edge research could unlock latent flexibility in the law and identify relatively uncontested centre-ground positions through iterative stakeholder dialogue[168].

[168] For examples where this has been done under United States fair use, see *Codes of Best Practices.* Center for Media & Social Impact. https://cmsimpact.org/report-list/codes/

Case Study 3: Oxford University Visual Geometry Group (Research group using open data practices)

1) Context

The Oxford Visual Geometry Group (VGG) is a leading research group in AI and computer vision at the University of Oxford. Their datasets and models have become global benchmarks in academic and industry research. However, legal uncertainty around UK copyright law—particularly Section 29A—(see *Section 3: Copyright, Designs, and Patents Act 1988*) has led to the removal of key datasets and restrictions on data sharing, impacting their ability to conduct and disseminate open, high-impact research.

Who is involved?

VGG, based in the Department of Engineering in the University of Oxford.

Jurisdiction(s)

Research is undertaken in the UK, with data gathered from various global locations via scraping and other means.

How does this use case reflect hybridity?

VGG operates at the intersection of multiple domains, including academic research, AI and computer vision technology, and open data practices. The group simultaneously act as researchers, technologist, data stewards, and service providers, navigating between knowledge production and knowledge dissemination.

Why is this working example important for hybrid practitioners?

This case highlights the legal and ethical tensions that hybrid practitioners—those navigating academic research, open data practices, and AI development—face under current copyright frameworks. VGG is internationally influential, and their challenges reflect systemic issues for innovation in regulated environments. Widely cited and used training datasets have had to be withdrawn from circulation despite the clear beneficial effects of their previous availability.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

Researchers must restrict or halt activities like sharing datasets or publishing trained models due to legal uncertainty. Research impact is lessened as a result, as is knowledge exchange.

The most important issues for practitioners include legal uncertainty around copyright, which creates risk and confusion, especially when sharing data or collaborating. This has led to the removal of widely used datasets and a chilling effect on innovation. Researchers face high legal costs and institutional pressure to avoid risk, all of which hinder open, impactful AI research.

2) Other challenges and opportunities

Socio-technical

Ethical research demands transparency and openness, yet copyright risks can compel secrecy and data hoarding. In addition, there is a moral imperative to enable assistive technologies (e.g., SynthText [169] aiding the visually impaired), which is hampered by data access issues.

Another key challenge is that non-commercial TDM is a vague concept, especially around defining "non-commercial" in academic-industry collaborations. It also restricts the sharing of datasets. Legal reviews could be useful to support how the group navigates legal uncertainties, but they are expensive and still don't ensure freedom from liability. As a result, VGG was forced to take down datasets due to ambiguous copyright status—even when they were widely used and cited in research. These restrictions in sharing extend to research outputs. For example, where AI models trained on copyright protected data. This undermines reproducibility, one of the principles for scientific data management and stewardship[170], and, in turn, scientific progress.

[169] SynthText is a synthetically generated dataset. Word instances are placed in natural scene images and consider the scene layout. See Gupta, A., Vedaldi, A. & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 2315-2324.

[170] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1). Also UK Research and Innovation (UKRI) have published guidelines in responsible research, such as guidance in best practice management of research data. See: https://www.ukri.org/publications/guidance-on-best-practice-in-the-management-of-research-data/

Technological measures, such as dataset filtering and content licensing (e.g., Creative Commons licensed images) have been proposed as possible solutions. However, they are insufficient replacements for broader, established datasets. The move toward “zero-risk” research environments limits innovation and excludes many institutions from cutting-edge work.

Case Study 4: Refik Anadol's Echoes of the Earth: Living Archive (Generative AI exhibition and large language model)

1) Context

Refik Anadol[171] is a Turkish-American media artist, formerly an artist in residence at Google. One of his first projects was using an open-source library in Turkey to create the Archive Dreaming project[172]. It was the first AI artwork to use public data to create art. Recently Anadol, working collaboratively with a team, developed an open-source Large Nature Model, which was behind the "Echoes of the Earth: Living Archive" exhibition at London's Serpentine Gallery (16 February - 7 April 2024).[173]

According to Anadol, his Large Nature Model is the first open-source generative AI model focused on nature images and sounds and took less than a year to complete. For the research and production, Anadol and team of workers from 10 countries speaking 15 languages harnessed more than half a billion open-source data images from cultural organisations such as the Smithsonian Institution, National Geographic and London's Natural History Museum, as well as from various rainforests around the world. [174]

Who is involved?

Anadol works with a team of people with diverse backgrounds, including architects, designers, AI engineers, data scientists, musicians, philosophers, neuroscientists among others. For some works, Anadol collaborated with institutions such as the NASA Jet Propulsion Lab (California) or MoMA (New York) to use their datasets. Zaha

[171] Refik Anadol. https://refikanadol.com/

[172] *Archive Dreaming — AI Data Sculpture.* Refik Anadol. Available at https://refikanadol.com/works/archive-dreaming/

[173] See *Refik Anadol: Echoes of the Earth: Living Archive.* Serpentine Galleries. https://www.serpentinegalleries.org/whats-on/refik-anadol-echoes-of-the-earth-living-archive/

[174] Schrader, A. (2024, January 16). Refik Anadol Launches the World's First Open-Source A.I. Model Dedicated to Nature. *Artnet*. https://news.artnet.com/art-world/refik-anadol-living-archive-nature-2419482

Hadid Architects (Seoul) and Casa Batlló (Barcelona) have also shared data. Most often, Anadol uses public data[175].

Jurisdiction(s)

The model was trained on data from National Geographic, the Smithsonian Institute, Cornell Lab, the Natural History Museum in London, and the Conservation Research Foundation Museum, as well as data Anadol's team has personally collected at locations around the world, including deep in the Amazon rainforest. Operating across multiple jurisdictions, the project integrated different methods such as LiDAR and photogrammetry, and captured ambisonic audio and high-resolution visuals of diverse ecosystems for the model.

How does this use case reflect hybridity?

This use case blends art, science, and technology. Anadol and his multidisciplinary team operate at the intersection of creative practice, technological innovation, data science, and environmental research.

Why is this working example important for hybrid practitioners?

Refik Anadol is a creative practitioner and technologist working across large public datasets in different jurisdictions to build a generative AI model to produce new creative outputs.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

Anadol expressed that as an educator at UCLA's Department of Design Media Arts, he values sharing knowledge to support others, but he has concerns that open sharing often leads to his work being copied without proper references or permission.

[175] Pillay, T (2025, February 4). Artist Refik Anadol Uses AI to Turn Data Into Dreams. *Time*. https://time.com/collection/time100-impact-awards/7212503/refik-anadol-ai-time-impact-award/

This, coupled with increasing value of his work among collectors, urges him to protect his work: "Without protection it becomes a free-for-all and nobody moves forward"[176].

Another important issue is understanding and managing data inputs in the creative process: from curating datasets to ensure alignment between the output and the artistic vision, and respect for IP rights. Similarly, having clarity on the data sources and obtaining the necessary permissions are crucial to protect both his own work and build trust with his audience and collaborators[177].

2) Other challenges and opportunities

Ethical

Anadol does not use personal human data in any of his creations[178]. In addition, he feels responsibility not to damage the environment further with the development of his AI where technologies are known to take immense computing power, using large amounts of energy and requiring the mining of rare earth metals[179]. "We don't want to damage nature while making a nature model", Anadol said, "[w]e worked with Google engineers, so we can now exactly see much impact to the environment our model has when it is running and can use only renewable energy when training A.I. model. So, we can do this extremely sustainable, which is possible."[180]

Socio-technical

Encouraging artists to engage with the broader community. Sharing knowledge and best practices can help to navigate the evolving world of generative AI and maintain the integrity and originality of the artworks.

---

[176] Nurton, J. (2024, December 1). 'Painting' with data: how media artist Refik Anadol creates art using generative AI. *WIPO Magazine*. Retrieved 30 October, 2025, from https://www.wipo.int/en/web/wipo-magazine/articles/painting-with-data-how-media-artist-refik-anadol-creates-art-using-generative-ai-67301

[177] Nurton (2024)

[178] Nurton (2024)

[179] Valdivia, A. (2024). The Supply Chain Capitalism of AI: A Call to (Re) think Algorithmic Harms and Resistance. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 1466-1466).

[180] Schrader (2024).

Case Study 5: Uppbeat (Music licensing platform for online content creators)

1) Context

Uppbeat[181] is a UK-based startup that provides a royalty-free[182] music licensing platform for online content creators, such as YouTubers, podcasters, and social media users. It offers a free music resource for the creator space, and to avoid copyright claims on music used within creator content. The platform claims that it offers a quality alternative to expensive music licensing platforms and free music options such as YouTube's Audio Library or Creative Commons (CC) music[183]. The service works on a *freemium* model, providing a basic version of the service for free, and premium features on subscription. The paid premium service removes download limits and "whitelists"[184] channels to prevent copyright claims on monetised videos.

Recently, the Uppbeat platform introduced a text-music playlist generator using ChatGPT technology. The Uppbeat AI Playlist Generator[185] helps content creators find copyright-free music by entering text prompts or descriptions of a scene, sound reference, or desired vibe. This generates a tailored playlist of tracks only from Uppbeat's extensive catalogue; it does not make AI-generated music.

Who is involved?

Uppbeat was developed by Lewis Foster and Matt Russell, the UK-based co-founders of music-licensing company, Music Vine[186], launched in January 2021.

---

[181] Upbeat. https://uppbeat.io/

[182] Royalty-free refers to material subject to copyright or other IP rights which may be used without the need to pay royalties or license fees for each use, per each copy or volume sold or some time period of use, for instance. https://en.wikipedia.org/wiki/Royalty-free

[183] Creative Commons (n.d.). *Legal Music for Videos*. https://creativecommons.org/legalmusicforvideos/

[184] Whitelisting is a security protocol that lists pre-approved entities, such as email addresses, websites, or applications, that are granted access to a system or network.

[185] Uppbeat. *AI playlist generator.* https://uppbeat.io/ai-playlist-generator

[186] Music Vine. https://musicvine.com/

Jurisdiction(s)

Uppbeat is a UK-based startup located in Leeds.

How does this use case reflect hybridity?

This use case profiles a service provider with a hybrid role as a data steward, hosting digital music tracks produced by creative practitioners, handling the licensing, and providing access on a dedicated “freemium” platform to these tracks for creative re-use with full copyright clearance.

Why is this working example important for hybrid practitioners?

It highlights the efforts to integrate generative-driven functionalities with transparent music licensing and fair revenue-sharing models. Uppbeat proposes an approach for hybrid ecosystems simultaneously supporting equitable access to creative resources, data and rights management, and sustainable creative economies.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

Musicians/artists retain the rights to their works while collaborating with the platform, and their earnings on Uppbeat are generated through a shared revenue model. Through this approach, the platform handles the master and sync licenses directly through user agreements. When a user downloads a song, they are granted permission to use it in their online content. In addition, to prevent copyright claims, free users are required to include a unique "Uppbeat Credit" in their video description, while paid users can whitelist their YouTube channels using the “channel safelisting” feature. Musicians on Uppbeat are paid through royalties generated from their music being licensed, which includes revenue from both free and premium users. The platform uses a system where revenue is pooled and shared among rights holders based on their usage share, with artists receiving a percentage of the total revenue earned from their tracks. Uppbeat’s licenses also allow users to use the music in videos shared across platforms such as YouTube, TikTok and Instagram, and other open distribution platforms. These types of services allow any artist to release their music to a wide range of online streaming services for a fee. However, users need a

different license if their content appears anywhere else, such as a live event or selling physical copies of a video.[187]

2) Other challenges and opportunities

Socio-technical

Uppbeat's AI playlist generator provides royalty-free music specifically designed to help creators avoid copyright infringement. The platform does not specify if users' prompts are subsequently saved and used in training and/or refining their AI playlist generator model.

In regard to copyright claims, the platform works by associating unique credits with each download, which helps platforms like YouTube proactively identify and prevent copyright claims. While the platform is designed to prevent claims, users are still advised to ensure correct crediting. Typically, copyright claims occur if a user doesn't *whitelist* their YouTube channel with Uppbeat, for example.

[187] Foster, L. (2024, July 3). *How Uppbeat's music licenses work.* Uppbeat. https://uppbeat.io/blog/royalty-free-and-copyright-free-music/uppbeats-music-licenses

Case Study 6: GLAM-E Lab (Web scraping of licensed digital cultural collections)

1) Context

GLAM-E Lab[188] is a joint initiative between the Centre for Science, Culture and the Law at the University of Exeter and the Engelberg Center on Innovation Law & Policy at NYU Law. The Lab studies issues affecting Galleries, Libraries, Archives, and Museums (GLAMs) and provides legal counsel to GLAM institutions and cultural organisations as they develop open access programs.

One of the latest initiatives of GLAM-E Lab consisted in capturing the impact that bots building datasets for AI model training are having on online cultural collections. The findings were published in the report *Are AI Bots Knocking Cultural Heritage Offline?*[189]. According to the report, thirty-nine surveyed institutions with online collections reported a recent increase in traffic in the past year, and 27 of these had identified AI bots as the cause (several others suspected bots were responsible, although they were uncertain). Several respondents reported a disruption in service to users as a result of the activity. The report indicates that cultural institutions are concerned by the aggressive harvesting of their content, and burden on institutional resources that large data-harvesting places on websites and online collections databases.[190]

[188] GLAM-E Lab. https://www.glamelab.org/

[189] Weinberg, M. (2025, June). Are AI Bots Knocking Cultural Heritage Offline? *Engelberg Center on Innovation Law & Policy*. https://www.glamelab.org/products/are-ai-bots-knocking-cultural-heritage-offline/

[190] The GLAM-E Lab survey results resonate with reported AI bot issues by the Confederation of Open Access Repositories based on the responses of 66 open access repositories run by libraries, universities, and other institutions (see, e.g., Shearer, K. & Walk, P. (2025, June 3). Open repositories are being profoundly impacted by AI bots and other crawlers: Report from a COAR Survey. *Confederation of Open Access Repositories*). Similar aggressive bot behaviour has been raised by, for example, The Wikimedia Foundation (see Mueller, B., Danis, C. & Lavagetto, G. (2025, April 1). How crawlers impact the operations of the Wikimedia projects. *Wikimedia.*) and the code documentation commons project ReadTheDocs (see https://about.readthedocs.com/).

Who is involved?

In addition to the GLAM-E Lab team, the project involved an anonymised survey of 43 organisations mainly from Europe, North America, and Oceania, with follow-up interviews from volunteer representatives.

Jurisdiction(s)

GLAM-E Lab is hosted at the University of Exeter and is a joint initiative with Engelberg Center on Innovation Law & Policy at NYU Law. The report captured responses from participants from Europe, North America, and Oceania. However, the audience for their initiative is potentially world-wide.

How does this use case reflect hybridity?

This use case profiles a hybrid service provider and facilitator in the creative practitioner and data steward community, providing legal counsel and collectively bringing together stakeholders on an issue of shared copyright and related rights concern; namely AI scraping bots ignoring data licenses.

Why is this working example important for hybrid practitioners?

Online cultural collections tend to offer easily accessible high-quality, well-structured and machine-readable images and research data. These collections are therefore highly desirable for training AI models. However, as highlighted by GLAM-E Lab in their report *Are AI Bots Knocking Cultural Heritage Offline?,* AI scraping bots largely ignore robots.txt (see *Section 4: Standardising copyright)* in the process of data harvesting across these online collections and can create problems with analytics and in ignoring different types of licenses that affect hybrid practitioners.

What are the most important issues from the perspective of the practitioner or other stakeholder(s)?

According to the report, survey respondents reported deploying, or increasing the use of, third-party services that offered the capacity to counter bots and the ability to track bots by the originating source (e.g., Google bots, Anthropic bots, Apple bots) and to provide more accurate analytics of bot activity more broadly. The overall findings highlight that responding to aggressive bot behaviour is complex and resource-intensive. Simple fixes, such as blocking IP addresses, or by domain, geography, or

user agent string, do not adequately reduce traffic, and updating robots.txt can have limited effect. This rise in aggressive bot activity could potentially result in open public and cultural repositories limiting access to their resources for machine agents, as well as human users.

2) Other challenges and opportunities

Overall, the GLAM-E Lab report argues that AI providers need to develop more responsible ways to interact with websites hosting cultural data and online collections, without which may lead to limitations in access and a diminishment of the social value and public good of the global open repository network. More specifically, the report urges to address ethical challenges and those in relation to crossing jurisdictions.

Ethical

A number of ethically related challenges are raised in the report, not least the current path for cultural institutions is not sustainable. GLAMs are not resourced to incorporate more servers, deploy sophisticated firewalls, and/or able to bring on board specialised operations engineers to tackle escalating and large-scale data harvesting. According to the findings, reporting abuse by identifiable bots can have some impact on lessening traffic.

Crossing jurisdictions

The GLAM-E Lab's work is primarily focused on making collections open in both the digital sense (available online) and in the legal sense (free from legal barriers to use and reuse, including in commercial context). This report includes data on both open (openly licensed) and online-but-closed collections. In part, this is because the bots do not appear to be altering behaviour in response to the presence, or absence, of legal restrictions. As highlighted in the report, "even if the licensing status was relevant, it is not always easy for bots to identify the licenses attached to a work it encounters online" (p. 21).

Of relevance to the report, The GLAM-E Lab previously outlined a number of challenges and opportunities in an official submission to the UK IPO, DSIT and DCMS

Copyright & AI Consultation, February 2025. [191] In the submission there is a suggested strategy to enable AI model training on out-of-copyright materials which can then more effectively drive resources and new investment to the cultural heritage sector.

[191] Wallace, A. (n.d.). *GLAM-E Lab Response to IPO, DSIT and DCMS Copyright & AI Consultation, February 2025*. GLAM-E Lab. https://www.glamelab.org/products/glam-e-lab-response-to-ipo-dsit-and-dcms-copyright-and-ai-consultation/

Section 7. Conclusion

The intersections between AI and copyright are marked by complexity, fluidity, and rapid technological evolution. AI's transformative role has led the domains of ethics, law, and technical standards, along with affected stakeholders into a shared grey zone; one that will need to be carefully navigated as technologies and governance structures continue to evolve.

The aim of the report has been to draw attention to these areas of uncertainty and interconnection by exploring how established frameworks are being challenged and reshaped by AI. The observations across the four grey zones are outlined below and they highlight considerations which can inform future policymaking and dialogues.

Ethics

Copyright rests on ethical and jurisprudential foundations that safeguard the market value and dignity of creative work. Emerging concerns regarding model collapse and cultural entropy further underscore the ethical complexities of integrating AI into cultural and creative domains. The comparison between AI training and human learning remains contested, particularly in relation to differences of scale, purpose and societal impact. Future reforms will have to consider how to strike a balance between public and private interests in a way that is ethically and jurisprudentially aligned.

Law

The comparison between AI training and human learning is not only an ethical challenge but also a legal one when it comes to arguments about memorisation and data regurgitation. It raises important questions about the difference between statistical pattern learning in a technical sense on the one hand and making copies of copyright works in a legal sense on the other. Legal complexities are further amplified by the geographical location of AI training as it influences legal liability. Additionally, the technical nature of AI training combined with the convoluted content hosting dynamic of the internet means it is not straightforward to know what works make part of a training dataset and who owns copyright in a particular work. This, in turn, can lead to unintended infringement and challenges with rights management.

Technical standards

The avoidance of unintended infringement and the management of rights are both currently being facilitated by technical standards, albeit imperfectly. Effective development and implementation of technical standards for rights management depends on alignment with evolving legal and industry needs. While standards cannot replace legal frameworks, they can complement them by establishing interoperable, community driven best practices that can promote consistency and trust. However, to avoid fragmentation, it is important for standards to have legitimacy through industry consensus and regulatory uptake. Moreover, engaging a broad spectrum of stakeholders, including rightsholders, technologists, hybrid stakeholders, etc. in the standards development process can ensure equity in participation. It can also ensure that interests of diverse groups are entrenched and catered for in the best practices that emerge. Finally, designing rights management tools and standards to be accessible and easy to use without the need for high levels of technical literacy can simplify the use and implementation of technical standards.

Hybrid stakeholders

As the boundaries between “creatives” and “technologists” become increasingly fluid, the adequacy of traditional frameworks is being questioned. Hybrid roles exist on a spectrum and illustrate the entanglement of technical, cultural, and legal systems in

shaping creative productions in the digital era. Rather than replacing existing categories, hybrid stakeholder roles surface important questions on how creativity and innovation can meaningfully intersect in a sustainable legal framework.

Annexure I. Emerging technical standards

Name	Description	Developed by	Focus
C2PA Specifications for Content Credentials	An open technical standard that provides a framework to communicate and verify the origins and history of digital content. C2PA Manifests (also known as Content Credentials) embed provenance metadata in the digital asset, including origin, modifications, and use of AI [192]. It also allows others to add further information, such as who created the content and whether they allow their content to be used for AI training. Content Credentials can be further implemented in the context of AI and ML through: • **AI-ML Model Content Credentials**. They provide information about model provenance and authenticity (e.g. model description, training data used to build the model, and other model training information, and explainability, transparency, and trustworthy-relevant indicators) and can also embed the model provider's preferences on the use of the model and model outputs to train other models.[193] • **AI-ML Training Data Set Content Credentials**. They provide credentials for the training dataset (including each partitioning) and how this was used during the training of the models.[194]	Coalition for Content Provenance and Authenticity (C2PA), a cross-industry standards development organisation.	Content and model provenance, authenticity, and rights expression

[192] Coalition for Content Provenance and Authority (n.d.) *C2PA and Content Credentials Explainer.* Retrieved 22 October, 2025, from https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html

[193] Coalition for Content Provenance and Authority (n.d.) *Guidance for Artificial Intelligence and Machine Learning.* Retrieved 22 October, 2025, from https://spec.c2pa.org/specifications/specifications/2.2/ai-ml/ai_ml.html#_ai_ml_training_data_set_content_credential

[194] Coalition for Content Provenance and Authority (n.d.) *Guidance for Artificial Intelligence and Machine Learning.* Retrieved 22 October, 2025, from https://spec.c2pa.org/specifications/specifications/2.2/ai-ml/ai_ml.html#_ai_ml_training_data_set_content_credential

JPEG Trust (ISO/IEC 21617)[195]	A formalised ISO/IEC standard aimed at establishing trust in digital media through secure and verifiable annotations throughout the asset's lifecycle. Many aspects of the framework are applicable to other image file formats, audio, and video, and is compatible with C2PA. Currently JPEG Trust consists of four specification parts: • **The Core Foundation** (ISO/IEC 21617-1, officially adopted) 'specifies aspects of authenticity, provenance, attribution, intellectual property rights, and integrity through secure and reliable annotation of the media assets throughout their life cycle' • **Trust Profile Catalogue** (under development) 'provides a catalogue of snippets that can be used for the purpose of constructing Trust Profiles, which can be used for assessing the trustworthiness of media assets in given usage scenarios.' • **Media Asset Watermarking** (under development) 'to support tools and mechanisms for content authenticity, provenance, integrity, labelling, and binding between JPEG Trust metadata and corresponding media assets.' • **Reference Software** (under development) 'provides a set of JPEG Trust reference software implementations'.	JPEG working group, a joint initiative between the International Organization for Standardization and the International Electrotechnical Commission.	Content provenance and authenticity
International Standard Content Code (ISCC) (ISO 24138:2024)	An open-source, ISO standard published in 2024 to enable reliable tracking of content metadata and rights information across digital content lifecycle. Unlike embedded metadata, which can be stripped during online sharing or file alternations, ISCC uses a soft-binding approach to maintain links to	Liccium with support from the European Commission	Content Provenance, authenticity, and rights expression

[195] European Union Intellectual Property (2025). The development of generative artificial intelligence from a copyright perspective. Doi: 10.2814/3893780
JPEG Trust (n.d.) *Overview of JPEG Trust. Retrieved 22 October, 2025, from* https://jpeg.org/jpegtrust/

	external metadata and opt-out declarations. It allows for clear identification of identical or similar content. Through scalable matching technologies it can detect higher degrees of similarity in content (either because of different versions or formats of the same content)[196]		
Text and Data Mining Reservation Protocol (TDMRep)	Originated from requirements set on the EU CDSM Directive, this web protocol allows content owners to signal their reservation of TDM rights on lawfully accessible web content using the *tdm-reservation* property, and, through its *tdm-policy* property give access to both contact information of publishers and authorisation conditions for mining content.[197] Initially designed to be used by publishers or other intermediaries controlling entire websites, it is less practical for individual creators whose content may be distributed across multiple platforms. The specifications outline five techniques tailored to different media and use cases: a JSON file stored on the origin server with a similar syntax to robots.txt, HTTP response headers, HTML metadata, metadata in EPUB2 and EPUB3, and XMP metadata in PDF files, allowing crawlers to detect TDM rights without accessing full content.	W3C Text and Data Mining Reservation Protocol Community Group (International standards organisation).	Rights expression and management
TDM·AI	A registry-based protocol for creators and rightsholders to inseparably bind their machine-readable usage preferences for TDM to their digital content using digital fingerprints and content-derived identifiers. It builds on top of ISO 24138 and W3C's Creator Credentials. It supports three metadata binding approaches: location/domain-based (e.g. robots.txt, HTML/HTTP metadata of the domain), asset-based (e.g. provenance metadata, C2PA), and registry-based (ISCC fingerprints and metadata in publicly accessible	Liccium	Rights expression and management

[196] European Union Intellectual Property (2025)

[197] World Wide Web Consortium (2024, May 10). *TDM Reservation Protocol (TDMRep)*. Retrieved 22 October, 2025, from https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

	registries), ensuring rights management remains effective even after content is shared or altered.[198]		
DECentralized Opt-in/out Registry for AI Training (DECORAIT)	A prototyped decentralized registry that allows content creators express their TDM preferences, trace provenance of generative AI training data, and recognise and reward content creators for their contributions. Provenance is traced via visual matching, leveraging C2PA, and consent and ownership registered using distributed ledger technology.[199]	Liccium	Rights expression and management
AI.txt[200]	A machine-readable file that defines permissions for commercial text and data mining. Placed in the root directory of a website, it specifies whether images, media, or code hosted on the domain may be used to train AI models. By default, all content is opted out, but site owners can adjust settings to allow specific content types. When permission is granted, data miners can identify which materials are available for AI training in accordance with the declared preferences.	Spawning	Rights expression and management
Media Manager	A tool under development that allows creators and content owners to specify how their works are included or excluded from AI training and research. It uses advanced machine learning to identify copyrighted content across multiple sources and reflect creator preferences.	OpenAI	Rights expression and management
CC Signals[201]	A machine- and human-readable framework to promote reciprocity in AI's use of large content collections (e.g. repositories),	Creative Commons, a non-profit organization	Reciprocity and content reuse

[198] TDM·AI (n.d.). *What is the TDM·AI Protocol?* Retrieved 22 October, 2025, from https://docs.tdmai.org/

[199] Balan, K., Gilbert, A., Black, A., Jenni, S., Parsons, A., & Collomosse, J. (2023, November). DECORAIT-DECentralized Opt-in/out Registry for AI Training. In *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production,* pp. 1-10.

[200] Spawning (n.d.). *Spawining's ai.txt. Retrieved 22 October, 2025, from* https://site.spawning.ai/spawning-ai-txt

[201] Hardinges, J., Pearson, S., & Ross, R. (2025). From Human Content to Machine Data. Introducing CC Signals. *Creative Commons.* Retrieved 22 October, 2025, from https://creativecommons.org/wp-content/uploads/2025/06/Human-Content-to-Machine-Data_Final.pdf

	whereby their stewards indicate their preferences on content reuse and/or on how they expect those benefitting from using such content contribute back. The draft framework includes four signals: • **Credit**: give appropriate credit based on the method, means, and context of use. At a minimum, it requires that the reusers cite the training dataset and that models using information retrieval techniques cite the collection as a hyperlinked source in the outputs. • **Direct Contribution**: provide proportionate monetary or in-kind support to the content stewards for their development and maintenance of the assets, based on a good faith valuation considering use of the assets and financial means. • **Ecosystem Contribution**: provide monetary or in-kind support back to the ecosystem from which the reuser is benefiting, based on a good faith valuation considering use of the assets and financial means. • **Open**: the AI system used must be open.	and international network	

The Alan Turing Institute

turing.ac.uk

@turinginst