

# A Practical Guide to AI and Copyright



Edition #11  
By Oliver Patel



Hey 🙌

I'm Oliver Patel, author and creator of **Enterprise AI Governance**.

This free newsletter delivers practical, actionable and timely insights for AI governance professionals.

My goal is simple: to empower you to understand, implement and master AI governance.

If you haven't already, sign up below and share it with your colleagues. Thank you!

For more frequent updates, be sure to follow me on [LinkedIn](#).

This special edition of Enterprise AI Governance contains everything you need to know to navigate the thorny issue of **AI and copyright**. It is packed full of practical resources, including:

- ✓ 5 practical steps to mitigate risks
- ✓ A primer on AI and copyright
- ✓ International regulatory snapshot **US EU GB JP CN KR**
- ✓ Relevant litigation and fair use arguments
- ✓ Generative AI and Copyright Cheat Sheet (*developed in partnership with Copyright Clearance Center*) - scroll to the bottom for this free **subscriber only** pdf download!

If this topic interests you, then check out this **LinkedIn Live event** I am speaking at today (Thursday 27<sup>th</sup> February) at 3pm GMT / 10am EST. It's a fireside chat on '*Bridging Innovation and Integrity in Responsible AI*', hosted by **Copyright Clearance Center**. We'll be covering how to build and implement an enterprise AI governance programme, as well as copyright and AI challenges. 800+ people have signed up so far and it would be great to see you there!

Thanks for reading Enterprise AI Governance!  
Subscribe for free to receive new posts each  
week.

## 5 practical steps to mitigate risks

Frequent readers of this newsletter know that I try to front-load value and actionable advice. Therefore, before my deeper analysis of the AI and copyright issue, here are 5 practical steps organisations can take today to mitigate copyright risks.

1. **Licensing:** use data and content explicitly licensed for use in AI systems. Where appropriate, renegotiate contracts to enable this.

2. **Contractual safeguards:** negotiate robust clauses, such as IP indemnity protection, in contracts with AI providers. However, be wary that these indemnity clauses do not apply in all circumstances, nor do they offer complete protection.
3. **Data lineage:** ensure copyright considerations are a core part of AI risk assessments. These assessments should track and record the lineage of all data used for AI training and development.
4. **AI literacy:** scale employee training and awareness, particularly on document and multimedia uploads and prompting. Now that AI is at the fingertips of all employees, anyone can exacerbate copyright infringement risk.
5. **EU AI Act compliance:** General-Purpose AI (GPAI) model providers must prepare for their obligation to publish training data summaries. This is applicable from August 2025 (or 2027 for models already on the market). The forthcoming [GPAI Model Code of Practice](#) and associated template will outline how to do this. AI deployers and downstream providers (which integrate GPAI models into new AI systems) should conduct robust due diligence and choose their providers wisely.

## **A primer on AI and copyright**

### **What is copyright?**

Copyright is a **form of intellectual property** that protects original works of authorship, including articles and books, to encourage and promote culture, science, and innovation.

Under copyright law, copyright owners, such as authors and publishers, have exclusive rights regarding how their work is used, shared, or reproduced.

Copyright laws attempt to strike a balance between ensuring fair compensation for creators of original work, whilst enabling wider society to benefit from that work, for example by permitting engagement, commentary, and transformation.

Throughout history, technological developments, such as the advent of radio, television, and the internet, have fundamentally altered how content is developed, distributed, and consumed.

Each new wave of technological change has been accompanied with meaningful changes in copyright laws and the way they are applied.

### **AI and copyright: what is the issue?**

Rapid advances in the field of machine learning, and generative AI in particular, raise important legal questions and challenges relating to copyright.

These challenges are relevant for all organisations which use AI.

Copyright infringement is a significant risk of AI development and deployment. This is because of the vast amount of published material protected by copyright, such as books, videos, music, news articles, and software, required to train certain AI models.

Those AI models can also reproduce copyrighted material in their outputs.

If AI-generated outputs closely resemble original works (which formed part of the training data) this can infringe copyright.

Copyrighted material can also be used—whether intentionally or unintentionally—to prompt, augment, or fine-tune AI.

Therefore, although the large foundation model developers are currently in the spotlight, many other organisations could get caught in the crosshairs of copyright litigation.

Organisations must also consider whether their AI-generated output (e.g., content) is protected by copyright. In most jurisdictions, the key factor is the level and nature of human involvement in generating this output.

AI-assisted works may be copyrightable, but purely AI-generated works may not be.

The copyright risks of AI are apparent in the dozens of lawsuits copyright owners have brought against unauthorised use of their protected works in AI systems.

## **Training AI models**

Training AI models, such as Large Language Models (LLMs), often involves the collection and processing of large amounts of copyright protected material.

LLM training involves breaking down this **text into tokens**, each of which is assigned a unique integer ID (i.e., a mathematical representation of the text, which the machine can process).

During training, the model processes these tokens, identifying patterns and structures, effectively compressing and retaining representations of the training data, including potentially copyrighted content, within its parameters.

During inference, the model predicts the next token which should appear in the sequence, based upon the statistical patterns and structures it has been exposed to.

This process allows the model to generate new text based on learned patterns.

The key question is whether the copying and use of this data, for AI training and

inference, infringes on copyright. There are strong arguments both for and against, which I summarise below.

### Copyright infringement risk across the AI lifecycle

- **AI model training:** many AI models are trained on large datasets, often incorporating copyrighted material.
- **AI model fine-tuning:** the fine-tuning and adjustment of pre-trained AI models can be done leveraging copyright protected material.
- **Retrieval-augmented generation:** RAG applications can incorporate or reproduce copyrighted material as part of their retrieval corpus or augmented outputs.
- **Inputs and prompts:** users can input or upload copyrighted material as part of prompts, such as images, documents, or text excerpts.
- **AI inference & outputs:** AI-generated content may reproduce or closely resemble copyrighted material, potentially infringing.

### AI-generated outputs



Whilst some organisations will primarily be concerned with whether their AI development activities are copyright infringing, many others will be keen to understand whether their AI-generated outputs are copyright protected (i.e., copyrightable).

The extent to which AI-generated outputs can be copyright protected varies. In most jurisdictions, [human authorship is required](#).

This is because copyright is a legal right afforded to humans. Therefore, without human authorship, there is no copyright.

This has been reinforced in [several legal cases](#), which determined that AI itself cannot be the author of a copyright protected work. In other words, AI-generated work is not copyrightable if no human was involved.

‘AI-assisted’ works may be copyrightable, whereas purely AI-generated works are most likely not. What matters is the nature and degree of human involvement and creativity.

For example, the [U.S. Copyright Office’s position](#) is that merely inputting a prompt is insufficient for copyright protection. This is because simple prompting does not usually enable the user to exert meaningful control over the nature of the output.

However, this changes if the human's work or creativity is perceptible in the AI output, or if the human can modify, arrange, or shape the AI-generated output in a meaningful way. In such cases, copyright protection may be granted for a human-assisted AI generated work. This will usually be determined on a case-by-case basis.

## **International regulatory snapshot**

There is no global copyright law. Rather, there is significant global divergence in copyright laws and their application to AI.

Also, the applicable legal framework is determined by the jurisdiction in which the copying happened.

Therefore, it is not possible to make blanket claims about whether AI development or deployment infringes copyright, as it will always depend on the relevant jurisdiction, the material which has been used, and exactly how that material has been used.

Crucially, there are exceptions in copyright law, such as 'fair use' in the U.S. These exceptions stipulate that, in specific scenarios, copyrighted material can be used without permission.

Given that there is broad international consensus on the requirement for human authorship for AI output to be copyrightable, the below will instead focus on whether the use of copyright protected material in AI training, development, and deployment is permitted under applicable exceptions.

**us USA:** There is no specific law or regulation regarding AI and copyright, so existing exclusive rights (like the copyright holder's right of reproduction) and exceptions and limitations (like fair use) apply. Whether AI model training constitutes fair use (and is therefore not copyright infringing) is determined by courts on a case-by-case basis. When evaluating whether an activity constitutes fair use, U.S. courts consider four key factors:

1. Purpose and character of the use.
2. Nature of the copyrighted work.
3. Amount of the portion taken.
4. Market effects.

**EU EU:** The [EU Copyright Directive](#) includes two text and data mining (TDM) exceptions—one for scientific research and another that allows broader use (including for commercial purposes), unless the copyright holder opts out. The TDM exception could apply to AI training, but this has not been tested in courts.

The EU AI Act reinforces that existing EU copyright law applies to AI systems and introduces transparency requirements regarding the training data used. Specifically, providers of general-purpose AI models will be obliged to publish summaries of their training data and implement a copyright compliance policy.

**GB UK:** The UK launched a [Copyright and AI consultation](#) in December 2024, which outlined a proposal to adopt a framework akin to the EU's (i.e., with a data mining exception covering commercial use). This would create a more permissive regime than is currently in place. The consultation closed earlier this week. The government's proposals were [widely criticised by artists](#) and publishers, but backed by AI companies.

**JP Japan:** Japan has a [nuanced approach](#). While Article 30-4 of the Copyright Act permits AI training in some cases, this is not a blanket exemption, and it has limitations. For example, if copying the material for "enjoyment purposes" or prejudice to rights holders are involved. Such prejudice includes "material impact on the relevant markets". Furthermore, AI-generated outputs that are "similar" and "dependent" on copyrighted works can still be infringing.

**cn China:** China has a stricter AI and copyright legal framework. There are no explicit exemptions for TDM or AI training. Enforcement by Chinese courts is expected to be relatively strict.

**KR Republic of Korea:** The Korean Copyright Commission encourages AI companies to proactively secure licenses, prior to using copyrighted material for AI training.

## **Relevant litigation and fair use arguments**

There are dozens of ongoing legal cases around the world adjudicating AI and copyright disputes.

One tracker lists 39 relevant cases, including Getty Images v Stability AI, The New York Times v Microsoft and OpenAI, and Farnsworth v Meta.

Many cases are in the U.S., where courts are determining whether various uses of copyrighted material in AI systems are infringing or constitute fair use.

There are strong arguments on both sides of the debate.

Those arguing that AI training does not meet the criteria for fair use highlight that original copyright protected materials (e.g., text) are copied for training and then stored, as mathematical representations, in model parameters, which is why LLMs can 'memorise' and reproduce training data content, thereby infringing the right to reproduction.