SLAUGHTER AND MAY/

TRAINING DATA IN, COURT CASE OUT?

Copyright infringement risks
in training Generative AI

OCTOBER 2025

/ INTRODUCTION

When ChatGPT was released in 2022, it started an explosion in the use of, and investment in, generative AI. The majority of the debate at that time was centred around doomsday theories and the possibility that it might someday replace workers in fields as diverse as consulting, business and the arts. Three years later, the technology has significantly matured, and so has the debate surrounding its use and development. Another, more nuanced question has taken centre stage – do developers need to obtain permission to use third party content to train their models?

This question has generated heated debate across the world, led governments to consider whether existing copyright laws are fit for purpose, and given rise to a number of copyright infringement claims against AI developers in multiple jurisdictions.

In this briefing we take a closer look at what the issue is, the approaches being adopted in the UK, the EU and the US to address it, and some of the key cases being litigated before the courts.

As with all things AI-related, this is an area of law that is developing almost as quickly as the technology itself, meaning that the below will likely only represent a snapshot in time. However, we regularly publish content on this topic, including via our blog The Lens, so please do keep an eye out for further updates from us in order to ensure you remain up to speed with the latest developments.

/ CONTENT

/ CONTACTS

LAURA HOUSTON
Partner
+44 (0)20 7090 4230
laura.houston@slaughterandmay.com

DAVID IVES
Partner
+44 (0)20 7090 4266
david.ives@slaughterandmay.com

RICHARD BARKER
Senior Knowledge Lawyer
+44 (0)20 7090 4522
richard.barker@slaughterandmay.com

RALIE VENCHEVA
Associate
+44 (0)20 7090 5743
ralie.vencheva@slaughterandmay.com

WHAT'S THE ISSUE?

Generative AI, or "Gen AI" for short, is a type of artificial intelligence, which is designed to analyse underlying patterns in training data, and generate outputs which resemble the patterns identified in the original dataset.

Training generative AI models requires a lot of data, which often comes from publicly available sources, such as open datasets, as well as from crawling and scraping the internet. That data will typically include copyright-protected content such as artwork, books, music or photographs. While training methods do vary, copies of the underlying content tend to be made at some stage of the training process. That copying, if carried out without the relevant rights holders' consent and without the benefit of a statutory exception, may give rise to infringement risk.

Generative AI providers believe that their current practices are (or at least ought to be) legal. Rights holders, however, are equally firm in their view that use of their content without a licence amounts to infringement. This debate, and where the balance should fairly lie, goes to the very heart of whether the current approach to training generative AI is workable. In turn, that has led people to question whether our current laws are fit for purpose and, in particular, whether any exceptions or defences to copyright infringement are available – or should be available – to generative AI developers.

In the remainder of this briefing, we will therefore focus on the key copyright exceptions that are currently available or being considered in each of the EU, the UK and the US, and we'll discuss some of the key cases being litigated before the courts in each of these jurisdictions.

THE EU APPROACH

TDM Exception

The EU's copyright framework includes two exceptions for "text and data mining" or "TDM" which were brought in under the Directive on Copyright in the Digital Single Market (Directive 2019/790) (the "**CDSM Directive**") – one for scientific research (Article 3), and a broader one which allows TDM for any purpose, including commercial purposes, subject to rights holders having the ability to reserve their rights and opt their content out (Article 4).

Article 3 (Text and data mining for the purposes of scientific research)

1. Member States shall provide for an exception…for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

Article 4 (Exception or limitation for text and data mining)

1. Member States shall provide for an exception or limitation… for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

(…)

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

As can be seen from the text box above, rights holders are able to reserve their rights and opt their content out of the broader exception, but not the exception for scientific research.

Perhaps surprisingly, however, in the first and only EU case to date that has considered the implications of including copyright-protected works in a dataset to be used for training generative AI tools (a decision of the District Court of Hamburg relating to the LAION-5B dataset), it was the narrower exception that was found to apply and which rendered the copying of an unlicensed photograph non-infringing.

In most cases, however, the debate will be around whether the broader exception applies. Whilst the opt-out provision in that exception is aimed at striking a fair balance between AI developers and rights holders, it's not yet settled what this requires and how rights holders should best express their opt out. Some guidance is given in Recital 18 of the CDSM Directive itself, which states:

// *In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service.* //

However, as noted in a recent study commissioned by the European Parliament's Committee on Legal Affairs, whilst the CDSM Directive was intended to harmonise the position across the EU, Article 4 has not been implemented by all EU Member States in a uniform manner. For example, some Member States require use of technical protocols (e.g. using the Robot Exclusion Protocol or metatags) for a valid opt out, whilst others do not. There is also debate about the meaning of terms such as "machine-readable", and whether it includes natural language terms and conditions - obiter comments by the Hamburg District court in the LAION case suggested that written terms and conditions might suffice, but ultimately the court did not have to determine the point as it had already found the Article 3 exception applied (see above).

All of this, combined with the fact that no uniform opt-out mechanism has emerged as industry standard, has led to significant uncertainty amongst both AI developers and rights holders about whether any given opt-out will be valid and enforceable.

Even if an opt out is done correctly, rights holders have expressed concern that the black box nature of many AI systems can make it difficult for them to identify where their works may have been used as training data in breach of their opt-out. They argue that this lack of transparency further undermines the effectiveness of the opt-out mechanism and raises concerns about accountability and compliance.

The EU AI Act

Partially in response to those concerns, the EU legislators ultimately included two copyright-focussed obligations in the EU AI Act.

Under those obligations, providers of General-Purpose AI ("**GPAI**") models that are placed on the EU market are required to:

1. put in place a policy to comply with EU copyright law, including any rights holders' TDM opt outs (Article 53(1)(c)); and

2. disclose a "sufficiently detailed" summary of the content used to train their GPAI models, based on a template provided by the AI Office (Article 53(1)(d)).

It's worth highlighting that these two obligations only apply to providers of GPAI models that are placed on the EU market. This means they do not apply to: (i) other actors in the AI supply chain, such as deployers, importers, distributors or users; (ii) AI systems or models that fall outside the definition of GPAI models; or (iii) GPAI models that are not placed on the EU market.

Both of these obligations came into effect on 2 August 2025, with the AI Office having the ability to enforce them from 2 August 2026. For models that were already on the market before 2 August 2025, however, there will be a two year grace period, giving providers of those models until 2 August 2027 to comply.

There is the potential for severe penalties for non-compliance, including fines of up to 3% of the provider's annual total worldwide turnover or €15m (whichever is higher). GPAI model providers do therefore need to take these provisions seriously.

Policy to comply with EU copyright law

The self-expressed aim behind the Article 53(1)(c) obligation is to prevent AI developers from forum shopping and to create a *"level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union"*.

The concern seems to be that, absent such a requirement and owing to the traditional territorial scope of copyright, AI developers might be able to circumvent EU copyright-holders' rights (including any TDM opt-outs), by carrying out the training of their models in a jurisdiction with fewer restrictions on the use of copyright works and subsequently importing these models into the EU.

Indeed, the EU AI Act's recitals make it clear that this obligation applies *"regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those [GPAI models] take place"*.

The obvious question then is: what exactly does the required copyright policy need to contain? The EU AI Office has provided guidance on this in the form of the copyright chapter of the General-Purpose AI Code of Practice ("**GPAI Code**"), which was published on 10 July 2025.

The GPAI Code is a voluntary code of practice, written by independent experts with input from stakeholders, which is designed to help GPAI model providers demonstrate compliance with Articles 53 and 55 of the AI Act.

The final version of the GPAI Code is made up of three chapters, one of which is on copyright. That chapter sets out five measures that signatories to the GPAI Code agree to implement in order to demonstrate compliance with Article 53(1)(c).

Measure 1.1: Signatories agree to draw up, keep up-to-date and implement a policy to comply with EU copyright law for all GPAI models they place on the EU market. Whilst signatories are encouraged to make a summary of their policy publicly available, that is not mandatory.

Measure 1.2: Regulates the mining of web-crawled content and aims to ensure that signatories only reproduce and extract lawfully accessible works for training purposes. This includes commitments not to circumvent technological measures (e.g. paywalls and subscription barriers) and to exclude certain piracy-focussed domains from their web-crawling.

Measure 1.3: Sets out signatories' commitments for identifying and complying with rights holders' TDM opt-outs. This includes employing web-crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol; engaging with rights holders to develop machine-readable standards for expressing a rights reservation; and publishing information about the web-crawlers used, their robots.txt features and any other measures adopted to identify and comply with rights reservations.

Measure 1.4: Sets out commitments to mitigate the risk that a downstream AI system generates infringing outputs. That includes implementing technical safeguards to prevent GPAI models from generating outputs that reproduce copyright-protected training material, as well as prohibiting copyright infringing uses in signatories' acceptable use policies and terms and conditions.

Measure 1.5: Signatories commit to designate a point of contact for affected rights holders to communicate with and to put in place a mechanism to allow affected rights holders to lodge complaints about signatories' non-compliance with the copyright chapter of the GPAI Code.

In addition to the five measures noted above, the copyright chapter highlights a very important distinction between compliance with the GPAI Code and compliance with EU copyright law. While compliance with the GPAI Code can help providers demonstrate compliance with Article 53(1)(c) of the AI Act, it will not necessarily equate to compliance with EU copyright law. Indeed, the GPAI Code itself makes it clear that it has no effect on the application and enforcement of EU copyright law.

Whether the GPAI Code will be a success remains to be seen and will depend, in large part, on how many GPAI model providers sign up to and adhere to it. At the time of writing, there are 27 signatories, including Google, OpenAI, Mistral, Anthropic and Microsoft. But some of the other big players, such as Meta, Perplexity AI, Midjourney and Stability AI are currently missing.

Template summary of training data

As already noted, the purpose behind the Article 53(1)(d) obligation is to ensure greater transparency in the data used to train GPAI models, thereby enabling rights holders to better identify when their works have been used and whether their opt-outs have been adhered to.

At the same time, however, developers of GPAI models often regard their training data as trade secrets or confidential business information and so, as the global AI race continues to intensify, they have become increasingly protective of that data.

The EU AI Act seeks to balance these competing interests by requiring GPAI model providers who place their models on the EU market to publish a "sufficiently detailed summary" of the content used to train those models, in the form of a template provided by the AI Office – the aim being that this will enable parties with legitimate interests, including copyright holders, to exercise their rights while at the same time giving due consideration to developers' needs to protect their trade secrets and confidential business information.

The Recitals to the AI Act provide some limited guidance on the scope of this disclosure obligation, but the key document to consider is the template that was published by the European Commission's AI Office, together with an explanatory notice and related FAQs page, on 24 July 2025. As the Commission's general FAQs make very clear, *"using the template is mandatory and serves as the sole guidance for providing those public summaries"*.

The template itself is broken down into three key sections.

1. **General information** – this covers basic information about the model, including the identity of the model and its provider, the modality of the model (text, images, etc.), a high-level indication of the amount of training data used to train the model and more general characteristics of that training data.

2. **List of data sources** – this section requires GPAI model providers to:

   • disclose the main datasets that were used to train the model;

   • provide a narrative description of data that has been crawled and scraped from online sources (either by the provider or on its behalf). This includes details of the crawlers used, their purpose and behaviour, the period of collection and a summary of the top 10% of all domain names scraped,

determined by the size of the content scraped (although there is a lower threshold for GPAI model providers who are SMEs); and

- provide a narrative description of all other data sources used, including user data (e.g. data collected from user interactions with the GPAI model or with the model provider's other products or services, such as social media or content platforms) and synthetic AI-generated data that was created by the model provider to train the model.

3. **Other data processing aspects** – this includes measures implemented by the provider to comply with rights holders' opt outs from the TDM exception and to remove illegal content.

The explanatory notice clarifies that the information to be provided should cover data used in all stages of model training, from pre-training to post-training. However, data used during operation of the model (such as through retrieval-augmented generation) is not covered, unless the model actively learns from that data.

Recognising the need to balance transparency with the protection of model providers' trade secrets, different levels of detail are required based on the source of the data in question, with greater detail required about the use of publicly-available datasets and more limited disclosure required for other sources such as licensed data.

Once complete, the summary has to be published on the model provider's official website, as well as on all of the model's public distribution channels, at the latest when the model is placed on the EU market (or by 2 August 2027 for models placed on the market before 2 August 2025).

Whilst it's clear that the AI Office has tried to achieve a balance between transparency on the one hand and protecting GPAI model providers' trade secrets and confidential business information on the other, initial reactions suggest that neither model providers nor rights holders are particularly enamoured with the outcome. Some model providers remain concerned that it may lead to them having to disclose confidential information and trade secrets; whilst many rights holders argue that it doesn't go far enough and won't facilitate them in enforcing their rights. Questions are also being raised about the scope and meaning of some of the requirements in the template, such as the need to summarise the top 10% of domain names scraped and how that threshold is to be determined.

Whether or not this template ultimately achieves its objectives remains to be seen. The real test of its efficacy will be in the approach leading providers take to populating it and, even more critically, how strictly compliance will be enforced and how the template's disclosure requirements will be interpreted.

The elephant in the room - does the TDM exception even apply to training AI?

Having said all of the above, an even larger question remains – do the activities undertaken to train a generative AI model actually amount to TDM at all? Since the emergence of popular AI models, the general assumption in the legal and policy world has been that they do. That assumption was arguably strengthened with the passing of the EU AI Act, which expressly refers to the TDM exception in Article 4 of the CDSM Directive and the need for GPAI model providers who place their models on the EU market to comply with rights holders' TDM opt outs.

However, a recent study report on Generative AI and Copyright commissioned by the European Parliament's Committee on Legal Affairs casts doubt on whether this is indeed the case, citing concerns that the TDM exceptions were aimed at the extraction of patterns from data for analytical purposes, while Gen AI is concerned with the synthetic reassembly of data to produce new content. The study concludes that *"generative AI training does not fall within the scope of Articles 3 and 4 of the CDSM Directive"*. It remains to be seen whether anything will directly come of this, but, fortunately, similar questions have now been referred to the CJEU in *Google v Like Company* so we should receive an answer on this critical point in the not too distant future.

That case (which concerns the production of text by Google's Gemini (Bard) chatbot which was partially identical to text contained on one of Like Company's websites) has arisen in the context of press publishers' rights, rather than copyright per se, but the CJEU's views on the questions referred will be instructive in the copyright context. Of particular relevance to this briefing, the questions referred include the following:

1. Is the process of training an LLM-based chatbot an instance of reproduction, where that LLM is built on the basis of the observation and matching of patterns, making it possible for the model to learn to recognise linguistic patterns?

2. [If yes] does such reproduction of lawfully accessible works fall within the exception provided for in Article 4 of the CDSM Directive, which ensures free use for the purposes of text and data mining?

A decision in this case is not expected until later in 2026 or perhaps early 2027, but the outcome is likely to be instrumental for the future of AI and copyright in the EU.

We might also see decisions on similar questions from the German and French courts, following recent claims filed by GEMA (a German music performance rights organisation) against Open AI and Suno AI, and by SNE (a French publishers' association) against Meta.

THE UK APPROACH

Current position and latest consultation

In contrast with the EU, as things stand, the UK only has a very narrow copyright exception for text and data mining which is limited to TDM for non-commercial research purposes. Given its non-commercial nature, it isn't particularly useful for most providers of generative AI.

Previous proposals by the UK government to introduce a broader UK TDM exception (with no ability for rights holders to reserve their rights or opt their content out) and to broker a voluntary Code of Practice on Copyright and AI have failed.

More recently, however, the UK government has been running a new consultation, which closed on 25 February 2025. In a bid to find a workable solution, the government put forward four policy options for consideration:

> I. Do nothing and leave our existing copyright laws as they are.
>
> II. Strengthen copyright by requiring licensing in all cases for training AI models.
>
> III. Introduce a broad TDM exception, with few or no restrictions.
>
> IV. Introduce a broad TDM exception, subject to rights holders having the ability to reserve their rights, which is underpinned by supporting measures on transparency.

The government's stated preferred approach in the consultation document was the last of these, which largely mirrors the approach taken in the EU. However, drawing on the EU's experiences, the government acknowledged that in order for this to be effective there needs to be a simple, standardised, machine-readable way for rights holders to reserve their rights and so it is also seeking views on whether and how it can support work to improve rights reservation tools and drive their adoption.

Highlighting the strength of views on this topic, the government received over 11,500 responses to the consultation and it's fair to say that neither side is fully

in support of the government's plan. The proposal has garnered high-profile criticisms from those in the creative sector (such as Elton John), while also drawing discontent from AI providers.

As a result, it has become very apparent that there is no easy way forward on this and that a political solution is likely needed - sooner rather than later. There is a concern, particularly from the rights holders' side, that the government is moving too slowly and that the opportunity to set an appropriate balance may be missed. That has already had a knock on effect on other pieces of legislation, such as the Data (Use and Access) Act, which was the subject of a lot of delay and controversy due to repeated attempts by those representing rights holders' interests to introduce provisions aimed at resolving aspects of this copyright debate in their favour. The Bill did eventually receive Royal Assent on 19 June 2025, but only after the controversial copyright-related aspects (which focussed on transparency) were stripped out. In their place, the UK government has agreed to produce, within nine months of 19 June 2025: (i) an economic impact assessment of the four policy options it put forward in the consultation; and (ii) a more general report on the use of copyright works in the development of AI systems. This suggests that we might have to wait until at least March 2026 for a proposed resolution on this issue in the UK – and possibly longer.

In the meantime, the English courts have been left to grapple with this difficult topic in the *Getty Images v Stability AI* case, which went to trial in June 2025.

Getty Images v Stability AI

The technology sitting at the heart of that dispute is Stability AI's generative AI tool known as "Stable Diffusion", which creates synthetic images in response to text prompts entered by users or in response to images uploaded by users (or a combination of both). Getty alleges that Stable Diffusion was trained using millions of copyright-protected images scraped from its websites without its permission. Getty originally asserted that those actions infringed its copyright and database rights. Getty also claimed that the outputs (i.e. the images) produced by Stable Diffusion infringed its rights by reproducing substantial parts of its copyright protected works or by bearing Getty's trade marks (in the form of its watermark). As a result, it brought proceedings against Stability AI for copyright infringement, as well as database right infringement, trade mark infringement and passing off.

At the trial, however, as part of its closing statements, Getty dropped its claims for direct copyright and database right infringement in the context of both the training of Stable Diffusion and the outputs generated by Stable Diffusion. Part of the difficulty on the training side lay in trying to prove that at least some training or development of Stable Diffusion took place in the UK. Copyright is a

territorial right, so there can be no direct infringement of UK copyright unless an infringing act has been committed in the UK. Ultimately, Getty clearly felt that it hadn't overcome the evidential hurdle to prove that.

The impact of this is that, from a copyright perspective, the main focus will now be on Getty's secondary infringement claim – in essence, can importing (or downloading) a pre-trained generative AI model, like Stable Diffusion, into the UK amount to secondary infringement of UK copyright?

The answer to that will ultimately come down to three things. Firstly, how the court construes the meaning of "article" and "infringing copy" in sections 22 and 23 of the Copyright, Designs and Patents Act 1988 (which provide that importing or dealing with an article known to be an infringing copy of a work is an act of secondary infringement). Historically, those provisions have only really been applied in the context of physical goods and so the court will need to decide whether this extends to intangibles. If it does, the second question is whether the court will construe those provisions as requiring the articles – i.e. the trained models - to retain the infringing copies within them. The final question is whether training Stable Diffusion would have infringed UK copyright if it had been trained here.

This is important because, if the court does find that Getty's case of secondary infringement has been made out, the question of where the model was trained – and the potential forum shopping around that – will become less important. Fortunately, we shouldn't have to wait too long to find out the answer to this as the judge (Joanna Smith) has said she should have the judgment ready just before or in the early stages of the next court term, which began on 1 October.

However, with the big questions around primary copyright infringement having fallen away in this case, and no other UK cases currently in the pipeline, there is even greater pressure on the UK government to provide clarity and to find a workable solution for both industries.

THE US APPROACH

In the US, there are no specific copyright exceptions targeted at TDM or AI more generally. Instead, the key debate in the US revolves around whether training generative AI falls within the scope of the "fair use" defence, which is set out in section 107 of the US Copyright Act 1976.

In determining whether that defence applies, four non-exhaustive factors need to be considered:

1. The purpose and character of the use – with a particular focus on whether the use is transformative and/or of a commercial nature;

2. The nature of the copyrighted work;

3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

4. The effect of the use upon the potential market for or value of the copyright work.

In May 2025, the US Copyright Office ("**USCO**") published a pre-publication version of a report setting out its views on the scope of this defence in a generative AI context (shortly following which the head of the US Copyright Office – Shira Perlmutter – was notoriously fired).

That report concluded that there is no "one size fits all" answer – whether the fair use defence applies will depend on the specific facts and circumstances of the particular case in question. On one end of the spectrum, for example, uses for the purpose of non-commercial research or analysis that don't enable copyright works to be reproduced in the outputs are likely to be fair. On the other end, copying expressive works from pirate sources in order to generate unrestricted content that competes with the original work in the marketplace is unlikely to qualify as fair use. Many uses will, however, likely fall somewhere in between these two extremes.

Separately, there have been a number of decisions handed down by the US courts, which have considered this same question.

Firstly, in *Ross Intelligence v Thomson Reuters*, the Delaware District Court rejected Ross Intelligence's attempts to rely on the fair use defence, after finding that it had used copies of over 2,000 Westlaw headnotes to train its AI-driven legal research tool. At the risk of over-simplification, the key factors in reaching that conclusion included that the use made of the headnotes was not transformative; and Ross Intelligence had used those headnotes to create a competing legal research tool, which negatively affected the market for Westlaw (meaning factors one and four above favoured Thomson Reuters). Importantly, however, the judge was keen to stress that this wasn't a generative AI case because Ross Intelligence's platform did not create new content.

More recently, the California District Court has handed down two decisions which did arise in a generative AI context.

In *Bartz v Anthropic*, the court found (on a summary judgment basis) that Anthropic's use of certain books to train its Claude LLM was transformative and did amount to fair use. However, Anthropic were not able to rely on the

defence of fair use in relation to its copying and storing of more than 7 million pirated books to build a central library – a separate use which was found not to be transformative. Following that decision, Anthropic agreed to settle the dispute for $1.5bn, a figure which the court filings suggest make it the "largest publicly reported copyright recovery in history".

Finally, in *Kadrey v Meta*, Meta was granted summary judgment in its favour on its fair use defence against claims that copying the claimants' books for use as LLM training data infringed the claimants' copyright. Whilst at first glance this would appear to be a big win for AI developers, the detail of the judgment arguably suggests otherwise, with the judge indicating that this decision was largely based on the fact that the authors hadn't presented enough evidence to convince him that Meta's model would dilute the market for their work. The judge went on to note that, in his opinion, using third party copyright works without permission to train generative AI would be unlawful in many circumstances and that "… *this ruling does not stand for the proposition that Meta's use of copyrighted materials to train its language models is lawful. It stands only for the proposition that these plaintiffs made the wrong arguments and failed to develop a record in support of the right one.*"

Whilst it is still early days, it is clear from these decisions and the USCO report that there is no bright line rule around when the fair use defence will apply and that a lot will depend on the particular circumstances of each individual case.

CONCLUSION

As all of the above highlights, the law relating to the use of third party copyright-protected materials in training generative AI remains complex and uncertain, with different jurisdictions taking different approaches and in many cases the law being subject to interpretation or leading to different outcomes based on the specific facts involved. However, with a number of material court decisions expected in the next 12-18 months, a response due on the UK government's consultation on copyright and AI, and the GPAI model rules under the EU AI Act now in force, the next few months could prove critical in shaping the future of how generative AI models are trained and how the rights of AI developers and IP rights holders are to be balanced.