Microsoft—written evidence (AIC0012)

House of Lords Communications and Digital Select Committee inquiry: AI and copyright

1. Are there uncertainties, gaps or barriers in the UK's current copyright framework that restrict innovation or rightsholders' ability to enforce their rights in relation to generative AI?

Yes. The UK's current legal framework does not meet the needs and expectations of UK industry, as it does not provide clear assurance that modern data analysis – searching, organising, and analysing large amounts of data - is legally permitted. Businesses require clarity that analysing legally accessed data, including publicly available online content, is permissible under copyright law and does not require a license from the rightsholder.

The lack of a Text and Data Mining (TDM) exception has widespread impacts beyond training conducted by AI developers. It impacts a business' ability to fine tune models or build machine learning applications in the UK using data they have legal access to. For example, a retailer wishing to fine tune a model by analysing large volumes of product feedback from customers to improve product research and development, or a media company wishing to fine tune its models on its own legally acquired archives.

This legal uncertainty as to whether data analytics is permitted negatively has impact beyond AI development. Modern data analytics (extracting information, patterns and trends from data – also known as Text and Data Mining or TDM) extend far beyond AI model development. Computational analysis is foundational to data analytics across every sector. Banks engage in TDM to search the web for relevant information to enhance lending workflows,[1] while educational AI tools used by teachers to help plan and shape lessons perform TDM on web content to bring fresh and relevant content to those plans.[2] UK companies are automating the scientific discovery process at massive speed and scale,[3] including through ARIA's investments to build automated "AI scientists".[4] A safe internet relies on TDM activity: moderation tools built by UK startups and labs rely on TDM activity to, for example, build classifiers and vision detection models. These systems rely on a content pipeline of public datasets to train an algorithm in order to evaluate new material against specific policy-based logic. Lack of clear exceptions that permit TDM can affect sectors across the entire economy and society, not just the AI sector. This undermines UK competitiveness, deters innovation, and exposes businesses to litigation risk.

---

1 https://www.microsoft.com/en/customers/story/19796-ubs-azure

2 https://www.microsoft.com/en/customers/story/20323-eani-microsoft-copilot

3 https://www.bioindustry.org/static/82163805-75d8-46dc-9fbb612255a4fba4/BIA-response-to-Government-Copyright-and-AI-consultation.pdf

4 https://www.aria.org.uk/ai-scientist

A further gap in the UK's copyright framework is that the current TDM exception under the law is narrow, applying to only TDM performed for research. This does not reflect the reality of how innovation occurs in today's economy, which routinely involves collaboration between universities, public sector bodies and the private sector, where the end goal is often commercialisation. As a result, the current exception has limited real-world application and provides little support for UK innovation.[5]

It deserves mention that the current copyright framework enables rightsholders to enforce their copyrights. Robust protections are available under UK copyright law to prevent infringing uses of copyright works, regardless of how these works are generated. A TDM exception that permits analysis of a work that is legally accessed, would not impact the ability to prevent infringing outputs. Alongside remedies available under copyright, AI systems are increasingly incorporating technical guardrails to prevent the output of copyrighted content.

For example, as part of Microsoft's broader Responsible AI approach, Microsoft embeds multiple layers of safeguards into its generative AI products such as meta-prompts, classifiers, and content filters, to prevent the generation of outputs that are similar to protected works.[6] Microsoft's Customer Copyright Commitment provides legal protection and indemnification for customers who use its Copilot and Azure OpenAI services, provided they do not disable guardrails and follow required mitigations,[7] reinforcing both responsible system design and responsible user behaviour.

a) If so, how could these be addressed?

The uncertainties and barriers presented by the current copyright framework in the UK could be addressed by introducing a copyright exception for commercial TDM. For the UK to deliver on its AI ambitions, UK industry must be given the opportunity to thrive on an international level. Providing certainty on a broad commercial TDM exception, which aligns with copyright reforms that other countries have introduced, will unlock UK businesses' ability to innovate, compete globally, and deploy AI to the benefit of citizens and public services.[8]

Introducing a commercial TDM exception does not diminish rightsholders' legitimate interests; it enables the use of non-expressive information, such as correlations, statistical patterns and common attributes across data sets. A clear TDM exception which includes commercial use, would therefore both preserve strong protections for rightsholders, and remove unnecessary

---

5 https://www.timeshighereducation.com/opinion/uks-copyright-laws-will-hobble-its-ai-ambitions

6 Microsoft (2025), Responsible AI Transparency Report, https://blogs.microsoft.com/on-the-issues/2025/06/20/our-2025-responsible-ai-transparency-report/

7 https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

8 Tony Blair Institute (2025), Rebooting Copyright: How the UK Can Be a Global Leader in the Arts and AI

barriers to lawful data analysis that underpins innovation across the UK economy. This approach aligns with other regions that have modernised their IP frameworks, or have fair use doctrines, such as Singapore, Japan, the US and the EU, and would ensure the UK is able to put AI to work in public services and the UK economy, and remains competitive internationally.

2. What practical and technical mechanisms for (a) rights reservation and (b) transparency would provide rightsholders with sufficient control over their work while being proportionate and administratively reasonable?

To be effective, proportionate and administratively reasonable, technical mechanisms for rights reservations and transparency should be straightforward to implement, widely understood, and scalable. They should integrate easily into existing practices for both TDM users and web publishers and be capable of broad adoption across markets.

Achieving this requires global approaches developed through technical norms that can evolve with the technology, rather than reliance on enforcement mechanisms that can be misaligned across jurisdictions and risk inhibiting consistent technical adoption.

On (a) rights reservation:

We recognise the principle that authors and rightsholders should have agency over how their works are used within AI training. Microsoft's early adoption of voluntary measures[9] that allow authors and rightsholders to make such a choice reflects our commitment to that principle. More recently, Microsoft has made more granular controls available to give publishers precise control over what appears in search and AI- generated answers, keeping the rest of their page discoverable.[10]

The data-nosnippet complements Bing's broader suite of metatags and HTTP headers that manage how content is indexed, cached and displayed in search results.

Common directives include:

- **noindex:** Prevents a page from being indexed.
- **nosnippet:** Blocks all text and preview thumbnails from appearing in snippets.
- **max-snippet, max-image-preview, max-video-preview:** Limit the size or duration of preview content.

[9] https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat

[10] https://blogs.bing.com/webmaster/October-2025/Bing-Introduces-Support-for-the-data-nosnippet-HTML-Attribute

- Apply **"data-nosnippet**" to any HTML element you want to exclude from Bing Search snippets or AI summaries. When Bing crawls your site, the marked content remains fully discoverable but won't appear in snippet text of AI-generated previews.

In addition, content blocked via "robots.txt (Disallow)" will not be crawled by Bing and therefore will not be indexed, surfaced in search or AI-generated previews.

Two important themes are relevant with any reservation of rights approach. First, globally harmonised frameworks for how the reservation of rights operates will need to continue to develop. This would provide clarity on what signals authors and rightsholders use to convey a preference and how developers should read and action such signals. Solutions to these questions can emerge though open consensus-based industry standards processes, including those taking place at the IETF and in other international convening forums.

Second, the application of the reservation of rights should be linked to specifically training generative AI models, rather than apply to TDM broadly. TDM activity goes well beyond training AI models and broad application of rights reservation to TDM could have unintended consequences. As mentioned in this response, performing computational analysis on data is critical to how the Internet, software, and most technologies operate. Microsoft's experience with developing a leading classifier to detect child exploitation materials involves making reproductions of photographs owned by others to prepare digital fingerprints.[11] This is one of many examples. If that process, and similar moderation tools that involve computational analysis on reproductions of works, must now account for opt outs it would impose additional compliance obligations and expose operators to risk.

The mechanisms that rightsholders can use today will continue to evolve. Many publishers currently rely on robots.txt as a location-based mechanism to indicate their opt out preference. However, because robots.txt was designed for web crawling rather than AI preference, work is being conducted to develop more granular, interoperable and machine-readable rights reservation mechanisms. These include updates to the Robots Exclusion Protocol currently being discussed at the IETF,[12] as well as emerging content-based rights reservation controls such as the International Standard Content Code (ISCC) which would allow rights reservations to be associated with the ISCC code that is generated from the content. It is important that these controls are carefully assessed by a broad group of stakeholders to ensure that they work for everyone.

On (b) transparency:

Technical measures for providing transparency should assist consumers to make informed choices and allow an assessment of the limitations and

11 https://www.microsoft.com/en-us/PhotoDNA

12 https://datatracker.ietf.org/wg/aipref/about/

capabilities of specific general-purpose AI models. Model cards are one example of existing industry steps to inform consumers about the attributes of a specific model. The information found on those cards also provides a general summary of the type of content used to train the AI model (example in footnote).[13] Generally, disclosure requirements regarding the sources of training materials are not an appropriate or effective approach to resolve copyright concerns or allegations because the question of copyright infringement should be assessed based on the similarity of an infringing work; there is no precedent for such disclosures within the UK Copyright Act.

Providing high level information about the crawlers or bots used for data collection (see example from Microsoft),[14] alongside a narrative description of the general types of data and large collections involved, offers a practical way for rightsholders to understand whether their choices such as the use of robots.txt or other technical controls have been respected. This approach supports accountability, avoids impractical source level disclosures, and aligns with emerging industry practices on transparency without undermining trade secrets or confidential information.

a) What, if any, legislative changes are needed to support rights reservation and transparency arrangements to function effectively?

A rights reservation, or "opt out", will function most effectively if implemented through evolving technical standards, rather than fixed regulatory requirements. The Government could instead consider putting in place guidelines that highlight the existence of these controls and encourage adoption by rightsholders who want to control the use of their works.

Legislative change could focus on introducing a commercial TDM exception for lawfully accessed works, while allowing the rights reservation mechanism to develop through industry-led standards that can adapt as technology evolves and establish global technical norms in the same way as compliance with robots.txt is observed globally without requiring legislative intervention. For decades, web publishers have relied on machine readable signals associated with robots.txt to declare whether a site prevents a crawler from indexing its pages and from presenting information in search results.

Comparable norms are now emerging in the context of AI training, where AI crawlers are expected to avoid a publisher's site, specific pages, or particular content for training purposes, based on information expressed using available controls.
Prescriptive legislative approaches to the opt out to introduce enforcement mechanisms risk fragmenting technical implementation across markets and inhibiting the development of technical standards.

3. What technical mechanisms or standards could support

[13] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

[14] https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat

reliable (a) labelling and (b) attribution of AI-generated content?

One example of provenance metadata that may be used for labelling AI-generated content and providing attribution for who generated the AI content, is Content Credentials, based on Coalition for Content Provenance and Authenticity's (C2PA) open technical standard. This standard - co-developed by Adobe, the BBC, Microsoft and others - enables cryptographically verifiable information such as when the content was created, which system generated it, and which organisation certified the credentials to travel with the asset in a tamper-evident way.

Microsoft attaches cryptographically signed provenance metadata to images generated with OpenAI's DALL-E 3 model in our Azure OpenAI Service, Microsoft Designer, and Microsoft Paint. Provenance information based on the C2PA standard is automatically ingested by and displayed on LinkedIn, helping users verify whether content was generated or modified with AI.

Microsoft adopts voluntary measures to disclose generative AI outputs as part of a broader commitment to protect consumers and preserve information integrity.

Microsoft draws on best practice in this space and works collaboratively with industry and civil society on frameworks, methods and means of disclosure. These efforts emphasise the importance of ecosystem-wide adoption to ensure provenance signals remain intact, usable and meaningful for end users.

At the same time, while provenance standards and techniques play an important role for generative images, audio, and video, current research indicates meaningful limitations across other forms of outputs, particularly software code, where persistent provenance information cannot be reliably embedded or preserved through routine transformation, reuse, or compilation. This underscores the importance of continued research and standards development to expand the effectiveness of provenance techniques across different media and use cases.

a) What role, if any, should legislation play in promoting or requiring such measures?

Continuing to support the development of, and adoption of, labels for content provenance and authenticity across audio, visual and image where appropriate can help build trust in AI technologies. It is important that any legislative approach recognises the modality-specific applicability of provenance tools, to avoid mandating technical requirements in areas where they are not yet feasible or effective, and is technology agnostic.

4. What are the opportunities for the development of a UK licensing market that would benefit rightsholders and AI developers, and how can these be maximised?

a) What structures and safeguards would be needed to ensure

that new licensing arrangements and revenue-sharing models are workable and that remuneration reaches individual rightsholders?

b) How would new licensing schemes handle complex, multi-party rights?

c) What role could the Government's proposed 'creative content exchange' play in this context?

A broad-based TDM exception is compatible with thriving markets for the acquisition and licensing of data in the context of AI. Developers and rightsholders are already partnering to acquire and license content in a variety of ways that one should expect to grow in tandem with a thriving UK-based market in AI development.

Microsoft's approach

Microsoft partners with rightsholders and publishers to access data for use in AI development and fine-tuning. Our 2024 data access arrangement with Informa, the UK publishing house, is one public example. Among other things, it provides access to content that would not otherwise be accessible to Microsoft for AI training purposes.

Microsoft has negotiated data access arrangements with other book publishers, academic journals, database providers, media companies, licensing platforms, news publishers, data aggregators, and coding platforms. These arrangements span from the use of archives to recent works.

Microsoft also licenses content for certain uses in products and services. For instance, Microsoft's Copilot Daily offers users an audio recap of the weather and current events within the AI assistant. We partnered with publishers like Reuters, Axel Springer, and The Financial Times[15] to acquire access to material and engage in specific uses that may otherwise infringe - e.g., outputs that re-use copyrightable expression, not just unprotected facts and ideas.

Observations about partnerships across the content ecosystem

While there are a range of approaches across the industry, there are many examples of other partnerships that also focus on access to content and/or uses to create new products and services. For instance:

- OpenAI's agreement with TIME provides access to over a century of archival material.[16]

- Perplexity is working with a range of publishers to display full content from articles and share revenue.[17]

[15] https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/do-more-with-ai/ai-for-daily-life/get-the-news-you-want-with-copilot

[16] https://time.com/6992955/time-and-openai-announce-strategic-content-partnership/

- Lionsgate has provided access to its movies to Runway, so that Runway can build the studio an AI model for use in its production process.[18]

These are a few examples of the ways developers are using AI to create new ways for people to engage with content and, in turn, create new revenue opportunities for rightsholders and publishers. Netflix, for example, has introduced new AI-driven experiences to deliver personalised recommendations, and Microsoft and Samsung are working on AI-driven features to help people find content that suits their taste through streaming services on their smart TVs.

Greater investment and innovation in UK-based AI development can drive a virtuous circle of partnership and collaboration. The key technical insight driving the AI era today is that approaches leveraging more computation and more data yield the best results.

Thus, opening access to data will remain a priority for developers, including through deals for access to data that would be otherwise inaccessible (e.g. behind a paywall, in a media companies' archives, or content associated with a "do not train" preference expressed in a Robots.txt file or meta tag).

The market also continues to evolve to remove transaction costs and other barriers to partnerships. For example, Microsoft is piloting a Publisher Content Marketplace programme that aims to make it easier for publishers who make otherwise inaccessible content available for use in AI products and receive compensation. Many other companies are also working on marketplaces and other infrastructure to streamline licensing (such as Tollbit).[19]

As market developments highlight, a foundation for licensing already exists. Government could play an enabling role by making information about controls for opting out content more easily available.

TDM exceptions remain necessary and complementary

As evident from what exists in the market, the introduction of a TDM exception will still enable rightsholders and publishers who do wish to actively commercialise their works to do so in myriad ways and create new revenue opportunities.
That said, content available for acquisition and licensing will remain a small fraction of websites, books, or other works that are needed to train today's high performing models and that will enable better models in the future. Models like GPT-4 are trained on a volume of works equivalent to around 14.4 billion news articles and LLAMA 4 on around 60 billion news articles (estimating 500-750 words per article); meanwhile, the New York Times' archive is approximately 13 million articles. The vast majority of all

[17] https://www.perplexity.ai/hub/blog/introducing-comet-plus
[18] https://runwayml.com/news/runway-partners-with-lionsgate
[19] https://docs.tollbit.com/publisher-marketplace/

copyrighted works are not actively managed and licensed. For instance, most of the 1+ billion websites that exist are not actively managed. AI developers cannot readily identify and contact the relevant owner, or find the author (and thus rightsholder) for a comment left on a site. Most books from the 20th century are still in-copyright, but are out-of-print and lack clear ways to identify and get licenses from the rightsholder.

Requiring licenses would leave these works inaccessible, benefiting neither rightsholders nor innovation, as AI models decline in performance and increase in cost.

5. What lessons can be drawn from the approaches taken to balancing the interests of rightsholders and AI developers in other jurisdictions?

Japan, the U.S., Singapore, Israel and the EU already provide broad exceptions that allow text and data mining, with other countries considering adopting similar approaches. These approaches are consistent with the purpose and function of copyright, and include protections for rightsholders.

Japan considers AI development or other forms of data analysis to be acceptable without permission because those types of activities are not considered an "exploitation of the work for enjoyment of the thoughts or sentiments expressed in the copyrighted work."

Singapore amended its copyright law in 2021 to include an exception for making and retaining copies of lawfully accessed works for purposes of computational data analysis. The exception applies to both commercial and non-commercial uses. The exception contains safeguards to protect copyright owners' legitimate interests, including requiring "lawful access" to works used in covered analysis.

In the US, the doctrine of fair use has been interpreted to permit large scale analysis of copyrighted works without permission. Courts have applied it to cases on reverse engineering software and more recently to cases involving text and data mining of millions of digitised books. Most recently courts have explicitly confirmed fair use applies to text and data mining on large book collections for the purpose of training models.[20]

The United States is not alone in adopting fair use. Israel, Liberia, Malaysia, the Philippines, Singapore, South Korea, Sri Lanka, and Taiwan have similar provisions in their laws. Unfortunately, the UK's laws lack a clear exception which undermines research and innovation in AI and generates legal uncertainty and litigation.

17 December 2025

[20] Kadrey v Meta Platforms Inc (N.D. Cal June 25, 2025) ; Bartz v Anthropic PBC (N.D. Cal. June 23 2025)