February 25th, 2025

Written Evidence Submission on Copyright and Artificial Intelligence Open Consultation

The Center for Data Innovation welcomes the opportunity to submit comments to the UK government's Intellectual Property Office, Department for Science, Innovation and Technology, and Department for Culture, Media and Sport on its consultation on proposed changes to the UK's copyright regime, the Copyright, Design, and Patents Act (CDPA) 1988.[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. Its mission is to formulate and promote pragmatic public policies designed to maximise the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence (AI), and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

EXECUTIVE SUMMARY

The UK government seeks views on four proposed legislative options related to copyright and AI to reconcile three objectives: supporting right holders' control of their content and ability to be remunerated for its use; supporting the development of world-leading AI models in the UK by ensuring wide and lawful access to high-quality data; and promoting greater trust and transparency between the sectors. The four proposed options are: do nothing ("Option 0"); strengthen copyright requiring licenses in all cases ("Option 1"); introduce a broad data mining exception ("Option 2"); or introduce a data mining exception which allows rights holders to reserve their rights, underpinned by supporting measures on transparency ("Option 3").

The consultation notes Option 3 is the government's preferred option. However, Option 3 is the most problematic option because it would not provide the legal certainty the government seeks, and it would penalise AI firms and others reliant on commercial TDM. Moreover, Option 3 would contradict the fundamental spirit of UK copyright law and fail to deliver a much-needed competitive advantage for UK AI development.

We recommend that the UK government carry forward Option 2 for four reasons:

1. Option 2 provides the most legal certainty to allow for lawful data scraping and data mining in the least burdensome way possible;
2. Option 2 strikes the optimal balance on incentivising creativity and fostering innovation;

[1] UK Government, Copyright and Artificial Intelligence Open consultation, 17 December 2024.

3. Option 2 ensures both commercial and non-commercial organisations are treated equally; and
4. Option 2 provides the UK with a competitive advantage over other regimes like the United States and European Union by offering predictability and lowering barriers to development.

OPTION 2 PROVIDES THE MOST LEGAL CERTAINTY TO ALLOW FOR LAWFUL DATA SCRAPING AND DATA MINING IN THE LEAST BURDENSOME WAY POSSIBLE

AI firms need access to high-quality datasets to train and refine models. They may gather these datasets themselves from publicly available online sources or acquire these datasets from commercial databases. UK copyright law explicitly permits data scraping and data mining for non-commercial purposes, but it does not have a similar carve out for commercial purposes. The CDPA does have exceptions for temporary copies (section 28A), which search engine companies rely on as the legal basis for indexing content.

However, firms relying on this exception for data scraping and data mining operate in a legal grey area because there are few legal precedents establishing the parameters of section 28A. Option 2 therefore provides the most legal certainty by explicitly articulating a legal basis for commercial and non-commercial text and data mining (TDM) with the fewest regulatory barriers. This clarification would permit AI firms to train models, as well as allow data scraping for other valuable purposes, such as fraud detection, credit scoring, marketing, and search.

In contrast, Option 0 would keep the law as it stands, which would not remove the legal uncertainty. Option 1 would require licensing in all cases, which would provide legal certainty, but would also unduly burden both commercial and non-commercial actors. Option 3 takes the worst of both Option 0 and Option 1: it would make rights holders and data users reliant on untested opt-out technical solutions, which would raise compliance costs, limit data availability for UK firms, and leave many legal questions unresolved until court cases or secondary legislation determine the exact parameters of an opt-out mechanism. Moreover, fully developing Option 3 would take years, but to be competitive, the UK technology sector needs legal clarity now in UK copyright law permitting both commercial and non-commercial TDM.

OPTION 2 STRIKES THE OPTIMAL BALANCE ON INCENTIVISING CREATIVITY AND FOSTERING INNOVATION

The purpose of copyright is to encourage creativity and innovation by granting creators exclusive rights over their works.[2] To achieve this purpose, copyright law must balance the interests of creators and the public—protecting the incentives to invest in creative industries while ensuring that copyright does not become a barrier to knowledge sharing, future creativity, and innovation.[3]

[2] The UK Copyright Service, "A brief history of copyright" [Accessed Feb 18 2025]. Blair, Tony. Copyright, Designs and Patents Bill [Lords], HC Deb 28 April 1988 vol 132 cc525-99.
[3] Hargreaves, Ian. "Digital Opportunities: A Review of Intellectual Property and Growth," May 2011.

Option 2 strikes this balance by protecting the economic interests of rightsholders without restricting the learning and creativity others derive from those protected works.

First, Option 2 does not impact the economic interests of creators. The government itself admitted there was almost no quantitative evidence provided during the 2021 consultation on AI and Intellectual Property that creators suffer economic loss when AI models learn from creative works, including on the introduction of a TDM.[4] The mere process of TDM to train an AI model does not deprive a creator of revenue in the way that piracy does because AI training does not reduce market demand for original creative works. Yet, clear evidence shows the positive economic impacts of AI.[5] Option 2 therefore reinforces the right balance between protecting creators' economic interests, and public interest knowledge sharing.

Second, Option 2 does not discriminate between human and machine learning. Indeed, there is no clear case for making a legal distinction between human and machine learning. Some creators argue that the "unlicensed use of creative works for training generative AI is a major, unjust threat to the livelihoods of the people behind those works, and must not be permitted."[6] Yet copyright law has never restricted the use of their works in the exact way that AI is using their works—to learn. Extending copyright to give creators new rights over the knowledge derived from a work would mark a fundamental and unprecedented departure from UK copyright principles.

AI training, at its core, is simply a scaled-up and automated process of learning and inspiration. Musicians do not have the right to prevent other artists from studying their compositions and incorporating similar styles into their own music. Writers cannot prevent future authors from being inspired by their narrative techniques. Painters cannot demand royalties because someone has analysed their brushstrokes and applied similar methods. Copyright protects the expression of ideas, not the ideas themselves—a principle that should allow both humans and AI to learn patterns and structures from copyrighted materials and then generate new outputs. If copyright law does not penalise human learning, it should not penalise AI learning.

Option 2 avoids introducing a new, inconsistent category of protection on learning that Option 3 would introduce. Option 3 raises questions over what constitutes lawful learning by selectively granting new rights over AI training whilst leaving other learning of creative works untouched. For example, a firm that has always lawfully used TDM to improve its language translation services—such as by analysing large collections of multilingual texts to build a translation database— may now risk violating copyright law simply by incorporating AI into the process. Option 3 may require the firm to obtain licenses, even though its core use of TDM has not changed. This inconsistent

[4] UK Government Open consultation on Copyright and Artificial Intelligence Summary assessment of options [Accessed Feb 18 2025].
[5] Google Economic Impact, "Google UK Economic Impact report: Fuelling Innovation across the UK", Google [Accessed Feb 19 2025].
[6] Statement on AI training [Accessed Feb 19 2025].

treatment unfairly penalises AI-driven learning, stifling innovation and creating compliance burdens where none previously existed.

Finally, by carving out AI training as a distinct category of learning that requires permission, Option 3 sets a precedent for expanding copyright control over other learning-based activities. For example, a UK publisher could argue for similar rights over the sale of online study materials that quote or analyse excerpts from their book, even though use of the works falls within the long-standing fair dealing exception of research and study. Option 3's selective expansion of copyright control risks creating incoherent legal distinctions over what type of learning is lawful. It would also deter investment into UK-based AI research and development by making it harder for UK AI firms to train AI models using public datasets and reducing their generative creativity. This outcome directly contradicts the government's broader aim of making the UK a global leader in AI and digital innovation. Option 2 therefore preserves the traditional concept of learning under copyright law while expanding it in a way that aligns with that tradition.

OPTION 2 ENSURES BOTH COMMERCIAL AND NON-COMMERCIAL ORGANISATIONS ARE TREATED EQUALLY

The CDPA exempts TDM for non-commercial purposes. Option 3 would expand the TDM exception to also apply to commercial purposes, subject to any opt-outs expressed by rights holders. Such a regime would create a two-tier system whereby non-commercial purposes are treated more favourably than commercial ones. Option 2, however, would remove any distinctions between commercial and non-commercial purposes, avoiding an artificial and counterproductive distinction that assumes that commercial actors are less deserving of TDM or that they pose a greater risk to copyright holders. This assumption is flawed for three reasons.

First, commercial AI development often serves the public good just as much, if not more, than non-commercial efforts. Many of the most impactful AI-driven advancements—such as medical diagnosis tools, financial fraud detection systems, and climate modelling technologies—are developed within commercial enterprises. Restricting their ability to train AI on publicly available data would limit technological progress that benefits society as a whole. This lack of progress is already seen in the EU with the Data Governance Act (DGA). The DGA encourages data for public good initiatives by enhancing access to data, but only amongst non-profit organisations. This cuts off the vast majority of for-profit enterprises that have historically driven technological progress for broader public benefit. AI is no different.

Second, public and private sector AI development are deeply interconnected. Many university research projects, which would qualify as non-commercial, receive funding from industry partnerships, making it difficult to neatly separate the two. For example, the Alan Turing Institute collaborates with the private sector to apply AI in fields such as healthcare and energy.[7] Under a

[7] Wood, Beth. "The Alan Turing Institute launches a strategic partnership with Roche to generate insights into disease, patient, and outcome heterogeneity using advanced analytics", The Alan Turing Institute, June 22, 2021; Future

two-tier system, these collaborations could face legal ambiguity, discouraging partnerships that drive innovation, and decoupling non-profit research from its commercial applications which would otherwise bring widespread prosperity to the UK. Such a distinction may question the efforts of non-profit data aggregators like Common Crawl, who offer free accessible repositories of web scraped data to all, including commercial entities.[8] The requirement of "commercial purpose" would likely place stronger burdens on these non-profits who would have to introduce checks on how entities might use their data to avoid falling foul of the legislation.

Finally, the UK government's broader policy goals emphasise reducing regulatory barriers, particularly in data-driven industries. Creating a rigid distinction between commercial and non-commercial TDM risks introducing legal uncertainties akin to those seen under the General Data Protection Regulation (GDPR)—organisations struggle to make full use of data due to GDPR's purpose limitation that restricts secondary uses of data. This lack of legal justification has led to compliance costs and barriers to innovation—problems the government is now actively trying to fix through the Data Use and Access Bill, which seeks to enable secondary uses of data to reduce unnecessary regulatory burdens. If the UK were to impose a two-tier TDM system, it would likely face similar challenges, requiring future interventions to mitigate unintended consequences. The delay until such fixes are made would create uncertainty, deterring investment in UK-based AI research and development.

Option 3 would unjustly penalise AI development for commercial purposes, creating an anti-business regulatory environment for AI firms in the UK. Ensuring a clear and consistent framework for AI training under Option 2's broad TDM would strengthen the UK's appeal as a destination for AI research and development.

OPTION 2 PROVIDES THE UK WITH A COMPETITIVE ADVANTAGE OVER OTHER REGIMES LIKE THE UNITED STATES AND EUROPEAN UNION BY OFFERING PREDICTABILITY AND LOWERING BARRIERS TO DEVELOPMENT

Copyright protection is limited to the jurisdiction in which copyright is granted. If the UK aims to position itself as a global leader in AI innovation, it should adopt a copyright framework that is not only competitive with but preferable to those in other major AI-developing regions. A permissive approach—such as a broad TDM exception without an opt-out—would provide the UK with a distinct advantage in attracting AI investment and fostering innovation, particularly as the UK has such a strong creator economy.

Currently, AI developers gravitate toward jurisdictions with fewer regulatory barriers, particularly regarding access to data. The United States exemplifies this with its fair use doctrine, which allows limited use of copyrighted material without permission under certain conditions, including

Energy, "ESO and The Alan Turing Institute use machine learning to help balance the GB electricity grid", National Energy System Operator, July 25, 2019.

[8] https://commoncrawl.org/

transformative use, the amount of material used, and the effect on the market. While fair use is more permissive than most other copyright regimes, it is also highly unpredictable, as its application depends largely on court rulings rather than clear statutory provisions. For example, in *Authors Guild v. Google*, the U.S. courts agreed Google's mass digitisation of books was fair use, but each case is assessed on its specific facts, leading to legal uncertainty and potential litigation risks for AI developers.

By contrast, the European Union takes a much more restrictive approach. Article 4 of the Directive on Copyright in the Digital Single Market provides an explicit TDM exception but allows rights holders to opt out, effectively limiting the scope of AI-relevant data. While this offers greater legal certainty than fair use, it significantly restricts access to data, raises compliance costs, and discourages AI development within the EU. The UK government's proposed Option 3 mirrors this restrictive EU model, and thus would create similar barriers to AI innovation.

The UK has the chance to address two key hurdles to current AI development: legal uncertainty, and data access. Option 2—a broad TDM exception without an opt-out—would position the UK as a more attractive destination for AI investment. It would offer a statutory foundation for AI development that is clearer and more predictable than the U.S. fair use system while being significantly less restrictive than the EU's opt-out model. This legal certainty would encourage companies to base their AI research and model training in the UK, fostering a more dynamic and competitive domestic AI ecosystem.