

## Comments of the International Center for Law & Economics

*UK Intellectual Property Office Copyright and AI Consultation*

*25 February 2025*

**Authored by:**

**Ben Sperry** (Senior Scholar, Innovation Policy, International Center for Law & Economics)

**Kristian Stout** (Director of Innovation Policy, International Center for Law & Economics)

## Introduction

Thank you for the opportunity to submit comments regarding the Intellectual Property Office's (IPO) consultation on copyright and artificial intelligence (AI).<sup>1</sup> The International Center for Law & Economics (ICLE) is a nonprofit, nonpartisan research centre with a roster of more than 50 academic affiliates and research centres from around the globe. Our mission is to promote the use of law & economics methodologies to inform public-policy debates, including those involving the subject of intellectual property. Our scholars have produced significant research on issues related to AI, including its interaction with copyright law and competition policy.

These comments offer a law & economics background that we hope will help the IPO to weigh the tradeoffs involved in balancing the need to protect the rights of copyright holders with promoting the growth and development of AI. We have concerns that the consultation's proposed approach to copyright and AI will unduly inhibit the growth of the AI sector in the UK by focusing on inputs without considering the broader nascent market that AI outputs can provide. In this broader market, the long-term interests of creators will likely be enhanced by providing new opportunities for commercialization that do not currently exist.

### I. The Economics of Copyright and Innovation

Copyright is an important tool to stimulate innovation by fostering incentives, in the form of exclusive property rights, for authors to invest in the production of creative works. Copyright holders can then use those rights to prevent commercial free riding by other entities. The economic justification for copyright is that promoting creative output enhances social welfare in the long run.<sup>2</sup> Because creative works are otherwise non-excludable and non-rivalrous, they are a public good that would be underproduced in the absence of copyright, as they would be too easily copied, thereby reducing the expected value of production.

On the other hand, copyright law does give the rightsholder some degree of market power, which includes the ability to raise prices and hold up others' ability to use copyrighted material. Thus, term limits and exceptions for fair use or fair dealing exist to provide some release valve for exceptional public-interest considerations. The need to grant both creators the ability to be compensated for their work *and* others the ability to use and enjoy that work (as well as the ability to realize the social benefits of fostering technological progress) is the basic tension inherent in copyright law.

This fundamental tension creates what we might call a 'hydraulic system'; when pressure is applied in one area of copyright protection, it necessarily creates corresponding effects elsewhere in the

---

<sup>1</sup> *Copyright and AI: Consultation*, UK INTELLECT. PROP. OFF. (December 2024), available at [https://assets.publishing.service.gov.uk/media/6762c95e3229e84d9bbde7a3/241212\\_AI\\_and\\_Copyright\\_Consultation\\_presentation.pdf](https://assets.publishing.service.gov.uk/media/6762c95e3229e84d9bbde7a3/241212_AI_and_Copyright_Consultation_presentation.pdf) [hereinafter 'Consultation'].

<sup>2</sup> For more on the economics of copyright, see Brent Luches, *Introduction 1-3*, in IDENTIFYING ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR COPYRIGHT POLICY (U.S. Copyr. Off., February 2025), available at <https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>.

system. Just as with actual hydraulics, where compressing fluid in one chamber causes movement in another, strengthening creators' rights in one domain may require greater flexibility in another to maintain the system's balance.

This hydraulic nature is particularly evident when disruptive technologies like AI challenge traditional frameworks. If we restrict AI systems at their input stage by limiting the materials upon which they can train, we may need to provide more flexibility at the output stage, or vice versa. Understanding this hydraulic relationship is crucial when considering how copyright law should evolve to accommodate AI technologies, while still fulfilling its foundational purpose.

Recent technological changes have introduced new challenges for policymakers seeking to strike the proper balance between promoting the production of creative works and permitting the consumption and use of copyrighted material as an input for further production. Allowing copyrighted material to be used as an input in AI models could, over the long term, change the incentives for creative output. But granting rightsholders the ability to reserve all rights could also unduly hold up AI developers' ability to make their models to be as useful as possible. While AI is the most recent challenge to the paradigm of copyright law, it is necessary to keep some basic principles in mind.

First, copyright protects the expression of ideas, not the underlying ideas themselves. Expressing underlying ideas in a different manner than a copyright holder has does not normally violate copyright law. The fair-use exception in the United States, for instance, arguably allows AI developers to use many copyrighted materials as inputs for their models. This does, however, remain a subject of controversy within U.S. legal circles. The many good-faith arguments raised on all sides of that debate demonstrate that the *sui generis* case of AI training leaves it far from clear how traditional copyright principles ought to be applied to this innovative technology.<sup>3</sup>

Second, requiring AI developers to obtain individual licenses for all pieces of copyrighted material used in AI training would pose enormous practical challenges and generate tremendously high transaction costs.<sup>4</sup> A single large AI model might be trained on billions of text snippets or images drawn from across the internet. Negotiating a separate license for each copyrighted work (or a blanket license with each rightsholder) in such a corpus would entail thousands or even millions of transactions.<sup>5</sup> The time and administrative overhead of this process would make it effectively impossible to license everything at the necessary scale. Thus, transaction costs would be 'prohibitively

---

<sup>3</sup> See, e.g., Kristian Stout, *AI Training Is Not Fair (According to One Court)*, TRUTH MARK. (11 February 2025), <https://truthonthemarket.com/2025/02/11/ai-training-is-not-fair-according-to-one-court>.

<sup>4</sup> See RICHARD A. POSNER, *ECONOMIC ANALYSIS OF LAW* 42 (7th ed. 2007), (discussing the transaction costs involved with copyright as including the tracing costs of identifying the copyright holder and negotiation costs of negotiating the license with the copyright holder).

<sup>5</sup> See Jorge Padilla & Kadambari Prasad, *Demystifying Licensing Debates: Should GenAI Developers Pay to Train Their Models on Copyright Protected Content?*, COMPASS LEXECON (25 February 2025), <https://www.compasslexecon.com/insights/publications/demystifying-licensing-debates-should-genai-developers-pay-to-train-their-models-on-copyright-protected-content>.

high if developers had to agree a licence with each and every copyright holder individually’, given the sheer number of works involved.<sup>6</sup>

Because individual work-by-work licensing would be so unwieldy, some experts have suggested collective-licensing models as a possible solution.<sup>7</sup> The idea is to have a central entity or clearinghouse negotiate blanket licenses on behalf of large groups of rightsholders, similar to how music performing-rights societies operate. The main appeal of such collective licensing is that it would lower transaction costs by pooling rights.<sup>8</sup> In theory, this approach could make access to training data more efficient and ‘affordable’.<sup>9</sup> Significant hurdles would, however, remain. First, not all content owners may join a collective, leaving gaps in the coverage.<sup>10</sup> Sources of AI training data are highly diverse and diffuse; a collective-licensing regime might cover only certain classes of creative works (e.g., those owned by major publishers or contained in large image libraries) but miss independent creators who aren’t represented.<sup>11</sup>

Second, if collective licensing became mandatory, it could introduce its own concerns. For instance, competition experts warn that forcing all licensors into a single pool might create a monopoly, eliminating competitive pressure among content suppliers.<sup>12</sup> In the absence of clear market failure, businesses might prefer voluntary solutions, as can be seen with some media companies already striking deals.<sup>13</sup> In practice, we are seeing a mix of approaches: some large content owners have, indeed, banded together (or opted out of free use) to demand licenses, while many smaller creators remain outside any collective mechanism.<sup>14</sup> This patchwork translates into uncertainty, while high costs persist for AI developers who try to assemble training datasets lawfully.

Third, calculating license fees will be very difficult. AI models use billions of inputs that contribute to an output. Determining the value of an input at the training stage may prove impossible.

---

<sup>6</sup> *Id.*

<sup>7</sup> *Id.*

<sup>8</sup> *Id.*

<sup>9</sup> *Id.*

<sup>10</sup> See Michael D. Smith & Rahul Telang, *The Effect of AI Ingestion on Rightsholders’ Incentives* 35-38, in IDENTIFYING THE ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR COPYRIGHT POLICY (U.S. Copyr. Off., February 2025), available at <https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf> (discussion of the limitations of collective licensing for AI training).

<sup>11</sup> See *id.* at 37.

<sup>12</sup> Padilla & Prasad, *supra* note 5; see also Press Release, *Runway Partners with Lionsgate in First-of-Its-Kind AI Collaboration*, LIONSGATE (18 September 2024), <https://investors.lionsgate.com/news-and-events/press-releases/2024/09-18-2024-140126979>.

<sup>13</sup> Padilla & Prasad, *supra* note 5.

<sup>14</sup> See, e.g., Christophe Geiger, *To Pay or Not to Pay (For Training Generative AI), That Is the Question*, INTELLECTUAL PROPERTY JOTWELL (18 December 2023), <https://ip.jotwell.com/to-pay-or-not-to-pay-for-training-generative-ai-that-is-the-question> (reviewing Martin Senftleben, *Generative AI and Author Remuneration*, 54 INT’L REV. INTELL. PROP. COMPETITION L. 1535 (2023)).

Moreover, given the sheer volume of input data needed, the marginal value of any piece of content—even content that may be quite valuable in other contexts—may be extremely small.

Once an output is generated and used commercially, it may be possible to assign value in novel output markets. Determining how much any given input contributed to that output, however, will depend on the degree of similarity. Such a system could theoretically work in much the same manner as YouTube’s ContentID system.<sup>15</sup> Further, there are likely new avenues for commercialization that would become available to creators if they were allowed to bargain with AI producers for the use of their name, image, and likeness in the content-generation stage.<sup>16</sup> But both of these potential solutions, as well as many yet-unconceived solutions, could be stymied by an overly aggressive application of traditional copyright principles to this new technology.

In sum, focusing on the input side may upset the balance of copyright law. A better approach would be to offer creators the ability to challenge outputs that are too similar to their own works. Below, we will consider in greater detail how this applies to the IPO’s proposal.

## II. Problems with the Reservation-of-Rights Approach

The IPO proposes ‘a data mining exception which ensures that rights can be reserved, underpinned by developer transparency’.<sup>17</sup> The outline of how this would work is as follows:

- (a) It would apply to data mining for any purpose, including commercial purposes.
- (b) It would apply only where the user has lawful access to the relevant works. This would include works that have been made available on the internet, and those made available under contractual terms, such as via a subscription. This would allow right holders to seek remuneration at the point of access – for example, in the price of a subscription to a library of research material.
- (c) It would apply only where the right holder has not reserved their rights in relation to the work. If a right holder has reserved their rights through an agreed mechanism, a licence would be required for data mining. Possible types of rights reservation, and the extent to which they can be supported using technology, are explored in more detail below.

---

<sup>15</sup> See *How Content ID Works*, YOUTUBE HELP, <https://support.google.com/youtube/answer/2797370?hl=en> (last accessed 25 February 2025).

<sup>16</sup> See Shane Greenstein, *Commercial Exploitation of Name, Image, and Likeness* 24-30, in IDENTIFYING THE ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR COPYRIGHT POLICY (U.S. Copyright Office, February 2025), available at <https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>.

<sup>17</sup> Consultation, *supra* note 1, para. 72.

(d) It would be underpinned by greater transparency about the sources of training material, to ensure compliance with the law and build trust between right holders and developers. Possible approaches to transparency are set out in more detail below.<sup>18</sup>

The legal effect would be to allow rightsholders to reserve their rights and prevent the use of their works for AI training.<sup>19</sup> If a reserved work is copied for an AI model, this would be considered an infringement of copyright law. The intent is to create a market for licensing copyrighted works as inputs in AI models.<sup>20</sup>

The problem with this approach is that it focuses too much on the use of inputs without adequately considering how this nascent technology could enable new modes of monetization for both creators and AI producers. As such, the proposal is more likely to prevent the growth of AI and reduce the social benefits that might flow from its use, while doing little to remunerate creators.

First, allowing rightsholders to reject the use of their materials as AI inputs would grant them a right that potentially exceeds the protections of traditional copyright. Copyright only protects the expression of an idea, not the idea itself. At this point, there is no legal consensus on what these AI systems are doing, even on a conceptual level. An argument can be made that unsanctioned use of copyrighted works in AI training amount to straightforward copyright violations, as the models are ‘memorizing’ creators’ content and using it to create new potentially infringing content.<sup>21</sup> But this is not obviously true, nor does it appear to be the consensus view. For example, research demonstrates that careful curation of data sets can lead to significant (if not total) reductions in the quantity of apparently ‘memorized’ content.<sup>22</sup>

Further, AI systems only make ‘true’ copies of works in the training phase; they do not literally copy them into their model weights. Instead, they analyse the statistical relationships among ‘tokens’ (smaller pieces of text or image chunks within a file) and learn the various patterns that human output takes for billions or trillions of similar input patterns.<sup>23</sup> This process is fundamental to how large language models (LLMs) process language, and it means the model never stores the original text as a readable sequence. Instead, each work in the training data is transformed into tokens and ultimately into abstract ‘weight’ adjustments. By the end of training, the model consists of billions of tuned parameters (essentially a complex array of numbers), or as Lee Gesner as described it, into

---

<sup>18</sup> *Id.* at para. 74.

<sup>19</sup> *See id.* at para. 76.

<sup>20</sup> *See id.* at para. 77.

<sup>21</sup> *See* Katherine Lee *et al.*, *Deduplicating Training Data Makes Language Models Better*, ARXIV (March 2022), <https://arxiv.org/abs/2107.06499>; *see also* Alex Reisner, *The Flaw That Could Ruin Generative AI*, THE ATLANTIC (11 January 2024), <https://www.theatlantic.com/technology/archive/2024/01/chatgpt-memorization-lawsuit/677099/>.

<sup>22</sup> Lee, *supra* note 21, at 1.

<sup>23</sup> Lee Gesner, *Copyright and the Challenge of Large Language Models (Part 1)*, MASS LAW BLOG (1 July 2024), <https://www.masslawblog.com/copyright/copyright-and-the-mechanics-of-large-language-models>, (offering a layman’s discussion of how, before training, AI models convert text into a numerical form through tokenization, breaking language down into small units (words or subwords) which are then represented as numbers).

‘a vast sea of numbers, with no direct correspondence to the original text’.<sup>24</sup> In other words, the model only retains statistical information about language usage.<sup>25</sup> Tokenization enables an AI model to learn the patterns and structure of language without reproducing creative expression in its permanent files.<sup>26</sup>

Thus, in the United States, the fair-use argument remains a live controversy. More broadly, policymakers around the world should pause before rushing too quickly to apply traditional assumptions about either how AI technology operates or the markets it could facilitate.<sup>27</sup> Moreover, while the exceptions for fair dealing in the UK are more limited than fair use in the United States, the underlying logic remains the same: there are otherwise infringing uses of copyrighted material that are permitted as in the public interest.

Second, the reservation-of-rights approach would create large transaction costs that will, in many instances, serve as an impediment to bargaining. While the consultation is hopeful about the possibility of collective licenses,<sup>28</sup> much of the internet content that is scraped for use by AI models is not subject to such management. AI developers would likely be unable to identify all the rightsholders with whom they would be bound to negotiate licenses. Even if they could, as noted above, the transaction costs would be onerous, likely leading to less-capable AI models due to a lack of inputs. Another alternative is that such models would be forced to rely primarily or completely on synthetic data, which could likewise have deleterious effects on model quality.<sup>29</sup>

Third, lacking the ability to assign value at the input stage would greatly reduce AI’s ability to foster a market for licensing. Even assuming that transaction costs are not insurmountable, there are two problems with focusing on input-licensing markets: the low marginal value of works in training sets and the difficulty of assigning value. There is no obvious way to calculate the monetary value of a particular work as an AI input.

This conclusion has been supported by analysis from the U.S. Copyright Office: because foundation models ingest millions or even billions of works, the influence on the model of any one copyrighted work is likely to be so diluted that even a small transaction cost arising from licensing negotiations would exceed the work’s share of the model’s utility.<sup>30</sup> While some data may be more important to

---

<sup>24</sup> *Id.*

<sup>25</sup> John Poulos, *Generative AI: How It Works, Content Ownership, and Copyrights*, INSIDE TECH LAW (24 May 2024), <https://www.insidetechl.com/blog/2024/05/generative-ai-how-it-works-content-ownership-and-copyrights>.

<sup>26</sup> *Id.*

<sup>27</sup> See Stout, *supra* note 3.

<sup>28</sup> See Consultation, *supra* note 1, paras. 94-96.

<sup>29</sup> See Maggie Harrison Dupre, *When AI Is Trained on AI-Generated Data, Strange Things Start to Happen*, FUTURISM (2 August 2023), <https://futurism.com/ai-trained-ai-generated-data-interview>.

<sup>30</sup> See Adam Jaffe, *Controlling the Use of Copyrighted Materials in Training 50*, in IDENTIFYING THE ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR COPYRIGHT POLICY (U.S. Copyright Office, February 2025), available at <https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf>.

a training set than other data, this is only on a relative basis. The entire collection of Roald Dahl's works represents an important contribution to English-language writing, but it remains only a drop in the ocean of the English-language corpus. Even these very valuable properties would be worth a pittance in the context of a vast training set. This makes traditional valuation methods impractical; one can't simply multiply a 'per-work' fee by millions of works without the total cost becoming astronomically high (and divorced from each work's actual impact on the model).

One reason that valuation is difficult is that the monetization in generative AI happens at the output stage, not when the data is ingested. The training data is not sold or consumed directly; rather, value is realized when the model produces useful output (text, images, etc.) for which users will pay, or that otherwise enables a commercial service. The contribution of any given training example is indirect and intertwined with countless others. Thus, without a means to determine the connection between a particular piece of generated content and a piece of content present in a training set, any method of assigning a monetary value to a piece of training data will be highly speculative. For most individual works (especially those by independent creators scattered across the web), there is no clear market rate for 'AI-training usage'. The value is context-dependent and likely *de minimis* on a per-work basis.

Relatedly, when considering restrictions on AI-training data, we must acknowledge the critical importance of dataset diversity to prevent algorithmic bias. In creating a system in which only commercially negotiable properties are readily accessible for training, we risk developing AI models that disproportionately reflect the perspectives, experiences, and cultural contexts of those with market power or established licensing frameworks. This would inevitably lead to AI systems with blind spots, particularly undermining representation of independent creators and non-commercial sources of knowledge or expression.

The reservation-of-rights approach could therefore inadvertently create AI systems that are not only less capable but fundamentally biased toward commercially dominant viewpoints. True innovation and society at-large would both benefit from AI trained on diverse datasets that reflect the full spectrum of human knowledge and creative expression, not just those segments with the resources to participate in licensing markets.

### **III. A Better Way Forward: Focus on Outputs**

A better approach would be to focus on finding solutions on the output side. Such an approach would recognize the 'hydraulics of copyright'—that is, when pressure is applied in one area of copyright law, it creates a pull in another. If we allow broader use of copyrighted works as inputs for AI training, we should counterbalance this by strengthening creators' rights regarding outputs that resemble their works.

When AI models are trained on copyrighted materials, they can produce outputs that are very similar to inputs. The IPO should, of course, consider ways to rebalance copyright law that would grant creators the ability to challenge outputs that are exceedingly similar to protected works. As it is put

in the consultation document, '[c]ontent generated by an AI model will infringe copyright in the UK if it reproduces a "substantial part" of a protected work'.<sup>31</sup>

While there are difficulties in policing AI, 'the copyright framework in relation to infringing outputs is reasonably clear and appears to be adequate'.<sup>32</sup> Promoting transparency in outputs generated by AI and adding additional protections for rightsholders when their creative works are substantially copied in AI outputs, would allow for AI models to grow while promoting the creation of new works.

This approach would acknowledge the tension inherent in copyright's purpose: both protecting creators and encouraging innovation. Training AI systems requires large-scale data ingestion, reflecting the significant potential social benefits that AI may offer; these include democratizing access to information, spurring creativity, and driving technological progress. Rather than restrict this process at the input stage, which could severely hamper AI development, we should focus on the market impact at the output stage and build upon the IPO's concern for AI-generated outputs.

For instance, in some jurisdictions within the United States, there are common-law protections for the use of an individual's name, image, and likeness. If an AI model were to reproduce someone's likeness or voice, this could possibly violate the 'right of publicity'.<sup>33</sup> While a model could train on a variety of voices and images, it could be a violation if the output is too similar to a known person. The IPO should consider how to adopt similar standards for individuals to control their own 'likeness'.

Beyond likeness rights, the IPO could explore revenue-sharing mechanisms tied to AI-generated outputs that substantially resemble copyrighted works, thereby ensuring that creators share in the benefits without stifling innovation. Such mechanisms would acknowledge that, while broad training may fall within fair-dealing exceptions, commercial outputs that compete with original works present a different consideration entirely.

A robust framework will need to evaluate whether the necessity of large-scale use of copyrighted works for AI training outweighs potential harms to creators, while simultaneously developing mechanisms to ensure fair compensation when outputs closely resemble original works. This approach would seek to preserve the foundational purpose of copyright: to protect creators while encouraging progress and innovation.

#### **IV. Conclusion**

The reservation-of-rights approach proposed by the IPO, while well-intentioned, risks undermining both the development of AI technology and the long-term interests of creators. By focusing primarily on input restrictions, this approach misunderstands the fundamental economics of copyright in the

---

<sup>31</sup> Consultation, *supra* note 1, para. 157.

<sup>32</sup> *Id.* at para. 161.

<sup>33</sup> See Greenstein, *supra* note 16, at 24-30.

context of generative AI and fails to account for the need to balance pressures throughout the copyright system.

As we have detailed, the proposed input-focused framework presents several critical problems:

- It potentially expands copyright beyond its traditional scope of protecting expression rather than ideas, applying old frameworks to a technology that functions in fundamentally different ways than traditional human consumption of creative works.
- It introduces prohibitively high transaction costs that would prevent effective bargaining between AI developers and the vast number of rightsholders whose works appear online, ultimately creating market inefficiencies.
- It overlooks the practical impossibility of calculating accurate values for individual works used as inputs in massive training datasets, where even culturally significant works represent only a tiny fraction of the total data.
- It risks embedding algorithmic bias by limiting training to ‘commercially negotiable’ properties, thereby potentially creating AI systems that reflect only the perspectives of those with market power and established licensing frameworks.

A more effective and balanced approach would preserve the essential purpose of copyright law—promoting both creative production and innovation—by focusing regulatory attention on outputs, rather than inputs. This would allow AI models to train broadly on diverse datasets while creating stronger mechanisms for creators to challenge, control, or be compensated for outputs that substantially resemble their protected works.

If it were to adopt this output-focused framework, the UK would seize the opportunity to position itself as a leader in AI innovation, while still respecting and protecting creators’ rights. Such an approach would better serve the public interest. It would enable a flourishing of the expected social benefits of AI technology—democratizing access to information, spurring creativity, and driving technological progress—while ensuring that creators can participate meaningfully in these new markets.

The future of copyright in the age of AI requires thoughtful recalibration, rather than restriction. We urge the IPO to consider how the hydraulics of copyright might be better balanced by strengthening creators’ rights at the output stage, while allowing for the broad training necessary for AI to realize its full potential for society.