Written evidence submitted To the Data (Use and Access) Public Bill Committee by Dr Sabine Jacques & Joseph Savirimuthu (DUAB32).

[1] This response has been written by Dr Sabine Jacques (Senior Lecturer in Intellectual Property Law and project co-lead on the UKRI/AHRC Creative Cluster, Music*Futures*) and Joseph Savirimuthu (member of the University of Liverpool's Cybersecurity Institute and recently conducted the AISC Masterclass, Joseph presents internationally on privacy, data protection, and emerging technologies).

[2] The Government's proposed measures regarding transparency obligations for the use of copyright works in training generative AI models, particularly in the context of the Data Use and Access Bill ('Bill'), present significant challenges for the UK's legal framework. This written evidence seeks to aid the House of Commons Committee in its scrutiny of the Bill by highlighting key considerations that warrant careful attention with regard to proposed amendments (61, 62, 63, 64 alongside amendment 44A) and 65, which seek to enhance transparency obligations for general-purpose AI models.

[3] The government's preference for implementing a pro-innovation approach to copyright policy raises two important concerns. First, established copyright law checks and balances may be reinterpreted to serve political priorities rather than their original purpose. Second, the emphasis placed on transparency as a regulatory mechanism to balance the tensions may underestimate the nuances of the interplay between copyright law and the AI lifecycle. For present purposes, the key points set out in this written evidence will be confined to addressing are as follows:

- The need to clarify the nature of the transparency obligations so that the multifaceted nature and complexity of AI data-driven decision making is fully captured;
- Understanding the challenges posed by AI systems and ethical requirements helps identify where the checks and balances already provided by copyright law may need adaptation.
- In general, the proposed transparency measures should be interpreted consistently with existing copyright principles and international obligations.

[4] Before examining the impact of these amendments on copyright law, it is important to outline key copyright concerns related to generative AI in the context of the current legal framework, the government's proposed changes, and the interplay with the Data Use and Access Bill.

Copyright and the AI Lifecycle

Phase 1: Input Phase - A Multifaceted Understanding

[5] The first phase of AI development anticipated by the proposed reforms under the Bill involves two key steps: (1) pre-training activities, such as web scraping, to collect data from

the internet, which is then stored in a large database for future training purposes, and (2) the training process itself, during which the AI model analyses the collected data to identify patterns and relationships. This process relies on machine learning and deep learning techniques to enable the model to make predictions about what should follow in a sequence. Adopting a multifaceted understanding of the use of this emerging technology is essential for developing effective transparency requirements. Such an approach offers several advantages:

[6] First, it provides a broader scope that encompasses the variety of data collection and training methodologies employed in generative AI development. Rather than focusing narrowly on specific techniques or platforms, this perspective recognises the diverse ways in which copyrighted works are incorporated into training datasets, going beyond simple technical distinctions to consider the qualitative aspects of how works are used.

[7] Second, it enables a more comprehensive governance focus by expanding regulatory considerations beyond just the technical aspects of data mining. This approach incorporates critical characteristics such as the potential impact on creative markets, the coherence with existing copyright frameworks, and the balance between innovation and rights protection. Such governance considerations are particularly relevant when assessing whether activities fall under the text-and-data mining exception outlined in section 29A of the Copyright, Designs and Patents Act (CDPA) or any new future text-and-data mining exception.

[8] Third, it acknowledges the inherent uncertainty and ambiguity that characterise emerging technologies like generative AI.[1] This recognition is crucial for policymakers as they navigate what might be described as a Collingridge Dilemma – the challenge of establishing appropriate regulations when the technology's full implications remain unclear yet delaying regulation until those implications become evident risks entrenching problematic practices.[2]

[9] Under current UK copyright law, some training activities may fall under the text-and-data mining exception in the CDPA. While this exception broadly encompasses pre-training and training activities necessary for AI models' development for non-commercial use, there is considerable uncertainty as to whether the proposals under the Bill will extend to all copyright-relevant acts in Phase 1. This uncertainty exemplifies why transparency requirements must be aligned with existing copyright principles and standards to address evolving technical methodologies and emerging legal interpretations.

[10] The UK government plans to introduce a new text-and-data mining exception for commercial use by AI models, with the provision for copyright owners to opt-out, likely through machine-readable means for publicly available online content (such as metadata and

---

[1] See The Alan Turing Institute—written evidence (LLM0081) to the House of Lords Communications and Digital Select Committee inquiry: Large language models

[2] Huw Roberts and others, 'Artificial Intelligence Regulation in the United Kingdom: A Path to Good Governance and Global Leadership?' (2023) 12 Internet Policy Review < https://policyreview.info/articles/analysis/artificial-intelligence-regulation-united-kingdom-path-good-governance > accessed 4 March 2025.

terms and conditions of websites). If these proposals are implemented, it will be similar to the approach taken by the EU.[3] That said, it is important to note that there are compelling arguments suggesting that the EU's text-and-data mining exception for commercial purposes was never intended to apply to AI development.[4]

[11] The *Kneschke v LAION* case in Germany is particularly relevant to this discussion.[5] This case examined whether the use of a copyright-protected photograph by Kneschke to train generative AI models was lawful. LAION, a non-profit organisation, provides datasets, tools, and models for machine learning research, which are subsequently used by commercial AI providers like Stable Diffusion. The court ruled in favour of the defendant, recognising that LAION, as a research organisation, was operating within the scope of both the EU's text-and-data mining provisions for non-commercial and commercial purposes. This ruling identifies potential shortcomings in the EU AI Act's copyright provisions, which the UK should keep in mind.

Phase 2: Processing Phase - Emerging Technology Challenges

[12] In the second phase, the AI model applies the 'knowledge' it gained during training to generate new content. This phase exemplifies why generative AI should be understood as a set of potentially transformative innovations at various stages of development, characterised by radical novelty that poses unique challenges for copyright frameworks.[6]

[13] For example, the processing phase introduces distinct legal questions that traditional copyright frameworks struggle to address. Primarily, these include whether the model's internal weights and configurations qualify as 'protected databases', and whether any 'memorisation' of protected content within the model constitutes unauthorised reproductions. These questions illustrate the radical novelty of generative AI technology, signalling that existing laws and regulations may be inadequate or inapplicable when addressing how copyrighted works are processed within AI systems. Furthermore, the relationship between the input and processing phases demonstrates why transparency obligations must acknowledge the varying stages of technological development. While input-phase activities like dataset creation have relatively clear analogues in existing copyright frameworks, processing-phase activities represent more advanced stages of development where the technology begins to transform the very nature of how creative works are utilised. This technological progression necessitates adaptive regulatory approaches that can evolve alongside the technology. Finally, the 'black box' nature of many AI systems further

---

[3] See João Quintais, 'Generative AI, Copyright and the AI Act' (2025) 56 *Computer Law & Security Review* 106.

[4] See for example news report: https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act

[5] Case No. 310 O 227/23, https://pdfupload.io/docs/4bcc432c

[6] Devyani Gajjar and Lois Jeary, 'Artificial Intelligence and New Technology in Creative Industries' < https://post.parliament.uk/artificial-intelligence-and-new-technology-in-creative-industries/ > accessed 4 March 2025.

characterises this emerging technology challenge. Even developers may not fully understand how specific inputs influence the model's internal representations.[7] This opacity creates fundamental difficulties for transparency as a governance instrument: meaningful disclosure becomes problematic when the relationship between inputs and internal processing remains technically inscrutable. The transformative nature of these technologies thus disrupts conventional understandings of how creative works contribute to subsequent creations.

[14] These characteristics of emerging technology in the processing phase have significant implications for the Government's proposed transparency measures. While disclosure requirements concerning training datasets might adequately address input-phase transparency, they provide little insight into processing-phase activities. Any comprehensive transparency framework must therefore acknowledge this limitation and consider alternative approaches to providing right-holders with meaningful understanding of how their works contribute to AI capabilities, without imposing technically impossible disclosure requirements.

[15] Consequently, the novelty of AI processing mechanisms suggests the need for creative legal thinking, rather than prescriptive procedural processes to grapple with issues that now transcends traditional copyright concepts of reproduction and derivation. Rather than attempting to reconfigure the challenges through the prism of the Bill and artificially skew these new technologies into existing legal categories, transparency obligations might better serve right-holders by focusing on what is both technically feasible and meaningfully informative about how protected works influence AI systems' capabilities and outputs.

Phase 3: Output Phase - Regulatory Complexity and Legal Uncertainty

[16] The output phase of the AI lifecycle presents perhaps the most visible and contentious intersection of copyright law, generative AI technology and the Bill. This phase involves the AI system producing new content based on its training and processing, introducing a complex matrix of legal questions that traditional copyright frameworks will now have to address and which the proposed amendments fail to fully grapple with, in any form of clarity or coherence. The primary question confronting policymakers and courts is whether AI-generated outputs qualify for copyright protection. This fundamental issue is currently under active consideration in the government's AI & copyright consultation, reflecting its centrality to future regulatory frameworks. However, such a question extends beyond simple categorisation to touch on foundational copyright principles about the nature of authorship, creativity, and the purpose of copyright protection itself. One regulatory implication is that transparency obligations must account for this uncertainty by facilitating meaningful disclosure about how the AI's outputs relate to its inputs, without presuming answers to these unresolved legal questions.

[17] A second layer of complexity emerges when considering whether AI-generated content constitutes derivative works based on copyright-protected materials used during the input

[7] Alan Turing Institute submission (fn 1).

phase. This question becomes particularly challenging due to the opacity of the processing phase discussed earlier. Even with perfect transparency about training data, establishing a direct lineage between specific inputs and outputs remains technically problematic.[8] This technical reality creates a fundamental tension in any transparency framework: how can meaningful information be provided about relationships that are, by their nature, difficult to trace or quantify?

[18] The potential for copyright infringement in AI-generated outputs and establishing violations further complicates the regulatory landscape. Current legal frameworks offer several potentially applicable exceptions, including those for quotation, criticism, review, parody, pastiche, or caricature. However, these exceptions were developed for human creative processes and may not translate effectively to AI contexts where the relationship between source materials, generative content and outputs differs fundamentally from traditional creative adaptation. Transparency obligations must therefore carefully navigate this uncertainty by providing right-holders with sufficient information to evaluate potential infringement, without imposing technically unfeasible requirements on AI developers.

[19] The ongoing UK legal case of *Getty Images v Stability AI*, scheduled for hearing this summer, exemplifies these challenges. This case will likely establish crucial precedents regarding how existing copyright frameworks apply to generative AI outputs, particularly concerning visual content that bears similarities to copyright-protected images in training datasets and the application of existing copyright exceptions to AI models.

Summary and Recommendation

[20] For transparency obligations to function effectively in this context, they must address the entire AI lifecycle rather than focusing exclusively on any single phase. The interconnected nature of input, processing, and output phases means that meaningful transparency cannot be achieved by considering each phase in isolation, as the proposed reforms presume. Instead, a comprehensive approach must provide right-holders with insights into how their works contribute throughout the AI lifecycle, while acknowledging the technical and conceptual limitations inherent in these emerging technologies.

Data Use and Access Bill (Bill 179, 2024-25)

Some General Observations

[21] The Data Use and Access Bill (Bill 179, 2024-25) introduces provisions in Part 7 that directly affect copyright considerations for AI systems. These provisions target operators of web crawlers and general-purpose AI (GPAI) models. The Bill attempts to distinguish between AI models (the mathematical/statistical algorithms trained to identify patterns, revise or generate new content) and AI systems (the technological infrastructure that integrate models

[8] David Leslie, Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector (Alan Turing Institute, 2019) pp 34-57.

with additional components). The ensuing conceptual ambiguity creates enforcement gaps by placing obligations primarily on 'operators' of models rather than 'deployers' of systems. The AI value chain typically involves multiple parties across development, training, deployment, and operation stages, making responsibility assignment unclear. Additionally, the exclusion of parties solely involved in pre-training activities (such as dataset creators) potentially exempts key actors whose work directly impacts copyright considerations. Crucially, the Bill's distinction between AI models and systems fails to reflect technological reality, where development processes increasingly blur these boundaries. Modern generative AI represents integrated solutions where models cannot be meaningfully separated from deployment infrastructure. Large language models involve continuous training, fine-tuning, and system-level optimisations that make traditional distinctions obsolete.

Summary and Recommendation

[22] To address these limitations, the Bill should adopt an end-to-end accountability approach that maintains responsibility across the entire AI lifecycle and not at the 'model' level. This framework would establish continuous lines of accountability from initial data collection through model training to system implementation and outputs. Such an approach would require:

1. Provenance Tracking: Implementing mechanisms to document when and how copyright-protected works influence AI throughout the development pipeline.

2. Shared Responsibility: Distributing accountability across all participants in the AI value chain rather than focusing exclusively on "operators".

3. Comprehensive Documentation: Maintaining records that connect training inputs to system outputs, creating auditable trails for copyright holders.

Interplay with existing copyright framework

[23] The relationship between the Bill and the copyright framework presents several complexities. While the Bill functions as a public law instrument designed through a pro-innovation lens that shape how risks are conceptualised and distributed, copyright law remains a private law framework providing right-holders with legal entitlements (considered a fundamental right under the European Convention on Human Rights, Article 1 First Protocol). These differing legal frameworks lead to distinct enforcement mechanisms and remedies. Although Section 135 of the Bill mandates compliance with the CDPA, the practical interaction between these frameworks remains ambiguous, not least because it is hard to predict what the outcome of the AI & Copyright consultation will bring.

[24] The Bill's current stance recognises that text-and-data mining techniques, which are essential for developing AI models, may require the use of substantial amounts of copyright-protected material. This would necessitate obtaining the authorisation of right-holders unless an exception applies, such as the existing text-and-data mining exception or any future

exception emerging from the government's ongoing consultation on copyright and AI. However, uncertainties remain regarding subsequent actions related to communication to the public or the making available of text-and-data mining results. Under the current text, the text-and-data mining exception only covers acts of reproduction and extraction. A case in point is the public availability of a dataset resulting from text-and-data mining activities. This act only has copyright relevance when the dataset reproduces protected works or infringes upon a restricted act. The *Kneschke v LAION* case, which dealt with a dataset of image-text pairs, is particularly instructive, as it largely consisted of hyperlinks to publicly accessible images.

Practical Implementation Challenges

[25] The Bill's disclosure requirements create several practical challenges for implementation that suggest the need for a balanced approach, and one which leads to a precautionary stance. First, operators must disclose specific information about web crawlers, including their name, legal entity, purpose, and recipients of scraped data. They must also provide a central contact point for copyright concerns and use distinct crawlers for different purposes. This information must be accessible and regularly updated. Second, these requirements will necessitate advanced auditing mechanisms that may prove resource-intensive, particularly for the UK's AI sector which consists predominantly of startups and SMEs. The technical oversight needed to enforce separate crawlers for different purposes remains underdeveloped at scale. Third, the Bill is unclear as to the standards for responsibility and how these manifest across the AI value chain. It remains unclear how obligations apply to 'downstream providers' who may lack direct access to or control over the AI models they utilise. This creates potential enforcement gaps where responsibility effectively remains with the original model operator.

[26] This creates several pressure points. First, it is unclear whether the obligation to respect copyright also includes moderating AI-generated outputs. Second, non-compliance with the Bill's copyright obligations could result in administrative fines rather than direct copyright infringement claims, although there may be regulatory overlap with the CDPA. Third, these obligations apply broadly, including to open-source models and providers of all sizes, which raises concerns about proportional compliance, especially for SMEs. Finally, enforcement will fall under the Information Commissioner's Office (ICO), which focuses on public regulatory oversight rather than private enforcement by copyright holders.

Rights reservation and territorial considerations

[27] A key focus of the Bill's copyright compliance obligations suggests the identification and respect for a potential future rights reservation mechanism, facilitated through state-of-the-art technologies. This raises important questions about implementation: Does the opt-out apply to original works or digital copies? When and how can it be exercised? Can it be implemented at the source page or training data level? Additionally, despite including extraterritorial provisions to level the playing field among operators, copyright protection

remains inherently territorial. The Bill takes a broader approach than the EU AI Act by extending obligations beyond pre-training and training to include development and operation of GPAI models. However, this creates potential conflicts between the legal regime for text-and-data mining (which applies to reproduction rights) and the act of making available a trained AI model.

Practical transparency approaches

[28] The transparency obligations in Clause 137 extend beyond copyright concerns to cover broader data use considerations. Given practical constraints, disclosure of copyright-protected material in training datasets cannot reasonably require fully itemised lists with clear ownership details. This would be impractical due to the low originality threshold for copyright protection, territorial fragmentation of rights, lack of mandatory registration, and insufficient ownership metadata. A more feasible approach would involve disclosing key categories of information: the overall size of training data, detailed information on datasets and sources (including origin breakdown), data diversity, and processing methods. This approach acknowledges the technical and practical limitations while still providing meaningful transparency to right-holders.

Recommendations for effective transparency obligations

[29] Several recommendations emerge for developing effective transparency obligations:

[30] First, transparency requirements should acknowledge the integrated nature of AI development rather than attempting to draw artificial distinctions between models and systems. A more effective approach would focus on the functions and capabilities of AI technologies, regardless of how they are labelled or structured. This would prevent entities from evading responsibility through technical or organisational arrangements that exploit definitional gaps and legal uncertainty. By focusing on *what the technology does* rather than *how it is categorised*, transparency obligations can more effectively cover the full spectrum of copyright implications.

[31] Second, transparency requirements should be calibrated to each phase of the AI lifecycle, recognising the differing technical realities and legal considerations at each stage. For the input phase, disclosure of training dataset categories and sources is both feasible and valuable. For the processing phase, transparency should focus on general model architectures and processing methodologies rather than attempting to trace specific works through inscrutable mathematical representations. For the output phase, developers should provide clear information about the potential relationship between outputs and training data.

[32] Third, policymakers should avoid a 'legislate and observe the results' approach that risks irreversible consequences. Instead, an iterative regulatory framework that evolves alongside technological developments would better serve both innovation and rights protection. This could include regulatory sandboxes where new transparency approaches can be tested before

widespread implementation. Such mechanisms reflect Julia Black's concept of responsive regulation, allowing for adaptation as technology and social expectations evolve.[9] They also recognise Rotolo's insight that emerging technologies require regulatory frameworks that can develop in parallel with the technology itself.[10]

[33] Fourth, any opt-out mechanism for copyright holders should be designed with technical feasibility in mind. The system should provide clear guidance on when and how rights holders can exercise these options, with standardised technical protocols that both content creators and AI developers can implement without disproportionate burdens. This approach acknowledges the technology-society-law interplay by considering how public trust in opt-out mechanisms can drive innovation rather than hinder it, as it creates clearer parameters for responsible development.

[34] Finally, enforcement mechanisms should acknowledge the distinction between public regulatory oversight and private copyright enforcement. A hybrid approach that provides both administrative remedies through bodies like the ICO and private actions for right-holders could create more comprehensive protection. This reflects Black's emphasis on understanding both formal and informal regulatory influences, creating a more holistic framework that leverages multiple governance mechanisms rather than relying on a single enforcement approach.[11] A key concern is the applicable sanctions or penalties for non-compliance. Although the Bill stipulates that failure to comply with these duties by relevant operators can be directly actionable by copyright owners who are adversely affected, entitling them to damages and injunctive relief, it is important to note that the penalty notice fine—which can reach up to £17.5 million or 4% of the undertaking's total global turnover in the preceding financial year (under Section 157 of the Data Protection Act 2018)—does not benefit copyright owners directly.

[35] Nonetheless, these provisions could indirectly benefit rights holders in two main ways. First, the threat of enforcement could pressure web crawler operators and GPAI model providers to adhere to UK copyright law, which could likely encourage licensing agreements and improve model design. This is already visible in the increasing number of deals between providers, rights holders, and aggregators in the EU. Second, regulatory spillover from the Data (Use and Access) Bill could potentially strengthen direct copyright infringement claims against GPAI providers, particularly for failing to meet the text-and-data mining exception requirements.

---

[9] Julia Black and Robert Baldwin, 'Really Responsive Risk-Based Regulation' (2010) 32 *Law & Policy* 181.
[10] Daniele Rotolo, Diana Hicks and Ben R Martin, 'What Is an Emerging Technology?' (2015) 44 *Research Policy* 1827 <https://www.sciencedirect.com/science/article/pii/S0048733315001031> accessed 4 March 2025.
[11] Julia Black, 'Regulatory Conversations' (2002) 29 *Journal of Law and Society* 163 <https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-6478.00215> accessed 4 March 2025.

Creators' remuneration and licensing models

[36] Regarding creators' remuneration, the Bill encourages GPAI providers to license content from large rights aggregators, ensuring access to high-quality datasets while mitigating legal risks. However, these agreements have yet to result in improved compensation for individual creators, as licensing rights are typically held by larger entities due to previous exploitation contracts. While collective licensing and bargaining could provide a solution, these mechanisms face both substantive and practical challenges.

Proposals for balancing copyright and AI innovation

[37] There is a pressing need for legal reform to better balance copyright protection, artistic freedom, and AI innovation. Several scholars and experts have proposed different approaches to address this issue. Geiger and Iaia suggest the introduction of a statutory license for AI training,[12] while Senftleben advocates for an AI output levy to ensure more equitable compensation.[13] Jacques and Flynn propose a dual licensing model along with an AI-royalty fund.[14] The proposed dual-licensing model detailed in this briefing includes two key elements: (1) licensing requirements for commercially exploited AI-generated content, akin to licensing human-produced songs, and (2) licensing AI services that train on copyright-protected material. An AI royalty fund, managed by a dedicated trust, should be established to address specific sector needs. This fund would prioritise support for creators from heavily impacted genres and those producing high-quality work, especially those disadvantaged in the algorithm-centric market. While these proposals are not without obstacles, elements of these ideas could inform future policy discussions and shape the trajectory of copyright policies related to generative AI.

March 2025

---

[12] Geiger & Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (October 6, 2023). *Computer Law & Security Review*, vol 52, 2024, 1-9., Available at SSRN: https://ssrn.com/abstract=4594873 accessed 4 March 2025.

[13] Senftleben, 'AI Act and Author Remuneration - A Model for Other Regions?' (February 24, 2024). Available at SSRN: https://ssrn.com/abstract=4740268 or http://dx.doi.org/10.2139/ssrn.4740268 accessed 4 March 2025.

[14] Jacques & Flynn, 'Protecting Human Creativity in AI-Generated Music with the Introduction of an AI-Royalty Fund' (2024) 73(12) *GRUR International*, 1137-1149; see Policy Brief: https://www.liverpool.ac.uk/media/livacuk/publicpolicyamppractice/pbseries3/PB3,18,FINAL.pdf accessed 4 March 2025.