Synthetic Media Research Network

Response to the Intellectual Property Office consultation on Copyright and Artificial Intelligence.

Synthetic Media Research Network, 25 February 2025

Authors: Dr Dominic Lees, Dr Mathilde Pavis

Question 4: ***Do you agree that option 3 - a data mining exception which allows right holders to reserve their rights, supported by transparency measures - is most likely to meet the objectives set out above?***

Option 3 is unlikely to meet the objectives in its current form. Our position is based on internal consultations within the Synthetic Media Research Network, a community of academic researchers, AI developers and creative industry stakeholders committed to tackling the social, legal and political challenges of Generative AI. We set out our reasoning to support our answer to this question under the three stated objectives of government policy.

1. Policy Objective: 'Supporting right holders' control of their content and ability to be remunerated for its use.'

1.1 A rights reservation regime designed to enable control by right holders of how their content is used must include a highly accessible, user-friendly and centralised mechanism for right holders to access. This is noted in the Ministerial Forward to the consultation ('simple technical means for creators to exercise their rights'), however we have not found a proven technical solution for such rights reservation anywhere in the world. Important to note is that the EU has begun to confront this problem after the passing of its AI Act, undertaking a feasibility study for 'a central registry of Text and Dat Mining opt-out as expressed by rightsholders' on 23rd January 2025. In the Consultation's section A1 paragraph 12, the technical challenge is acknowledged, and we hope that the UK may avoid the EU's error of mis-sequencing the necessary preparation for an Opt Out system.

We note that the category of rightsholder is very broad, from major libraries and text collections, to individual creative practitioners. Any mechanism for rights reservation must function equally for large rights holding businesses or collecting societies, and for sole traders who lack the knowledge, skills and business bandwidth to engage in opt out procedures.

However, this issue is of critical concern to government policy success: without a technical solution for centralised rights reservation that has been beta tested by

rightsholder organisations and individuals, the government's preferred option will be incapable of winning the confidence of creative producers and rightsholders.

1.2 Remuneration. Paragraph 71 must be developed to give a much fuller understanding of how remuneration will be managed between AI developers and rightsholders. How will negotiation be managed? Will minimum rates be applied? We note that market rates for the licensing of creative datasets are very varied: currently, the value of one minute of video is between $1 and $4. How will the government support individual rightsholders who do not have market knowledge, or skills in negotiating licenses of their work?

2. Policy Objective: Supporting the development of world-leading AI models in the UK by ensuring wide and lawful access to high-quality data.
We cannot conclude that Option 3 will have a major impact on the development of world-leading AI models in the UK. AI developers in the Synthetic Media Research Network include businesses based in Europe, the US and the Middle East. These colleagues do not see that a liberalised copyright regime in the UK will stimulate business development in this country; it will not encourage them to invest in the UK; some comment that it might lead them to acquire datasets in the UK, but only to support their businesses located elsewhere.

The proposal for rights reservation as currently expressed is too simplistic. How will an AI Developer distinguish between a dataset that has been opted out with the intention of never being used for training, and a dataset that has been opted out with a view to negotiating a licensing agreement? It will be necessary to develop a clear system of 'licensing signalling', a mechanism for rightsholders to demonstrate their willingness to license that is easy for AI developers to read, leading to a quick and accessible pathway to securing the dataset.

As we discuss below, the path towards 'the development of world-leading AI models in the UK' is not through a change to the UK's copyright laws. Instead, it is through the careful construction of a centralised resource of data available for license under pre-agreed terms, which we outline in detail under Q4. This proposal provides reassurance to AI developers that they will be able to access licensable UK datasets at scale from a single resource point that is easily accessible. The experience of AI developers across the world is of multiple data aggregators offering datasets for licence, confusing the process of acquisition by developers. The UK has the opportunity to offer a transformative alternative, as we describe below.

3. Policy Objective: Promoting greater trust and transparency between the sectors.
The breakdown of trust between the AI developer sector and the rightsholder community is based on the ongoing breach of copyright law by those businesses involved in unlicensed Text and Data Mining. The reaction of individual and collective rightsholders to the Consultation has intensified this distrust, as will be evidenced in

stakeholder responses to these questions. Government has an important role in rebuilding trust and we recommend a strategy developed by the Synthetic Media Research Network, as cited in POST's horizon scanning report on 'AI and new technology in creative industries' (2024). This involves independently chaired roundtables between rightsholders and selected representatives of the AI industry in a non-confrontational forum.

Question 5. ***Which option do you prefer and why?***

As an independent research organisation, we find that the policy proposals are not addressing the future landscape of the Generative AI industry. None of the four policy options will sufficiently address the policy objectives set out by the government. We think that current copyright law – Option 0 – should be the basis of future policy and that government should create new structures to enable the rapid and consensual licensing of datasets, instead of hoping that a change to the UK's copyright regime will lead to intended policy outcomes.

Our discussions with stakeholders and AI developers participating in the Synthetic Media Research Network reveal that UK government thinking is not keeping up to date with the needs of innovative AI businesses. The important observation is that AI developers need large, reliable and clean datasets – scraping the internet for this resource is only one method, not the only means available. The preferred policy option is designed to facilitate a simpler framework for Text and Data Mining, however many advanced AI developers are moving away from this mode of dataset sourcing. Internet 'scraping' is frequently seen as a clumsy and quite costly means of acquiring data: those with experience of this process describe four business motivations for seeking alternative methods of dataset acquisition:

1. Up to 50% of a developer's budget for dataset acquisition through scraping is spent on 'cleaning' the mined material.
2. Competitors may be mining the same material: using TDM for training a model makes it hard to offer a unique GenAI product to the market that will output responses superior to other providers.
3. Scraped datasets may include material that is AI-generated, undermining reliability and the scope for truly original outputs from the resulting model.
4. Issues of IP and provenance create potential legal risk.

The consultation seeks to solve the fourth of the points, but the proposal for a major change to UK copyright law would only facilitate the current business model of large AI developers, which we think will be redundant in just a few years' time. We predict a future in which TDM will be the least-favoured option for acquiring datasets and recommend that government policy be trained solely on creating a world-leading facility for the mass licensing of data.

Another key development in Generative AI is Small Language Models. AI developers in the Synthetic Media Research Network have demonstrated

remarkable achievements with SLMs, targeted at specific areas of the market and trained exclusively on data that has been licensed. These SLMs outperform Large Language Models, indicating that a major portion of the AI market will, in the future, require limited datasets of quality material, confounding the current assumption that the output performance of models is based on the quantity of data used in training.

Question 6: *Do you support the introduction of an exception along the lines outlined in section C of the consultation?*

We do not support the introduction of Option 3 without the UK Government first undertaking and publishing a detailed and comprehensive impact assessment. The existing impact assessment accompanying the consultation does not sufficiently address several critical areas of concern. Below, we outline key issues the government's impact assessment must cover before progressing any further with Option 3.

1. Complexity of the UK copyright framework

Option 3 risks complicating the UK copyright framework. Option 3 would introduce a third mode of content access, beyond current options of either obtaining explicit upfront permission from rightsholders or benefiting from clearly defined statutory exceptions (when no upfront permission is required). By creating an additional category where content is free to use for data mining or AI training unless the rightsholders explicitly opt-out, the proposed reform may significantly complicate UK copyright law. The government must assess whether increased legal complexity will lead to widespread non-compliance due to confusion, lack of knowledge or resources to manage compliance.

2. Practical implementation and rights reservation protocols

The effectiveness of Option 3 hinges entirely on rights reservation mechanisms that are accurate, reliable, and easily implementable. The impact assessment must address:

- The availability of standardized, easy-to-use technological tools for rightsholders. There are no effective tools at present.
- The resources and knowledge required by rightsholders and AI developers to identify, apply, and respect these restrictions.
- Comparative analysis of similar systems, such as Creative Commons licenses, highlighting issues of misunderstanding and misapplication, even among educated user groups (e.g., heritage institutions).
- The necessity and feasibility of large-scale awareness and education campaigns for global stakeholders involved in data mining.

3. Market impact and economic consequences

The government must evaluate the likelihood and consequences of a "chaotic rush" by rightsholders to opt-out, as observed within the EU, where entities like SACEM and GEMA have opted-out their members en masse. Such widespread opting-out may significantly impact AI training dataset availability and market dynamics, potentially discouraging innovation rather than fostering it.

4. Increased legal uncertainty on rights clearance and risk of misinterpretation

Option 3 risks creating legal uncertainty regarding rights clearance for AI developers. The impact assessment must clarify:

- The extent to which Option 3 clearly communicates to AI developers that other restrictions (e.g., data protection and privacy rights) continue to apply despite the proposed copyright exception.
- How to prevent inadvertent breaches of these other obligations by stakeholders misinterpreting the scope of Option 3.

5. Rationale for Option 3 and sector-specific implications

Unlike the EU's targeted rationale for introducing expanded text and data mining (TDM) exceptions to support specific sectors such as medical and health research (Recital 11, CDSM; Guadamuz, Scanner Darkly 2024, p. 121), the UK’s proposed rationale broadly refers to generic AI innovation. The assessment must:

- Clearly articulate sector-specific impacts, especially distinguishing between the creative and cultural sectors versus scientific and medical research.
- Evaluate whether creative and cultural datasets substantially contribute to sectors beyond entertainment and culture (e.g., healthcare) and whether this justifies lowering copyright protection for creators.
- Examine the risk of undermining UK rightsholders’ ability to leverage their IP assets for participation in the AI innovation market.

6. International Law Compliance

Finally, the detailed analysis of Option 3's compatibility with international intellectual property law is crucial. This must be thorough and address two key international standards.

6.1. Rights Reservation as Rights Assertion

International treaties binding on the UK explicitly prohibit imposing formalities for securing copyright and performers’ rights protection (Berne Convention, Article 5(2)); Performances and Phonograms Treaty, Article 20; Beijing Treaty, Article 17).

An argument can be made that the broad scope of the exception proposed, in relation to economic activities (data mining and AI training) is so critical in the age of Generative AI, combined with strict requirements attached to the form of the opt-out – which may be necessary for the scheme to be workable in the first place– converts a formality for ‘right reservation’ into a formality ‘right assertion’ or ‘right subsistence’, prohibited under international law.

The impact assessment must:

- Evaluate whether requiring rights reservation via technical formats (metadata encoding) may constitute ‘formality’ prohibited under international intellectual property law. The broad scope of the proposed exception—covering economically crucial activities such as data mining and AI training in the era of Generative AI—combined with stringent requirements for opting out (which may be essential for Option 3 to function well), risks transforming what should be merely a formality for 'rights reservation' into a prohibited formality for 'rights assertion' or right subsistence.
- Provide comparative international examples exploring the nuanced spectrum of formalities and rights assertion mechanisms that may be permissible. For example, under US copyright law, ‘works’ may be protected without registration but registration with the copyright office is required to introduce legal proceedings.

6.2. Three-Step Test Compliance (Berne Convention)

The government must rigorously test Option 3 against the three-step test (Berne Convention Article 9(2); TRIPS Agreement, Article 13; WPPT Article 9(2); Beijing Treaty, Article 13), examining each step comprehensively:

- **Step 1 (Special Case):** Confirm if Option 3 qualifies as a "special case." Currently, its broad allowance for data mining and AI training for any purpose likely fails this step. By contrast, the existing copyright exception for non-commercial data mining classes as a special case.
- **Step 2 (Normal Exploitation):** Evaluate how Option 3 conflicts with the normal economic exploitation of rights-protected content, especially within sectors heavily reliant on usage-based licensing. A wide range of UK stakeholders have expressed an interest, and began industrial negotiations, to establish licensing agreement on commercial data mining, AI training, prompting and generating. This would suggest the exception would interfere with the normal exploitation of rights for those stakeholders.
- **Step 3 (Legitimate Interests):** Analyse whether the option to opt-out sufficiently protects creators' legitimate interests, considering the alleged (GEMA filing for copyright infringement against Suno) documented failures in similar rights reservation protocols (e.g., Longpre et al., 2024).

The assessment should acknowledge potential international treaty violations, offering a clear justification if the government anticipates economic or other benefits that may outweigh compliance concerns.

How did the EU's opt-out regime pass the three-step test?

The 'opt-out' regime in the 2019 CDSM Directive specifically addresses data mining, not AI training, suitably limiting its scope. Although it covers both commercial and non-commercial data mining, the EU legislator intended it primarily to support scientific discovery and health research, particularly through public-private partnerships (Recital 11, CDSM; Guadamuz, Scanner Darkly 2024, page 121)—not general commercial research as interpreted by the UK government. As originally designed, this narrower scope aligns with the first step of the three-step test under international law.

However, the subsequent EU AI Act expanded these provisions to include AI training without restricting the purpose or context to scientific or health research. Combined, these two regulations now likely fail the first step of the three-step test.

Unless it can be demonstrated that rightsholder opt-outs are practically effective and enforceable, these EU provisions may face legal challenges for breaching international treaties. Such challenges are particularly likely in EU Member States like France, where domestic courts can directly enforce ratified treaties, including the Berne Convention.

7. Risks of rights restrictions replicating Digital Rights Management (DRM) failures

Relying on rights restrictions for data mining and AI training could unintentionally restrict lawful uses, such as non-commercial research, criticism, parody, or education. Historical evidence from DRM systems highlights the risk of similar negative impacts on activities permitted under copyright exceptions or limitations. This point is further expanded in our answer to Q8 of the consultation.

In conclusion, the UK Government must undertake a thorough and nuanced impact assessment addressing the above considerations. Only after fully evaluating these factors should any further steps towards adopting Option 3 be considered.

Question 7. ***If so, what aspects do you consider to be the most important? If not, what other approach do you propose and how would that achieve the intended balance of objectives?***
As above.

Question 8. ***If not, what other approach do you propose and how would that achieve the intended balance of objectives?***

We do not think that the copyright exception will achieve the policy objectives, which we think can be achieved with alternative policies. The central business requirement of AI developers is a means of acquiring quality datasets at scale, to which Text and Data Mining is a clumsy solution. Developers have turned to this method only because there is no simple and accessible alternative, not because it delivers the best results for resourcing their model training requirements. Data aggregation businesses now form a rudimentary market for the acquisition of datasets, however AI developers are finding that the material offered often lacks the qualities that they require.
We recommend a UK solution to this problem that we think will win wholehearted support from rightsholder community, while smoothing the process of dataset acquisition for AI developers. This would be achieved through a government-regulated centralised rights/data storage facility, the 'UK Licensed Dataset Bureau'. Allied to government's plan for a National Data Library (NDL) of government and open access datasets, this body would handle datasets with complexities of rights: copyrighted creative material in the form of text, voice, image, sound, and video. A centralised system, highly visible to international AI developers, with guaranteed ease of access, will secure a competitive advantage for the UK.

This proposed arms-length body, the 'UK Licensed Dataset Bureau', would be responsible for the management of developers' access to licensable creative and media copyrighted datasets in the UK. This body's need to handle complex rights issues will require a different management skillset to the administration of the NDL.

This proposal would also secure success in policy objective 1: 'Supporting right holders' control of their content and ability to be remunerated for its use.' The body would explicitly occupy a neutral role, guaranteeing to rightsholders that their rights will be managed securely.

The arms length body would have responsibility for:

1. Compiling a list of datasets whose owners have reserved their rights but are willing to licence on agreed terms collective licencing.
2. Liaising with collecting societies' and large holders of copyrighted material to establish lists of confirmed datasets available for licencing.
3. A dataset aggregation service for individual rightsholders and SMEs.
4. Ensuring qualities of data 'cleanliness' and metadata standards, complying with the needs of AI developers.
5. Ensuring compliance with the UK's treaty obligations and GDPR.
6. Provision of legal advice to individual or SME creative rightsholders, in order to secure accessibility to the licencing scheme.

The arms-length body must be resourced so that it can provide a tailored response to AI developers. It requires the legal and negotiating expertise to nuance the varied needs of AI developers and rightsholders.

Why set up this arms-length body?

AI developers will be attracted to the UK by a smooth process of acquiring the specific datasets needed by their operation. AI developers have expressed to us their need for guarantees of the 'cleanliness' of licensed datasets, metadata standards, rights clearance, and forward usage agreements. These are requirements that developers often find lacking in commercial data aggregation services. By providing government-assured safeguards through this national arms-length body, the UK will establish itself as a premier location for the AI industry.

The proposal envisages a mature future for the AI industry, in which uncontrolled scraping is replaced by a structure mutually beneficial to both rightsholders and AI developers. AI developers will come to the UK to acquire datasets from both open source NDL and the UK Licensed Dataset Bureau. We are certain that most AI developers would favour the ease of accessing licensable datasets at scale, replacing the need to run bots to scrape for these resources.

Question 9. ***What influence, positive or negative, would the introduction of an exception along these lines have on you or your organisation? Please provide quantitative information where possible.***

Not applicable.

Question 10. ***What action should a developer take when a reservation has been applied to a copy of a work?***

A reservation of rights by a copyright holder would be a declaration of one of two positions:

1. This work may never be used as AI training data;
2. This work may be licensed with mutual agreement.

The action that a developer takes would be different in these two cases:

1. Respect the copyright and not use the work in AI model development;
2. Seek a license.

The preferred policy, Option 3, thus creates complications for AI developers, not clarity. A much easier alternative for AI developers is to understand that all works are protected by UK copyright law and they should seek licensing of datasets from a 'one stop shop', a centralised rights and data holding body that has been mandated by government. Our proposal for an arms-length body to facilitate dataset licensing would meet policy objectives and industry needs.

Question 11. ***What should be the legal consequences if a reservation is ignored?***

Circumventing or disregarding rights reservations relating to data mining and AI training should logically constitute copyright infringement. However, the effectiveness of enforcement by rightsholders is limited due to several practical challenges: parties may not have enough knowledge of the law to understand their rights have been infringed; parties may not be aware of the infringing activities; the costs of litigation are too high to seek redress, or the damages a court might award may be too modest to be worth pursuing.

Under current UK law, damages for copyright infringement are usually calculated based on the licensing fees that would have been agreed upon by the parties if a proper licensing arrangement had been in place at the time of infringement. As a new licensing market, it will likely take years for rates applicable to data mining or AI training to become a reliable point of reference for accurate compensation. Only in exceptional cases are damages calculated according to the actual benefits or enrichment gained by the infringer, and this measure is applied only when no other reasonable method of licensing rate calculation is available.

In the specific context of data mining and AI training, it is uncertain whether such damages would be substantial enough to serve as an effective deterrent. In other jurisdictions, effective deterrence typically includes the possibility of awarding punitive damages, which are designed to penalize and discourage deliberate or reckless infringement. Without a similar mechanism, or alternative means to disincentivise infringement, the UK copyright framework may not have what it takes to encourage compliance on scale. On this, we circle back to the needs of providing a convenient means of compliance to AI developers, such as a centralised point for accessing and licensing datasets.

Question 12. ***Do you agree that rights should be reserved in machine readable formats? Where possible, please indicate what you anticipate the cost of introducing and/or complying with a rights reservation in machine-readable format would be.***

1. Requirement for rights restrictions to be ‘machine-readable’

1.1. Limitations of 'Machine-Readable' as a Concept

The term ‘machine-readable’ is overly generic and does not offer sufficient clarity for rightsholders or AI developers. Traditionally, ‘machine-readable’ meant information encoded specifically for software retrieval and execution. However, recent technological advances, particularly in natural language processing, have made such encoding unnecessary, as machines can now understand instructions in plain text. Consequently, the distinction between encoded formats and plain text has become unclear or irrelevant. This ambiguity extends to the EU Directive, which vaguely references rights restrictions in metadata or website terms as 'machine-readable,' thus creating confusion rather than clarity for stakeholders.

1.1. Alternative Solutions for Effective Rights Disclosure

Rather than relying on a broad and unclear concept like ‘machine-readable,’ it may be more effective to focus on where and how rights reservations are disclosed. The UK IPO could look to existing frameworks for moral rights assertions as potential models. However, caution is needed, as international treaties prohibit conditioning the enforceability of rights upon formal declarations or assertions (e.g., Berne Convention Articles 5(2), 5(3)). Any proposed approach must ensure compliance with these international obligations while providing clear guidance on how rightsholders can effectively communicate their restrictions.

2. Unintended costs of machine-readable rights restrictions

We are concerned that a system relying on rights restrictions for data mining and AI training may have the unintended cost of reducing access to content for non-commercial research or other purposes allowed by copyright law such as review, criticism or parody.

The UK Intellectual Property Office will be aware of the unintended negative consequences associated with Digital Rights Management (DRM) systems and tools, which rightsholders deployed widely during the initial rise of internet and digital technologies. Scholars worldwide, including in the UK, have extensively documented how DRM adversely impacted the lawful access to, and reuse of, copyright-protected works. Specifically, DRM tools diminished users' rights to engage with copyrighted content for purposes permitted under copyright exceptions, supporting education and free speech.

To briefly summarize key points highlighted by this scholarship, the harmful effects of DRMs on the public’s lawful access to content primarily arose from:
(1) The inability of DRM tools to accurately distinguish between lawful and unlawful uses of copyright-protected works.
(2) The legal prohibition against circumventing DRM, regardless of whether a copyright exception applied.
(3) Users' insufficient understanding of copyright law, which limited their ability to recognise and exercise their legal rights.

As the UK Intellectual Property Office will be aware, copyright exceptions and permitted acts under UK copyright law do not make for a long list and have been hard earned by stakeholders defending them. We should be mindful to proactively protect and preserve those copyright-free zones. We strongly advise that any proposed rights-restriction regime for data mining or AI training carefully avoid replicating the historical issues around DRM described above. Researchers engaging in non-commercial research should be allowed to mine rights-protected content under the conditions set by the law. Any new provision around machine readable rights reservation should not inadvertently prevent a writer, an artist or a student from using rights-protected content to train an AI model for the purpose of engaging in criticism, parody or pastiche.

TECHNICAL STANDARDS

Question 13. ***Is there a need for greater standardisation of rights reservation protocols?***

Yes, this would be an important role of government in the case of implementing a rights reservation policy. Evidence in the EU on the implementation of the Directive and EU AI Act confirm the need for more standardisations, and this being a point of failure. Initiatives such as Creative Commons have tried to provide similar solutions over several years, on more narrower points of law, but this has still proved challenging. The experience of similar tools of rights signalling is relevant to consider here. CC labels and tools are routinely mislabelled leading to licences being granted without the appropriate permissions, or restrictions applied on public domain content.

The preferred option only speaks to rights restriction or permission on copyright, whereas most content (image, audio, video) will also carry other rights like GDPR. It is important to note that rights restrictions applied by rightsholders often lead to more misleading/misrepresentation of the law and usage permissions.

Question 14. ***How can compliance with standards be encouraged?***

Compliance with standards could be encouraged by: (1) the copyright exception being conditional on the mining agent being registered and instructed to respect rights restrictions communicated in standardised formats; (2) prohibiting the use of unregistered mining agents.

Question 16. ***Does current practice relating to the licensing of copyright works for AI training meet the needs of creators and performers?***

AI developers in the UK lag behind their international peers in embedding systems of licensing and remuneration for copyright works. On July 9th 2024, the Synthetic Media Research Network convened a Roundtable that brought together UK rightsholders and three international AI developers from Israel, Ukraine and the US – companies that train their models exclusively on licensed datasets. The dialogue that we hosted revealed to UK rightsholders that there is a section of the AI industry committed to an ethical business model, a strategy that is both supportive in principle of copyright and sets these companies apart from their rivals that are committed to TDM without licensing. We see a role for government in promoting the best practices that have been developed by ethical AI developers.

Question 17. ***Where possible, please indicate the revenue/cost that you or your organisation receives/pays per year for this licensing under current practice.***

Not applicable.

Question 18. *Should measures be introduced to support good licensing Practice?*

Yes. Our policy proposal (cf Q4) would create a centralised bureau responsible for implementing standards of licensing.

Question 19. *Should the government have a role in encouraging collective licensing and/or data aggregation services?*

Yes.

Question 20. *If so, what role should it play?*

Our proposal for an arms-length body, which we have given the working title 'UK Licensed Dataset Bureau', would fulfil government's role in centralising the licensing of aggregated data. There is an essential role for government to ensure that those rightsholders with less resources, such as certain museums, individual artists and microbusinesses, are included in the opportunities for creative dataset licensing. Currently, the opportunities are being enjoyed by large news organisations but not small scale rightsholders.

Government should fund the setting up of the proposed Bureau, which cannot be afforded by small rightsholders, including support for the preparation of approved legal standards for licenses.

Question 21. *Are you aware of any individuals or bodies with specific licensing needs that should be taken into account?*

Yes: Micro-entities, self-employed artists and performers, start-ups and SMEs.

TRANSPARENCY

Question 22. *Do you agree that AI developers should disclose the sources of their training material?*

Yes, disclosure of the sources of training material is part of an ethical business strategy that has been proved to be effective for AI developers as well as creating trust between the AI industry, rightsholders and the users of GAI tools.

Government policy looking to the long-term future should consider the growing discussion of 'eXplainable AI' (XAI) (Pavlidis, 2024), as well as the EU AI Act's restrictions on the AI 'Black Box'.

Promoting best practice: a case study

We propose a conceptual alternative to the 'Black Box': the 'Glass Box', which is a model of AI development seen in the Israeli company, BRIA. We believe that developing policy around this business model will enable government to fulfil its three policy objectives.

BRIA is an AI developer that has createde a prompt-based image generator similar to that of Midjourney. The business is now moving into video and sound generation. From its launch, BRIA has only used licensed material as its training data and has a policy of complete transparency, as well as an advanced system for the remuneration of copyright holders whose work is used to generate outputs. When the Synthetic Media Research Network invited BRIA to its Roundtable with copyright holders in July 2024, the company demonstrated its Attribution Technology. When a new image is created from a prompt, a single click reveals the original content within the AI model that most impacted its generation. In the example we saw, five photographs that had been used as training data were identified; the copyright holders of those photographs will be remunerated by BRIA for the use of their work.

This case study demonstrates that transparency of training material is possible, makes business sense for the AI developer, and an equitable system of remuneration for copyright holders can be a simple procedure. Owing to its approach of respect for copyright and remuneration, BRIA now finds that rightsholders come with offers of datasets to be used as training data. The company says that the quantity of such training data is ample for its business purposes, confounding the voices of some AI developers who claim a dearth of data necessitates TDM.

The transparency offered by such a 'Glass Box' system is of vital importance to end users of generated material: there is no legal risk involved with using such outputs within, for instance, a film that must secure complete copyright clearance of all its embedded IP.

Question 23. ***If so, what level of granularity is sufficient and necessary for AI firms when providing transparency over the inputs to generative Models?***

The required level of granularity depends on the intended objectives of transparency. Transparency serves three main purposes:

1. Purpose 1: Enabling Rights Enforcement: Transparency intended for rights enforcement requires highly detailed information. This includes identifying individual input files by title, source organisation, or specific URL if accessed online. Upon request from a rightsholder, this information should be provided in a timely manner, ideally within one month, and in a user-friendly format, such as a spreadsheet, suitable for individual freelancers, employees, or organisations specialised in rights enforcement.

The burden of managing detailed content records should be reasonable for AI developers, as handling large datasets aligns closely with their core expertise. Such record-keeping practices are already common among technology companies for

compliance purposes, such as GDPR requirements, or during asset evaluations for investment or acquisition processes. Therefore, documenting data provenance should not represent a significant new challenge for developers. Concerns over commercial sensitivity in disclosing this information to competitors can be effectively addressed through confidentiality agreements.

In situations where an AI developer is also a rightsholder with competing interests, conflicts can be managed by referral to the Intellectual Property Office or a suitable independent body tasked with balancing interests. While currently rare, these scenarios may become more common as AI technology matures and rights-holding organisations increasingly engage in AI development.

EU Member States are actively considering this matter within the implementation framework of the EU AI Act. Notably, a task force from the French Ministry of Culture recently published a detailed report and provided a transparency template for AI developers (see page 30, Bensamoun et al., Report of the Task Force on the Implementation of the European Regulation Establishing Harmonized Rules on Artificial Intelligence, Conseil supérieur de la propriété littéraire et artistique, 11 December 2024).

2. Purpose 2: Verification by End-users and Consumers: Transparency allows end-users and consumers to verify AI companies' claims regarding data sources or model performance at the point of purchase. For this purpose, transparency does not require the same level of detail as for rights enforcement. Instead, it should provide a clear overview of the input datasets, including the proportions of crawled versus licensed data, and whether specialised or general datasets were used. Enabling verification for end-users is essential for rights-based sectors like the film, music or publishing industries. Professional end-users who monetise their intellectual property by transferring titles onwards need legal certainty in the intellectual property rights they have accrued and will transfer. For those users, using AI tools with unclear sources of training data in their creative process can generate risks of intellectual property “pollution” (by carrying infringement risks from the input to the output data) or “denuding” of the intellectual property (by ‘thinning’ the copyright they have created to the parts created without the assistance of AI) they created and wish to monetise.

3. Purpose 3: Encouraging Fair Competition: Transparency can serve as a differentiating factor for AI companies, enabling fair competition by accurately representing training inputs. Clear disclosure helps prevent anti-competitive practices such as 'ethical AI washing', where companies might make unsubstantiated ethical claims about their data sourcing and usage.

For purposes (2) and (3), transparency can remain at a higher descriptive level, without close granularity. Nevertheless, end-users and consumers should retain the ability to request additional technical details where necessary, ensuring informed decision-making regarding technology suppliers.

Question 24. ***What transparency should be required in relation to web Crawlers?***

Key features of transparency by Crawlers should be enforced by government:
1. No crawlers without disclosed agents;
2. Crawlers should be registered to access benefits of a copyright carveouts; and they should keep a record of what they've crawled for rightsholders to access.
3. Crawler traffic information should be easily findable; the individual or collective copyright holder has the right to know that they have been crawled.

Question 25. ***What is a proportionate approach to ensuring appropriate Transparency?***

In our response to Question 17, we have demonstrated how transparency can become a business advantage to an AI developer. The approach of government should be to incentivise the Glass Box approach and to disincentivise TDM and Black Box strategies. Where commercially sensitive training datasets are involved, simple systems of anonymisation can be deployed.

Question 26. ***Where possible, please indicate what you anticipate the costs of introducing transparency measures on AI developers would be.***

No response.

Question 27. ***How can compliance with transparency requirements be encouraged, and does this require regulatory underpinning?***

Yes, there must be a regulatory underpinning of transparency requirements: the right to crawl must be dependent on transparency.

Wider clarification of copyright law

Question 29. ***What steps can the government take to encourage AI developers to train their models in the UK and in accordance with UK law to ensure that the rights of right holders are respected?***

Our proposal for a single, centralised source of licensable training data will be world-leading, establishing the UK as the best source for reliable, clean and legally assured datasets. We caution that while establishing such a resource will make the UK very 'AI friendly', it cannot be guaranteed that this will lead AI developers to relocate to this country; as we have described above, our consultations with international AI developers in the Synthetic Media Research Network indicate that neither will the preferred Option 3.

Question 30. ***To what extent does the copyright status of AI models trained outside the UK require clarification to ensure fairness for AI developers and right holders?***
No response.

Question 31. ***Does the temporary copies exception require clarification in relation to AI training?***
Generally, no. In discussions and interviews conducted by the Synthetic Media Research Network, very few rightsholders—and no AI developers—have suggested that the temporary copies exception should apply to AI training activities. Both the literal wording and broader legislative context of this provision clearly indicate that it does not cover acts related to AI training. To interpret it otherwise would distort both the language of the exception and the legislative intent behind it.

We wish to bring attention to one caveat on the position outlined above, in relation to the fine-tuning of an AI tool for the purpose of creating the digital replica of person's voice, face or body. The process of fine-tuning a base model so that it can consistently generate outputs in the voice or likeness of a performer often requires training a specialised AI system on recorded speech or performances. Often these recordings will not only capture a person's speech or performance, but also the underlying in-protection copyright works that may be interpreted (eg, the narration of a literary work, the performance of a musical work or that of a work of dance). In this context, a performer may wish to fine-tune an AI model to generate new content in their likeness or performance style, using as wide a range of previously recorded work in their portfolio (the rights to which they do not own or control). Here, the fine-tuning process may engage in temporary and incidental reproduction of the underlying copyright works as the AI system analyses the likeness or the performance.
There are, prima facia, valid and reasonable arguments to regard this activity as making temporary copies of the underlying copyright works. For example, we may wish to see performers embrace AI technology and remove barriers to creating high-performing digital replicas of their likeness by allowing fine-tuning on their portfolio of previous work. In practice, it is unlikely that performers will have retained or obtained the rights to fine-tuned models for this purpose.

For these reasons, we recommend that the UK Intellectual Property Office investigates and assesses whether the temporary copies exception may apply to the use of sound recordings or film for the purpose of creating the digital replica of a person's voice, face or body.

Question 32. ***If so, how could this be done in a way that does not undermine the intended purpose of this exception?***
No response.

ENCOURAGING RESEARCH AND INNOVATION

Question 33. ***Does the existing data mining exception for non-commercial research remain fit for purpose?***

The original intention of the UK legislator to encourage research across all fields of science by introducing the text and data mining exception for non-commercial research remains valid. Recent AI developments do not invalidate its purpose or relevance.

Question 34. ***Should copyright rules relating to AI consider factors such as the purpose of an AI model, or the size of an AI firm?***

The statutory framework of copyright as expressed in the Copyright, Designs and Patent Act 1988 should apply regardless of the size of an AI firm. The purpose of an AI model may be considered, in the same way that the purpose of certain activities is already considered when determining whether a copyright exception or defence applies (see for example, exceptions based on non-commercial research, making accessible copies, education, archiving, parody, criticism or review). We could envisage, too, that an institution may train an AI model in the context of non-commercial research under a differentiated copyright regime in contrast to organisations performing the same activity for commercial purposes. However, and wherever possible, those differentiations are best handled by the industry though licensing, following a sector-by-sector approach. This would allow tailored, flexible and easily updated terms to be set by the relevant market players. The UK IPO could establish mechanisms to prevent unfairness or paralysis in the collective bargaining process, should this be a concern.

Computer-Generated Works (CGW): protection for the outputs of generative AI.

CGW Policy Option 0: No legal change, maintain the current provisions

Question 35. ***Are you in favour of maintaining current protection for computer-generated works? If yes, please explain whether and how you currently rely on this provision.***

Yes.

Our discussions and interviews with industry rightsholders and AI developers, the regime related to computer-generated works under UK copyright law presents no significant challenge or barriers. In this regard, we have no evidence to suggest or justify a change of the law on this point. Uncertainty in rights subsistence for creative practices relying heavily on automated means is inherent to the originality condition of both ‘traditional’ copyright and the provisions of computer-generated works. In practice, matters of rights subsistence or ownership are managed by contract to

reduce or remove legal uncertainty. Rightsholders have warned that if rights are removed from computer-generated works, this could lead to undesirable effects by making computer-generated works more attractive than human-generated works in certain markets, as the absence of rights may lead to lower transaction costs. This may inadvertently ‘devalue’ human-generated works. This concerns remains theoretical at this point as there is no empirical evidence – that we are aware of – to either validate or disprove it.

AI Output labelling

Question 45. ***Do you agree that generative AI outputs should be labelled as AI generated? If so, what is a proportionate approach, and is regulation required?***

Yes.

Government must protect the consumer from potential harms caused by invisible Generative AI outputs. This is particularly important for online content and media communications. We note that the October 2024 Party Political Broadcast by Reform UK was partly created using AI tools but made no disclosure of this to its viewers. Lack of regulation opens the potential for damage to the UK ‘s democratic culture and social cohesion and a system of labelling is required. For media content, this could be similar to the ‘BAFTA Albert’ sustainability certification on broadcast programmes and films.