Response

Submitted to Consultation on Copyright and Artificial Intelligence
Submitted on 2025-02-25 09:25:20

Basic Information

1 What is your name?

Name:
Paul Keller

2 What is your email address?

Email:
paul@openfuture.eu

3 What is your organisation?

Organisation:
Open Future Foundation

Copyright and Artificial Intelligence

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?

Yes

Please give us your views:

We have long held that the European Unions approach to Text and data mining represents a sensible balance between the interests of rightholders to control certain uses of their works on the one side and the interests of the general public and developers of AI systems on the other hand. We therefore welcome the fact that the UK is now considering to follow this example.

The introduction of a general text & data mining exception to UK copyright law, which would apply unless the use of a work has been expressly reserved by the rights holder in a machine-readable manner, is likely to balance the respective interests of users (including AI developers), rights holders, and the public interest in access to knowledge. However, in order to meet the objectives of control, access, and transparency, the UK should precisely align the legal mechanism used for text & data mining with international precedent.
In this regard, option 3 as outlined in the consultation document includes some ambiguities that should be avoided:
The consultation document considers AI developers as the main beneficiaries of a text & data mining exception. While it is true that this exception is critical for the training of AI models, it should be borne in mind that a copyright exception does not govern “access to works by AI developers”, but rather provides for a justification for specific copyright-relevant acts. AI training is not a copyright-relevant act as such, but it involves copyright-relevant acts of reproduction. AI training is but one of many beneficial use cases of a general text & data mining exception, which should cover, at minimum, acts of reproduction for the purpose of any analytical technique aimed at analysing text & data in order to generate information. This definition aligns with EU law and is sufficiently technology-neutral to encompass potential future use-cases.
AI developers will not have more access to training data purely as a consequence of the introduction of a copyright exception. The exception will merely allow them to use data to which they already have access. To ensure that the exception can meet the UK government’s objectives, legally ambiguous conditions on the enjoyment of the exception, such as the requirement that the user have “lawful access”, should be avoided. Instead, in line with EU copyright law, the text & data mining exception should apply when the works are “lawfully accessible”. Lawfully accessible works are those that a user can access without breaking any laws, whereas a requirement to “have lawful access” could be interpreted as a requirement on the user to investigate the conditions under which works have been made accessible, a requirement that is fundamentally incompatible with the UK government’s stated objective of allowing AI developers “to train on large volumes of web-based material without risk of infringement”.

5 Which option do you prefer and why?

Option 3: A data mining exception which allows right holders to reserve their rights, supported by transparency measures

Please give us your views.:

We have long argued that the European Union's approach to text and data mining strikes a reasonable balance between the interests of rights holders in controlling certain uses of their works, on the one hand, and the interests of the general public and developers of AI systems, on the other. We therefore welcome the fact that the UK is now considering following this example.

In our view, such a solution is preferable to a data mining exception that would unduly privilege AI model developers to the detriment of authors and other rights holders who would be prevented from exercising their copyright rights against AI model developers. On the other hand, requiring licences in all cases of text and data mining would have the effect of making many routine forms of data processing dependent on licences. This is particularly problematic as - due to the excessively long duration of copyright protection and the low threshold of originality - the majority of works that are publicly available online are not actively managed by their rightholders. This would create a massive orphan works problem and severely limit many of the core

functions of modern societies, which are increasingly dependent on all forms of data processing.

A general exception for data mining combined with a rights reservation mechanism strikes the right balance. In particular, it does not put rightholders who wish to exercise their rights in a worse position than if all text and data mining activities required explicit permission.

Our proposed approach: Exception with rights reservation

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

Yes

Please give us further comments.:

We support the introduction of a general text & data mining exception that is closely modeled after Art. 4 EU Copyright in the Digital Single Market Directive. To the extent that the proposal diverges from the EU approach, it should be aligned to promote legal certainty, innovation and cross-border collaboration in research, industry and civic data projects.

7 If so, what aspects do you consider to be the most important?

Please give us your views.:

The text & data mining exception should apply to copyright-relevant acts recognized by current copyright law (acts of reproduction) and not create a new copyright-relevant act (access and use for AI training).

The exception should not be limited to situations in which the user “has lawful access” to the works. If a similar requirement is considered, it should apply when the works in question are “lawfully accessible” (see justification in response to q1).

The exception should apply unless the use of a work has been expressly reserved by the right holder in a machine-readable manner.

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

Please give us further comments.:

n/aAn opt-out from TDM or AI training shall require the AI model developer to remove the opted-out work (including all instances of the opted-out work) from, or not include it in, its training data sets. This means that the AI model developer should be required to stop using the opted-out work to train new AI models. However, it shall not require the AI model that has already been trained to “unlearn” the work. In other words, AI model developers shall not be required to commit to remove opted-out works from already trained models, since that is not technically feasible. Instead, they should be encouraged to record and publicly communicate the training dates, after which new opt-outs will no longer have to be complied with.

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.

Please give us your views.:

n/a

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

Please give us your views.:

An opt-out from TDM or AI training shall require the AI model developer to remove the opted-out work (including all instances of the opted-out work) from, or not include it in, its training data sets. This means that the AI model developer should be required to stop using the opted-out work to train new AI models.

However AI model developers shall not be required to commit to retroactively remove opted-out works from already trained models. Instead, they should be encouraged to record and publicly communicate the training dates, after which new opt-outs will no longer have to be complied with.

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

Please give us your views.:

The user would no longer be able to rely on the proposed text & data mining exception for acts of reproduction of copyright-protected works the use of which has been expressly reserved by the rightholder. Other copyright exceptions would remain unaffected. Rightholders would be able to use remedies against copyright infringement if no exception applies.

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.

Please give us your views.:

Yes. Text & data mining typically involves large amounts of data. Due to the lack of a copyright registry and the need for a case-by-case assessment of originality, it is not possible to identify all copyright-protected works included in a text & data mining operation at scale. This makes proactive rights clearance infeasible. By introducing a text & data mining exception with a mechanism for express rights reservation, the EU legislator tried to account for this difficulty by making text & data mining permissible by default. Where rightholders wish to reserve their rights, they have to do so expressly and, in the context of online resources, in a machine-readable manner in order to allow beneficiaries of the exception to exclude such works from their text & data mining operations in bulk. This is only possible if the rights reservation is machine-readable.

Please refer to our 2024 paper on "Considerations for implementing rightholder opt-outs by Ai model developers" available at https://openfuture.eu/publication/considerations-for-implementing-rightholder-opt-outs-by-ai-model-developers/ for a more detailed discussion of the criteria for machine readability.

Technical Standards

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

Please give us your views:

Yes. While the Robot Exclusion Protocol (REP) is currently the only technical solution used at a large scale, this protocol has a number of conceptual shortcomings that make it unsuitable as an expression of rights reservations, including the following: 1) REP policies can only be set by entities that control websites/online publishing platforms (in many cases, these entities will not be the rightholders themselves); 2) REP has limited usefulness for types of content (such as music or AV content) that are not predominantly distributed via the open internet; 3) REP is unable to deal with embedded media files; 4) most importantly, REP does not allow opting out of TDM or specific applications of TDM (e.g. AI training). Currently, REP does not allow a web publisher to indicate a horizontal opt-out that applies to all crawlers crawling for a specific type of use. Instead, web publishers are required to target each individual crawler individually, often lacking information about the crawlers in use and the purposes of individual crawlers.

Further standardization is therefore recommended to streamline opt-out processes and increase legal certainty. Standardization of rights reservations benefits all affected parties, as it provides clarity over the practical application of the standard of machine-readability, facilitates compliance with the limits of the text & data mining exception and makes it easier for rightholders to expressly reserve their rights, trusting that a rights reservation expressed in accordance with a standard will be respected.

There are many other technical solutions for expressing opt-outs, but none of them can currently be considered a widely used industry standard. As this is a highly dynamic area, it is important to ensure that the reservations of rights / opt-outs expressed by different solutions are interoperable at a conceptual level. This can be achieved by aligning them with a vocabulary that defines the scope of different types of opt-outs. Open Future is working on such a vocabulary with stakeholders from across the spectrum.

14 How can compliance with standards be encouraged?

Please give us your views.:

Compliance with rights reservation standards can be encouraged through legislative intervention. AI model developers should be required to comply with all standardized machine-readable means to express rights reservations that may emerge over time.

15 Should the government have a role in ensuring this and, if so, what should that be?

Please give us your views.:

See our answer to question 13. In addition, the government could be responsible for maintaining an up-to-date list of rights reservation standards (or cooperate with the EU in maintaining such a list). Finally, a public registry for recording rights reservations is also advisable, although such an effort should be undertaken through international cooperation or within the context of a set of federated registries. Rightholders should not be required to register opt-outs on a country-by-country basis.

Transparency

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

Please give us your views.:

Yes. Greater openness and transparency in the development of AI models can serve the public interest and facilitate better sharing by building trust among creators and users. As such, we generally support more transparency around the training data for regulated AI systems, and not only on training data that is protected by copyright.

Simply requiring AI developers to publicly disclose the data sources used for training of the AI model is not enough. AI developers should also be required to release a summary of the internal compliance policies followed during the scraping and training stages, including a list of the rights reservation protocols complied with during the data gathering process.

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

Please provide further comments:

Please refer to our 2024 policy paper containing a proposal for implementing the AI Act's
training data transparency requirement for GPAI available at: https://openfuture.eu/publication/towards-robust-training-data-transparency/ and the blueprint for the training data template available here:
https://openfuture.eu/wp-content/uploads/2024/09/240919AIAtransparency_template_requirements-blueprint_v.2.0.pdf

24 What transparency should be required in relation to web crawlers?

Please give us your views.:

AI developers should be required to publicly disclose at least, the following elements:

(1) list of all crawlers deployed by the model developer, and
(2) list of other solutions for expressions of rights reservations honoured by the model providers including information on the period of time as of which these solutions have been honoured.

25 What is a proportionate approach to ensuring appropriate transparency?

Please give us your views.:

Please see the paper referenced in the answer to question 23.

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

Please indicate the anticipated costs of transparency measures.:

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

Please give us your views.:

Training data transparency-related requirements should be introduced via regulatory intervention. Without regulatory underpinning, it is unlikely that AI firms will provide meaningful and comprehensive information to all relevant stakeholders.

28 What are your views on the EU's approach to transparency?

Please give us your views.:

We support the transparency obligation in the AI Act. We consider the public transparency provision for training data to be one of the key elements of the EU regulatory approach, which has the potential to address many of the issues arising from the 'black box' approach currently taken by most AI model developers, which is highly problematic for a technology that is likely to be widely adopted by society. If the UK were to consider departing from the EU example, the approach should be to strengthen public transparency provisions by requiring more detailed training data transparency obligations.

Wider clarification of copyright law

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?

Please give us your views:

By ensuring that the UK legislative approach is aligned with the EU legislative approach.

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?

Please give us your views:

31 Does the temporary copies exception require clarification in relation to AI training?

Please give us your views:

The temporary copies exception is a critical element of any balanced copyright regime. It serves important public interest purposes such as accessibility (every digital hearing aid performs acts of reproduction that are not economically significant) or access to knowledge (web browsing). This exception should be maintained in its current form.

32 If so, how could this be done in a way that does not undermine the intended purpose of this exception?

Please provide further comments:

Encouraging research and innovation

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

Please give us your views:

The existing text & data mining exception for research could better serve its purpose of promoting scientific research by removing the “non-commercial” criterion, which frequently leads to interpretation challenges. Private third-party funding is a common occurrence in the research sector, as are public-private partnerships in research collaboration. Researchers often face difficulties when judging whether their research qualifies as being non-commercial. The EU exception for text & data mining for research has taken a different approach, which doesn’t require the research to be non-commercial in nature, as long as it is performed on a not-for-profit basis, or by reinvesting all the profits in its scientific research, or pursuant to a public interest mission. This approach is more attuned to the realities of research funding, however the EU exception has the disadvantage of not applying to individual researchers. The UK could improve its exception by replacing the non-commercial criterion with a public interest criterion.

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

Please give us your views:

It is important to note that text & data mining has many purposes other than AI. For example, any time a journalist performs data analysis in the context of their reporting, they are performing text & data mining. A journalist at a news company cannot rely on the existing text & data mining exception for non-commercial research. When introducing a new text & data mining exception, the UK government should ensure that it is formulated in a technology-neutral manner that allows for its application in contexts unrelated to AI.

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

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.

No

Please give us your views:

No. Copyright protection should be granted to human authors only. Outputs without a human author should be in the public domain. Considering the increasing amount of AI-generated materials in circulation, we believe that awarding exclusive right protection to all of these materials bears the risk of disincentivising human creativity and makes it very hard for other parties to understand who owns the rights to any AI-generated output, particularly if there is no party that has an active interest in exploiting their rights. As far as AI providers are concerned, sufficient protections and incentives exist in the form of other exclusive rights, such as trade secrets and other intellectual property rights.

36 Do you have views on how the provision should be interpreted?

Please give us your views:

CGW Policy Option 1: Reform the current protection to clarify its scope

37 Would CGW legislation benefit from greater legal clarity, for example to clarify the originality requirement? If so, how should it be clarified?

Not Answered

Please give us your views:

38 Should other changes be made to the scope of CGW protection?

Please give us your views:

39 Would reforming the CGW provision have an impact on you or your organisation? If so, how? Please provide quantitative information where possible.

Not Answered

Please give us your views:

CGW Policy Option 2: Remove specific protection for CGWs

40 Are you in favour of removing copyright protection for computer-generated works without a human author?

Yes

Please give us your views:

As laid out above, copyright protection should only be granted to works of human authorship. This does not mean that the assistive use of AI precludes copyright protection. On the contrary, where AI plays an assistive role in the creative process, but where there is significant human involvement, the work should be eligible for copyright protection. While this determination will not always be easy to make in practice, all the necessary tools exist in the copyright framework.

41 What would be the economic impact of doing this? Please provide quantitative information where possible.

Please provide further comments:

42 Would the removal of the current CGW provision affect you or your organisation? Please provide quantitative information where possible

Not Answered

Please give us your views:

Infringement and liability relating to AI-generated content

43 Does the current approach to liability in AI-generated outputs allow effective enforcement of copyright?

Not Answered

Please give us your views:

44 What steps should AI providers take to avoid copyright infringing outputs?

Please give us your views:

Any measure to avoid copyright infringing outputs has the potential to affect legitimate uses and should not be adopted without proper users rights safeguards.

AI system-level measures to prevent output similarity, such as keyword filtering or other filtering measures that are triggered by an interaction with the end-user, entail non-negligible risks to fundamental rights. Users rights considerations come into play when output similarity is the result of an intentional act of “extraction” (e.g. specific instructions) by the end-user to cause an AI system to generate outputs similar to copyrighted works. The user may employ selective prompting strategies to elicit an AI system to generate an output similar to a copyright-protected work without infringing copyright. A similar output can only be qualified as an infringing output if 1) the output triggers the application of copyright law, which is not always the case (e.g. stylistic similarity has no copyright relevance, since artistic style is not protected), 2) the copyright-relevant output does not qualify as an independent similar creation and 3) no copyright exception or limitation (e.g. quotation, caricature, parody and pastiche) applies to the similar output.