

Hugging Face, Inc. 20 Jay Street, Suite 620, Brooklyn, NY 11201

Hugging Face Response to the Copyright Office Notice of Inquiry on Artificial Intelligence and Copyright

Hugging Face commends the Copyright Office on its extensive work framing the different aspects and new questions pertaining to copyright in Al systems inputs and outputs. The following comments are informed by our experiences as an open platform for state-of-the-art (SotA) Al systems, working to make Al accessible and broadly available to researchers for responsible development. We first provide a summary answer, further comments are then organized by section and by question. If a section or question is not highlighted, we do not have specific, actionable feedback.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI.

Summary: Artificial Intelligence and Copyright

Machine Learning (ML) techniques and AI systems have received increased attention in recent years. In particular, the growing popularity of commercial products described as "generative AI" has raised new questions about the societal impact of this technology, and about the distribution of its benefits. Among these questions, the role of the original data creators – including artists, journalists, and all categories of media creators – and the impact that these new tools will have on their livelihood and ability to keep creating new original media have been at the forefront of conversations. In order to help navigate these conversations, and especially the role of copyright in answering these questions, we provide comments based on our experience developing, studying, and sharing the various components that make up AI systems – with a view to helping regulators understand the likely consequences of different policy decisions on all stakeholders.

Generative AI systems are created in multiple stages. AI systems are first and foremost a representation of their **training datasets**. In order to extract the correlations and information that will support the systems' capabilities, developers typically start by curating very large corpora combining publicly available digital media with private collections. These datasets are then used to train **pretrained models** that encode some of the statistics of the training data. The pretrained models are then **adapted to specific uses** through a process called fine-tuning that may make them better suited to specific tasks, like translation or summarization, or help align them to the



expectations of a user interacting with the model through, *e.g.*, a chatbot interface or image processing software. Finally, the fine-tuned models are **deployed in commercial products or demos** that handle user requests, computational resources and hardware, and model outputs to allow the vast majority of users to use them for specific purposes.

Understanding this development chain is particularly important. In practice, <u>copyright</u> and <u>Fair</u> <u>Use assessments</u> will be very different at each of the stages – whereas sharing training datasets and pre-trained models <u>enables the very kind of beneficial uses</u> that the Fair Use doctrine aims to enable, answering questions about the impact of the technology on the market of the creators will often depend more on choices made at the fine-tuning and deployment stages. We caution that making requirements of all actors along the development chain based on decisions made at the deployment stage may end up taking compliance out of the reach of anyone but a few very well resourced companies, concentrating their authority and eventually harming the very content creators the requirements aim to protect – *e.g.*, by limiting their ability to <u>negotiate for application</u> <u>conditions of the technology</u> that are more aligned with their needs. We note in particular that early approaches to licensing data for generative Al training have shown very limited benefit to data creators given the scales involved – with <u>median pay-outs at under 0.01\$ per image</u> in one estimation, even as the 4M\$ total price tag for that deal takes it outside of the scope of most academic and smaller actors' budgets.

As an alternative approach to balancing the needs of different stakeholder groups, we recommend instead focusing on **transparency requirements** and supporting content creators' ability to **opt out of ML training** within a clear legal framework to support open development of generative AI datasets and pre-trained models by a broader range of actors. Open-source development of and open research into generative AI systems <u>support broader participation</u> in the development of this new technology, <u>foster more responsible AI</u>, help <u>avoid market</u> <u>concentration</u> within a few well-resourced actors, and allow the development of tools that <u>let</u> data creators directly control whether and how their data is used. Transparency and opt-out based approaches are also important features of <u>recent statements on generative AI</u> by organizations like the Software Heritage Foundation, and can foster <u>international consistency</u> with regimes such as the EU directive on Copyright in the Digital Single Market and proposed AI Act.

Recent discussions have also focused on the intellectual property status of the output of an AI system. Given our own experience with AI tools, we feel that the current requirements for a work of human authorship to be eligible for copyright protection are already well equipped to handle this question. As generative AI tools become part of the toolbox of most digital artists, the amount of work by the artists to elicit, process, and integrate the output of the systems into their creative process will remain the most important factor. Granting automatic rights to model developers or users of AI systems regardless of the amount of human work supporting each creation would facilitate displacement from the original creator communities, rather than enabling them to leverage these new tools to support their own creativity.



General Questions

Question 1

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

Generative AI systems can support a range of new tools to support and facilitate creative processes. These tools can make some forms of media generation more broadly accessible, and support artists and creators in their activity. Text-to-image systems make it significantly easier to create prototypes of images from language inputs. Image-to-image systems enable new kinds of image manipulation that can support new genres of artistic expression.

These tools also have the potential to displace labor and further destabilize an already precarious creator economy. Even though generative AI tools cannot by themselves replace an artist's voice, creativity, and thoughtfulness, companies may still choose to behave as if they did - creating lower quality media for much cheaper, and directing financial resources to the developers of the tools rather than the groups of workers who ensure that new art and media are created. This risks making being a working artist less sustainable, and have a drastically negative effect on the diversity of the artistic commons.

Specific development choices matter in ensuring that the technology's benefits are more evenly distributed without having a negative impact on creativity and creators. These choices are spread all along the development chain, from selecting specific training data, choosing how to use it to train models, providing sufficient transparency about the role of training data in the model outputs, etc. We address the specific impacts of each of these aspects in the rest of this response.

Question 3

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

- The paper introducing the original GLAZE technique, which allows artists to limit the utility of their published work when training generative AI systems with current techniques:
 - https://arxiv.org/abs/2302.04222
- <u>AI Art and its Impact on Artists</u> studies the broad impact of generative AI on practicing artists and makes several recommendations for making the technology more sustainable for artists:
 - o https://dl.acm.org/doi/10.1145/3600211.3604681



- <u>Foundation Models and Fair Use</u> analyzes several development and deployment scenarios for foundation models, including generative AI models. In particular, their analysis showcases how fair use determination have to consider the specific deployment and use of the technology:
 - https://arxiv.org/abs/2303.15715
- <u>Talkin' 'Bout Al Generation: Copyright and the Generative-Al Supply Chain</u>
 <u>https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551</u>
- Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and <u>Other Technologies</u>
 - <u>https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4564304</u>
- SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
 <u>https://arxiv.org/abs/2308.04430</u>
- The <u>BigCode</u> approach to training a Large Language Model for code generation uses an opt-out process that relies on GitHub authentication for verification
 - https://hf.co/datasets/bigcode/governance-card#consent-of-data-subjects

Question 4

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? (40) How important a factor is international consistency in this area across borders?

Recent regulatory developments in the EU can provide inspiration for possible approaches to managing Content under copyright in training datasets.

The EU directive on Copyright in the Digital Single Market provides for broad use of publicly available content for Machine Learning, including content under copyright, through a Text and Data Mining exception with exemptions that aims to strike a balance between the needs of data creators and the benefits of allowing broad use of public information for technology development. Specifically, the directive allows developers to use analytical techniques on digital data to "generate information which includes but is not limited to patterns, trends and correlations" as long as it respects the ability of data creators to opt out of having their content included. Operationally, this approach is currently limited by the lack of broadly recognized protocol for enacting that opt out, and the lack of transparency (and trust) on whether exemption requests are respected by developers – although recent approaches to <u>collecting centralized</u> preference signals and relying on <u>governance mechanisms of the data hosts</u> have shown promise. Legally, it provides data creators with a minimum level of agency by guaranteeing their right on content they produced, providing a basis for individual and collective negotiation.

More recently, the EU Parliament version of the EU AI Act introduced a requirement that Foundation Model developers provide a "*sufficiently detailed summary*" of the content under copyright in the model's training dataset. Such a summary would serve several purposes to enable better governance of creative content. First, it builds trust that opt-outs and expressed wishes of data creators are indeed respected. Second, and more importantly, it allows copyright assessments in the context of ML development and deployment to remain future proof. Relevant aspects of copyright assessments, including the propensity of a model to copy items from its



training data verbatim, are not a fixed, uniform, and homogenous property of all ML approaches - they depend on specific development choices. Transparency about copyright data used in various versions of the technology will give courts access to necessary information when assessing what constitutes relevant case law, rather than having to rely on information provided by interested parties.

Training

Question 6

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

Different types of organizations provide different amounts of transparency into the training materials used to develop their AI models. Unfortunately, the overwhelming majority of better-resourced companies developing these systems – especially companies developing systems for the purpose of integrating them in their own commercial offerings – tend to provide limited information about their training corpora. In recent years, organizations like Google, OpenAI, or more recently Anthropic have gone from providing high-level information about these datasets in the form of datasheets (Google PaLM's first iteration) or short dedicated sections in technical reports to withholding all information in more recent releases (PaLM version 2, Dall-E 3, GPT-4, Claude). These organizations may see the composition of their training data as a competitive advantage, fear legal exposure for data uses of uncertain legality, or simply choose to deprioritize the work involved in sharing and documenting datasets.

This lack of transparency negatively affects data creators and rights holders in multiple ways. First, it deprives creators of the ability to make informed decisions about whether and how to publish their work, and from leveraging the full range of social (such as preference signals) and technical tools (e.g., GLAZE, Nightshade, Kudurru) that can help them control how their work is leveraged in technology development. Second, it puts data creators at a disadvantage when negotiating conditions for use of AI systems with the technology's developers and deployers (see the recent WGA strike and agreement). Third, it makes determinations related to Copyright and Fair Use all the more difficult. As we argue in the rest of this response, determinations about most of the Fair Use factors are contingent on specific development choices throughout the development choice, starting with the dataset selection and curation. For any scope of copyright protections granted in the context of the use and development and use of generative AI tools, leveraging those rights will require a minimum level of transparency to the rights holders.

In this context, most of the information available about the prevalence and nature of copyright-protected materials in AI model training comes from a handful of open and collaborative research efforts typically hosted by non-profit organizations, smaller companies, and open-source communities. Corpora such as <u>The Pile</u>, <u>The Stack</u>, <u>ROOTS</u>, <u>DoLMA</u>, and <u>LAION</u> were developed with a focus on documentation, transparency, and re-usability, and have been used to train models that have often shown themselves competitive with their counterparts trained on more obscure datasets – albeit with less public visibility. These corpora are often provided along with more detailed analysis, tooling to support investigation by non-technical stakeholders, and publicly available information about the sources included. Analysis about copyright questions in generative AI contexts has typically used this information as a proxy to



make up for the lack of transparency of commercial systems. While this approximation has been necessary to enabling important investigation into the likely role of copyright-protected data in a new range of commercial systems, it can unfortunately also put undue pressure on research and open development efforts by abstracting away the meaningful differences between systems developed for the commons and systems developed for private use.

For the organizations that do have a consistent track record of providing substantial transparency into the data collection and curation process, it is common to train models on a wide array of copyright-protected texts such as books, academic papers, computer code, and more. These materials are typically obtained from publicly available sources. These sources include re-purposed previously published ML datasets, pre-crawled web sources (typically by processing data obtained by CommonCrawl), and web sources that are scraped by the dataset curator directly.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

The curation, sharing, and re-use of Al training resources has different stakes and consequences for research organizations and open-source and open science actors than for larger companies. In collaborative research and development settings, curating a dataset can represent a significant resource investment for a single actor that is typically amortized by ensuring that it can easily be analyzed and re-used. These dynamics encourage the curation of datasets that fill existing resource gaps in the commons, prioritize transparency, and enable external scrutiny into the biases encoded in a model by looking at their sources in the training dataset (for example on how biases scale with dataset sizes). They also make model training accessible to smaller actors whose work is typically more aligned with fair use conditions. Conversely, while better-resourced companies may choose to prototype on private datasets, they can typically afford to re-create datasets "from scratch" behind closed doors, limiting their exposure to copyright suits. This opaque situation is also untenable if opt-outs and preference signals, since every company may use different standards for what data they include or which opt-out signals they respect.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

Little information is publicly available about the specifics of licensing agreements with copyright owners. The two most visible such agreements have occurred between OpenAI and the Associated for text, and ShutterStock for images. The Adobe Firefly system was trained on stock images whose copyright was owned by Adobe. A <u>recent analysis of the Shutterstock/OpenAI</u> <u>licensing deal</u> finds an estimated median payout of under 0.01\$ per image and 18.5\$ per artist for their entire portfolio.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

This is highly dependent on the policies and procedures of individual labs. There is no inherent requirement to do so, however we expect it to be common. Multiple different models are frequently trained on the same data and so retaining copies of the training data decreases duplicative data processing work. Data processing is a substantial amount of work and time, so discarding datasets after training would meaningfully increase incurred costs in the long term. It



is common for dataset processing to be a multi-step endeavor, with each step addressing a different issue (e.g., removing low-quality documents, removing PII, improving formatting, combining datasets from multiple sources). As a result, it is common to store not only the final dataset but also many intermediate stages of the dataset with different amounts of preprocessing performed. This enables researchers to decide to do preprocessing differently for a future model without having to start again from scratch.

There are some important fields of research for which having access to pretraining data is important beyond merely training the models. For example, research into memorization studies the tendency of large language models to generate long exact sequences from their training corpus and the possible causes and mitigations. It is widely accepted that the gold standard for this research requires access to the training data. For example, recent work by (Biderman et al., 2023) provides passages from the training data as an input to the model to study if the model can continue them correctly. Deleting the training data would make this type of research very difficult. We see the difficulties in doing research into memorization without access to the training corpus in papers like "Speak. Memory: An Archaeology of Books Known to ChatGPT/GPT-4" (Chang et al., 2023) where the authors seek to identify what books OpenAI's ChatGPT has memorized. A key difficulty identified in the paper is the need to make a guess as to what data is in the training corpus and the fact that the quality of this guess can have a substantial impact on the results of the analysis. This challenge is reiterated in many other papers that seek to study memorization without access to the training data. In "Quantifying Memorization Across Neural Language Models" (Carlini et al., 2022) the authors write "We leverage access to each model's original training set to provide order-of-magnitude more precise bounds on the amount of extractable data than in prior works." and "Our research would not have been possible without EleutherAl's complete public release of The Pile dataset and their GPT-Neo family of models."

Question 7

7.2. How are inferences gained from the training process stored or represented within an AI model?

Inferences are stored as model weights. Model weights apply transformation based on those inference to inputs at use time. While their final state is shaped by interactions between the AI system and data points during training, currently used techniques do not store copies of these data in the weights. Some weights and combinations of weights ("neurons") have been shown to represent specific words, but the relationship is usually more complex. The impact of a weight or training example on the model behavior is always stochastic - we have limited causal understanding to support most counterfactuals (i.e., how would the model have differed if a specific training point had been omitted or modified), especially about general questions spanning multiple use cases. In other words, while a model may be found *a posteriori* to reproduce sequences or images that are featured in its training dataset, especially if these sequences are repeated multiple times, it is next to impossible in general to predict *a priori* how a specific data item or document will influence a model's output in a given situation.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?



This is an open technical question and there is not conclusive evidence in any direction at the present. There is some preliminary evidence that models can "unlearn" inferences gained from training, but these results need substantially more study before they can be accepted. In particular, the robustness of these techniques to changing prompts and changing deployment context is largely unknown. There are also some promising approaches to accomplish similar goals to machine unlearning that do not involve retraining the model. Research into ethics and bias of machine learning models has studied the extent to which a model can be made to "not use" certain types of information. While this doesn't remove the inference gained from training from the model itself, it would prevent those inferences from being leveraged in user-facing applications.

Whether or not unlearning is technologically and economically feasible, it is not a desirable policy solution to models learning undesirable inferences from training data. There is no way for an organization training a model to conclusively demonstrate to a third party, regulating agency, or the public that any unlearning methodologies they may have applied consistently produces the desired outcome, i.e., that the model behaves as if it had been trained without the "unlearned" data.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

In some cases yes (Carlini et al., 2022; Chang et al., 2023; Biderman et al., 2023), but it's very hard to know how reliable or complete these results are. In particular, **such processes are unable to give negative results**: either they find evidence that the model was trained on a particular price of material or it is inconclusive. One cannot rule out that a model was trained on data without examining the training data. This relates to the discussion in 6.4 of the study of memorization and how it is made much easier by access to the underlying training data.

Question 8

8. Under what circumstances would the unauthorized use of copyrighted works to train Al models constitute fair use? Please discuss any case law you believe relevant to this question.

When properly articulating the different stages of AI development, from dataset curation, to model (pre-)training, to model adaptation and deployment, the earlier stage of dataset curation and general training of AI models as typically done and documented are typically aligned with the fair use framework.

Much of the recent success of AI, both "discriminative" and "generative", can be traced back to the growing popularity of pre-training methods. While earlier ML models used to be trained from the ground up for each specific application up until the early 2010s, multiple innovations over the last decade have combined to shift the field as a whole to a new paradigm where a single model is first "pre-trained" on a very large amount of data to encode generally useful statistics about a modality (text, speech, images, etc.). These statistics give models a "leg up", so to say, when they're being adapted to specific applications - allowing developers to build functional systems for such tasks as automatic summarization, object detection, or to be used as a chatbot with many fewer annotated examples.

As the size of the models and pre-training datasets have grown by orders of magnitude, these statistics have become more complex. For example, while early text-based pre-training methods



focused on simple co-occurrence counts between words or sentence-level phenomena in text, modern Large Language Models can encode probability distributions over longer sequences well enough to generate entire documents that look not unlike what a person could write. Models of software code, images, speech, music, and multimodal data have seen similar progress in recent years - enabling models that reach higher performances on multiple question answering, transcription, image classification, and many other benchmarks. Notably, modeling higher-order statistics at the level of an entire image is what allows models to generate new images by sampling from its learned distribution. In some cases, the model may additionally capture correlations between features of the works of a specific data creator, allowing models to be prompted to generate images that are reminiscent of the style of the specific creator, but this behavior is hard to predict *a priori* and depends on specific technical choices in model training and data processing.

Insofar as this approach primarily aims to encode broadly useful statistics and common patterns across its training examples, we broadly think that the inclusion of data under copyright among these examples aligns with existing law. The data is observed in training and not "stored" in the final model in any applicable sense. More generally the use of a given work in training is of a broadly beneficial purpose: the creation of a distinctive and productive AI model. Rather than replacing the specific communicative expression of the initial work, the model is capable of creating a wide variety of different sort of outputs wholly unrelated to that underlying, copyrightable expression. For those and other reasons, generative AI models are *generally* fair use when they train on large numbers of copyrighted works. We use "*generally*" deliberately, however, as one can imagine patterns of facts that would raise tougher calls.

While "the market for the work" in factor 4 has never swept so widely as to cover an entire domain or occupation (e.g., the market for singing as opposed to the market for a particular communicable expression), cases may get closer to the line and that have relevant impacts on a person's labor. For instance, one might imagine a model trained on a highly narrow set of works specifically for the purpose of replacing those communicative expressions in the market. A model might be trained on only a particular singer's copyrighted songs, or in a way that makes it more likely to generate items that are particularly similar to specific training examples, with the express purpose not to parody or comment but rather to create replacements. Importantly, copyright law has the ability to address these edge-cases, through case-by-case analysis under fair use. Some facts may only be present and able to be fully dealt with at later stages of an evaluation – i.e., assessment of how a particular model was deployed or made available as a product for sale.

Ultimately this needs to be a decision made by a court on a case-by-case basis. As stated above, training in commercial settings may still be fair use, such as for the purpose of learning general fact patterns. We recognize that ensuring that fair use remains consistent can put an undue burden on copyright holders, and decrease their ability to advocate for their broad interests. In order to mitigate those risks, <u>categories of stakeholder</u> and other jurisdictions have turned to transparency (<u>EU AI Act</u>) and opt-out (<u>EU CDSM</u>) requirements.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

The quantity of data depends on the modality of the models being trained and the application context. For text generative models, it is common to train on 100 billion to 5 trillion "tokens"



(units of data), which is approximately 600 million to 30 billion pages of single-spaced Times New Roman English text. For text-to-image generative models, it is common to use between 300 million and 5 billion images. Finetuning a model requires far less data, sometimes requiring as few as 30 pages of text or 100 images for finetuning to perform a specific task.

The quantity and diversity of data used to train so-called "general purpose" Al systems is an essential component of ensuring that they do not adhere too closely to any particular datum. The primary reason that text models require billions of pages of text is that if models are trained on much less data (perhaps repeated multiple times) as is common in some fields of machine learning, the model doesn't learn the general purpose objective and cleaves too closely to the training data to be useful in diverse application contexts. Similarly, finetuning models on small quantities of text is most effective when the goal is to create a model that performs a specific task following a specific style such as producing art that resembles the work of a specific artist.

This speaks primarily to the purpose and character of the use. Large and diverse training corpora enable the model to use the training data for the purpose of learning general representations and concepts from the data. Smaller, more application-specific finetuning corpora causes the model to use the training data for the purpose of learning and mimicking specific stylistic elements of the training data. While the size of the training corpus does not determine the character of use in isolation, it should play a role in determining whether the purpose that the data was put to was learning general representations or mimicking style.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? (46) Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

The fourth factor analysis centers on the communicable expression of a given work (or works). It has not to our knowledge previously been interpreted as preventing competition generally among users and developers of new tools.

In Google Books, the court focused "on whether the copy brings to the marketplace a competing substitute for the original, or its derivative, so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original." Nor did the possibility of a licensing market for text mining undermine the finding of fair use in Hathitrust, given that "[I]ost licensing revenue counts under Factor Four only when the use serves as a substitute for the original and the full-text-search use does not."

The narrow interpretation of the market for or value of a work is made clear in Campbell vs Acuff-Rose, where the Court considered whether a parody song sold commercially is a market substitute for the original song. The Court finds "[a]s to parody pure and simple, it is unlikely that the work will act as a substitute for the original, since the two works usually serve different market functions." The market for a parody song vs the song it is based on is substantially more similar than the market for an AI model vs its training data.



Question 9

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

While opting into the use of work as training material may be a medium to long-term goal, it is not currently feasible to seek opt-ins for already published data - especially as the majority of data under copyright on the web does not have an easily identifiable rights holder.

Opt out can be achieved through a combination of recognized machine-readable opt-out formats and contestation rights - where rights holders may request that their data be removed from new versions of a maintained dataset (see e.g. <u>BigCode</u>).

<u>Several standards are currently being used</u> on different platforms to enact opt-outs. The success of opt-outs as a governance mechanism will depend on having sufficient operational guidance on the formats accepted to express reservation. In order to facilitate the adoption of formats that work for various platforms (code repositories, discussion forums, etc.) without leading to unnecessary fragmentation, we recommend that the Copyright Office maintain a repository of recognized formats.

In order to make opt-outs functional and trustworthy, we additionally recommend that each platform use a single format, and that appropriate transparency requirements allow rights holders to check whether their reservations have been respected for a given dataset. We note that this can be achieved even when the dataset is not made directly available by hosting a <u>search index</u> or a <u>membership test</u> to check whether a data item is present in the dataset, or by providing <u>sufficient metadata</u> about the items included in a dataset. We advise against letting dataset curators and model developers specify their own formats for opt-outs, as this can put a disproportionate burden on rights holders to keep track of all possible formats.

Questions 12-14

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

This is an open technical question. However at present there is no known way to do so reliably. While some techniques do exist that allow the outputs of a generative AI system to be distinguished from "natural" data, they are not secure against someone who wishes to adversarially conceal that a work is machine generated.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

Given the quantity of works used in current training of generative AI systems, we are concerned that a licensing requirement would lead to significant market concentration without sufficiently benefiting data creators. An outcome where licensors pay millions of dollars to train on hundreds of thousands or millions of works under copyright would constitute a "worst of both worlds" outcome in our assessment, as such a deal would be costly enough to exclude any but the very largest companies from training new models, while still providing negligible additional income to the original data creators. Cheaper datasets would provide even less to the creators, and more expensive ones would further restrict the set of possible developers. Additionally, we are



concerned that the need for a more immediate return on investment for this data would motivate developers to capture more of the market for the base data, and to develop tools that replace rather than support the original data creators.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

Given the state of the technology, sufficient training data documentation is the only way to gain certainty regarding what works were used. It is therefore in the interest of rights holders that this information is made available.

Transparency and Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

Yes

15.1. What level of specificity should be required?

A broad description of all training data sources should be considered part of standard model documentation. While it's not practical to provide a list of the millions of works and authors that appear in the dataset (let alone list all the untitled content with no identified author), it is possible to create search tools that allow rights holders to look up relevant information without displaying the entirety of the work. Such a hybrid system has already been executed in practice and provides relevant information to all interested parties (search index, metadata explorer).

15.2. To whom should disclosures be made?

First, different groups of stakeholders should be informed about the data: the users of a given model, the people who created the training data, and relevant regulatory agencies. For models trained on widely available Internet data, we believe data documentation should be made public. The only exceptions to this should be carefully limited to protect the privacy of data subjects in domains such as healthcare and medicine.

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

The entire model supply chain needs to be available to the users to allow them to understand the limitations of a given AI system and to make informed consumer decisions.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

The cost of proper documentation is marginal when done along with the initial research, but can be significant when done *a posteriori*. We believe the broad benefits are substantial, both from a copyright perspective and aligning with broad recommendations about responsible Al development and public interest.



Generative AI Outputs

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?

As machine learning practitioners, we find that very little to no innovation in generative AI is driven by the hope of obtaining copyright protection for model outputs. The incentives for innovation already exist without modifying copyright law. On the other hand, lack of copyright protection for wholly AI-generated material gives leverage to the professionals who might otherwise find themselves displaced by generative AI tools. This suggests that legal protection for AI-generated material is neither desirable nor necessary for innovation.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs? (53)

Open AI models pose different challenges, and introduce different opportunities with respect to the copyright status of their outputs. Open models released without sufficient documentation can make it difficult for users to evaluate whether a given output is substantially similar to an example in the dataset. Conversely, open models released with sufficiently documented datasets open the possibility for users to assess similarity more broadly by examining the closest training examples themselves, and thus lower the risk of producing similar outputs to their training data that would be missed by an automatic filter. This is a concern as notions of substantial similarity will have to take the continuous nature of model outputs into consideration, which makes existing tests insufficient.