gettyimages®

iStock™ Unsplash

195 Broadway, 10th Floor, New York, NY 10007
(800) IMAGERY (800) 462-4379 |www.gettyimages.com

legal@gettyimages.com for related inquiries

December 5, 2023

Reply Comments for USCO Inquiry on Artificial Intelligence and Copyright

Docket No. 2023-6

Getty Images appreciates the opportunity to respond to the initial round of comments submitted in response to the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright (the "NOI"). As with our initial response to the NOI ("Initial Comments"), our reply comments focus on Generative AI Models and Generative AI Systems that are trained with copyrighted images and video and the corresponding captions and metadata. In particular, we focus on arguments offered in support of the claims that:

(i) developers of Generative AI Models and Generative AI Systems do not infringe copyrights when they reproduce and distribute copyrighted materials without permission from the copyright owners in the course of training models, even when those models and systems are deployed for commercial purposes and generate outputs that compete with the works on which they are trained, and

(ii) when Generative AI Systems deliver outputs that are substantially similar to copyrighted works on which the underlying model was trained, and therefore infringing, the users are solely to blame and there should be no liability for the model developers.

Both of these propositions are incorrect as a matter of copyright law and misguided as a matter of copyright policy.

That the copyrighted materials on which a Generative AI Model was trained were "publicly available" is not a defense to copyright infringement.

A number of responders observed in their initial comments that their Generative AI Models were trained on "publicly available information." But being "publicly available" and being in the "public domain" are very different propositions. The public availability of a work is irrelevant to the questions of whether the work is copyrighted and whether reproducing the work in the course of training a Generative AI Model without the permission of the copyright owner constitutes copyright infringement. Getty Images does not waive its copyrights or grant any implied licenses by making its copyrighted images available (albeit with prominent watermarks and copyright notices to deter infringement) for review by customers and prospective customers on its public-facing websites. Nor does Getty Images waive its copyrights when it grants a license to a customer to display one of its images on the customer's public-facing website. While the unauthorized reproduction of copyrighted works from a copyright owner's own website or another source authorized by the copyright owner to display the works infringes the copyright owner's exclusive rights absent a fair use defense or other legal justification, we note that some commenters are not only conflating "publicly available" with "public domain" but going so far as to include within the scope of "publicly available information" both works that are obtained from known piracy sites and works scraped in violation of express terms of use. While the public may be able to access such works, they are not "publicly available" for the taking. Courts routinely find defendants liable for infringement when they copy "publicly available" copyrighted photographs from websites for commercial purposes without permission from the copyright owner.[1]

The Copyright Office should emphatically reject efforts to elide the distinction between works that are "publicly available" and works that are in the "public domain" for purposes of assessing whether unauthorized copying is legally defensible.

Training a Generative AI Model – at least in the context of text-to-image or image-to-image generation – involves reproducing the expressive content of the materials on which the model is trained.

Some commenters have attempted to claim that training a Generative AI Model does not even involve acts of copying within the meaning of the Copyright Act. While there may be room to debate, in some circumstances, whether particular acts of copying can fall within the scope of a valid fair use defense, there is no question that training a Generative AI Model capable of text-to-image or image-to-image generation involves reproducing the expressive content of the images and associated captions and metadata on which the model is trained. Indeed, training a diffusion model capable of image generation typically requires at least the following acts of reproduction:

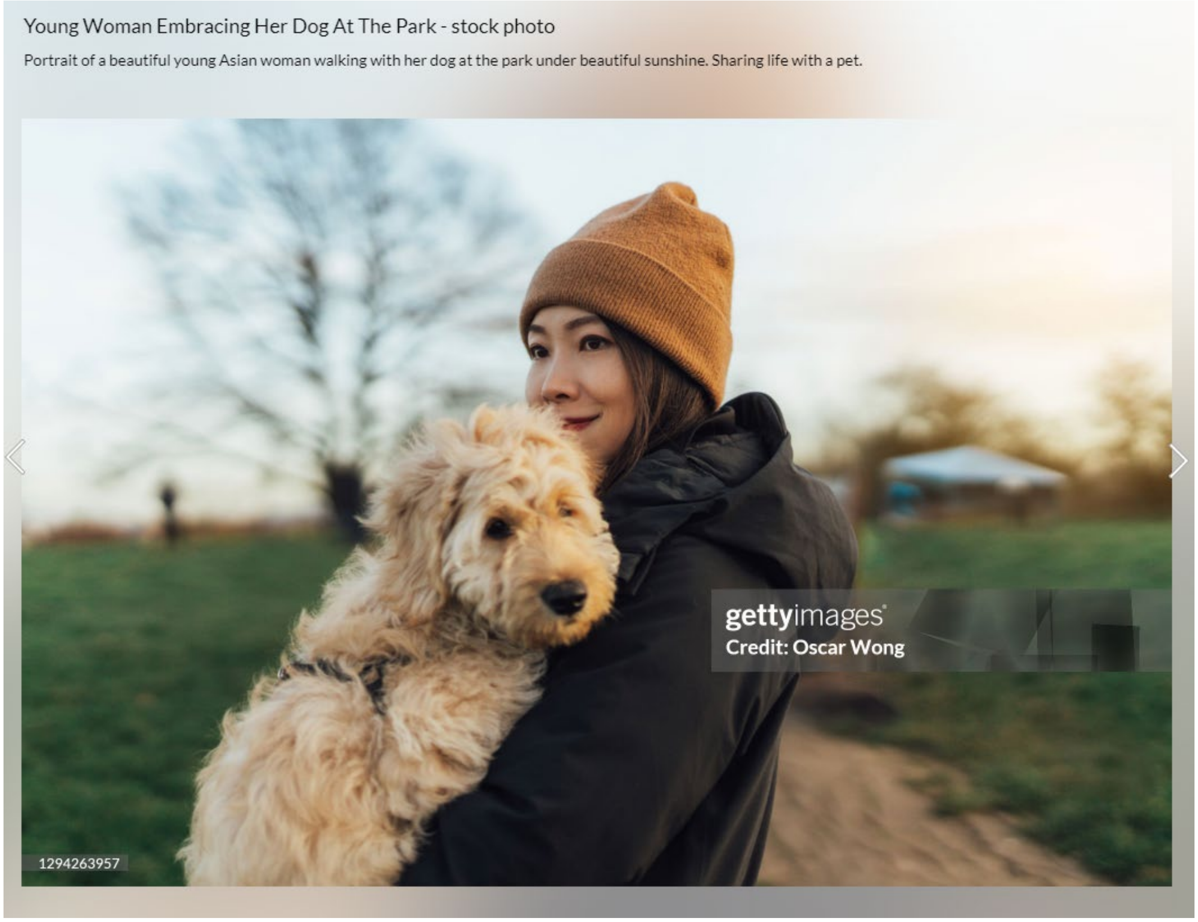
- Copying text-and-image pairings that the developer selects for its training set from the sources from which those pairings are obtained;
- Loading those text-and-image pairings into computer memory to train the model;
- Encoding the images to create smaller versions of the images that take up less memory and separately encoding the paired text, with the encoded images and text retained and stored as an essential element of training; and

[1] *See, e.g., Mango v. Buzzfeed*, *Inc.*, 970 F.3d 167 (2d Cir. 2020); *FameFlynet, Inc. v. Shoshana Collection, LLC*, 282 F. Supp. 3d 618 (S.D.N.Y. 2017); *BWP Media USA Inc. v. HipHopzilla, Inc.*, No. 1:14-CV-0016-AT, 2016 WL 4059682 (N.D. Ga. Mar. 1, 2016).

- Incrementally adding visual “noise” to the encoded images so that it is incrementally harder to discern what is represented and, in the process, creating new copies that, notwithstanding the addition of noise, are substantially similar to the original images (at least until so much noise has been added that they are no longer recognizably similar).

In each instance, the model developer is copying expressive content, not merely some fact or idea contained in the original work.

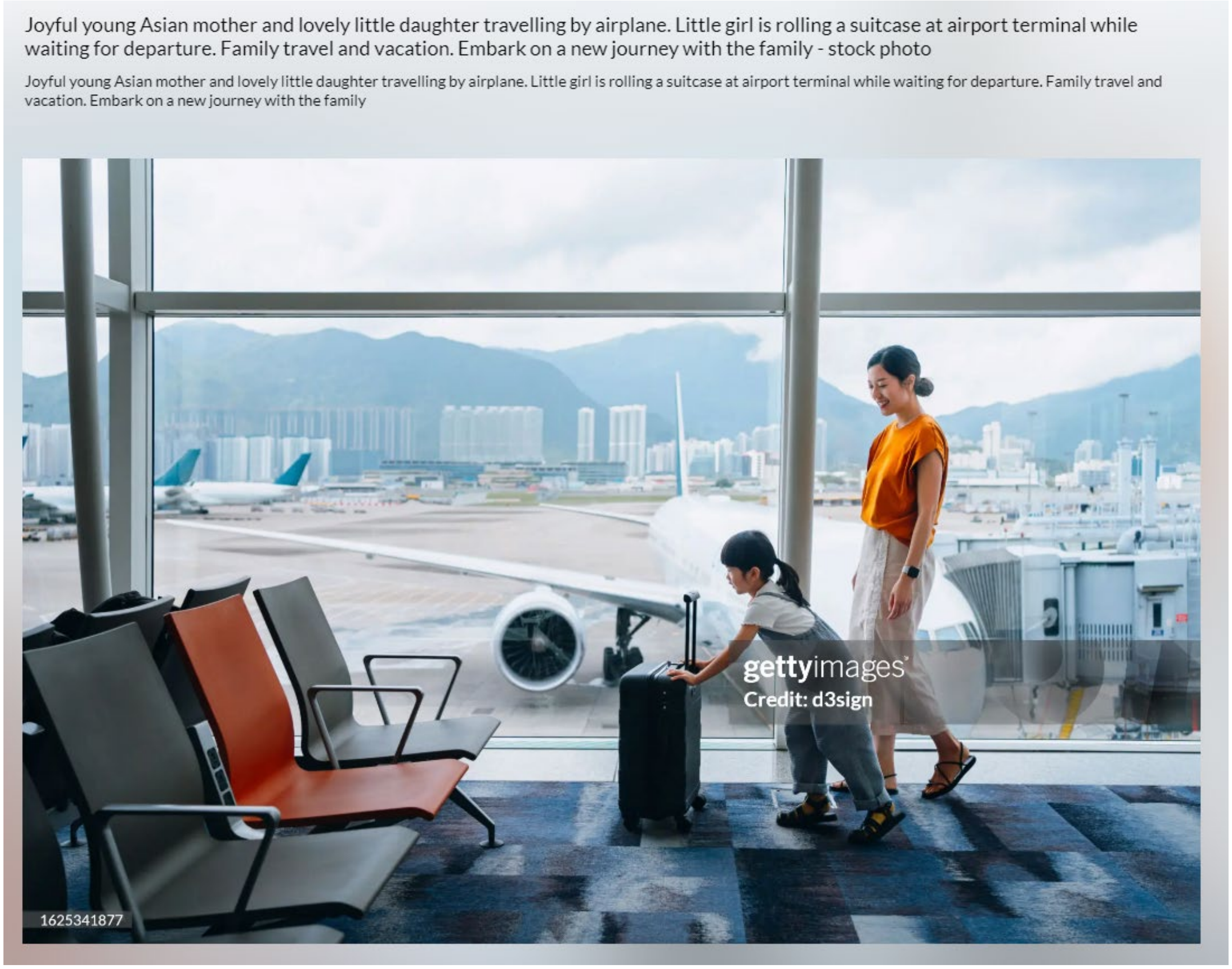
The expressive components of Getty Images’ copyrighted works are why those works are particularly desirable for use in connection with training Generative AI Models capable of image generation: the photographs are high quality, and the accompanying captions and metadata are richly descriptive. For example, if one were to use the text-image pairing set forth below to train a Generative AI Model, the photograph is much more helpful to the training process than a stick-figure drawing of a dog or a woman with a dog. And it is much more helpful to inform the model that this is a “[p]ortrait of a beautiful young Asian woman walking with her dog at the park under beautiful sunshine” and that this image is a reflection of someone “[s]haring a life with a pet” than merely that this is a picture of a “dog” or a “woman and a dog.”



Young Woman Embracing Her Dog At The Park - stock photo

Portrait of a beautiful young Asian woman walking with her dog at the park under beautiful sunshine. Sharing life with a pet.

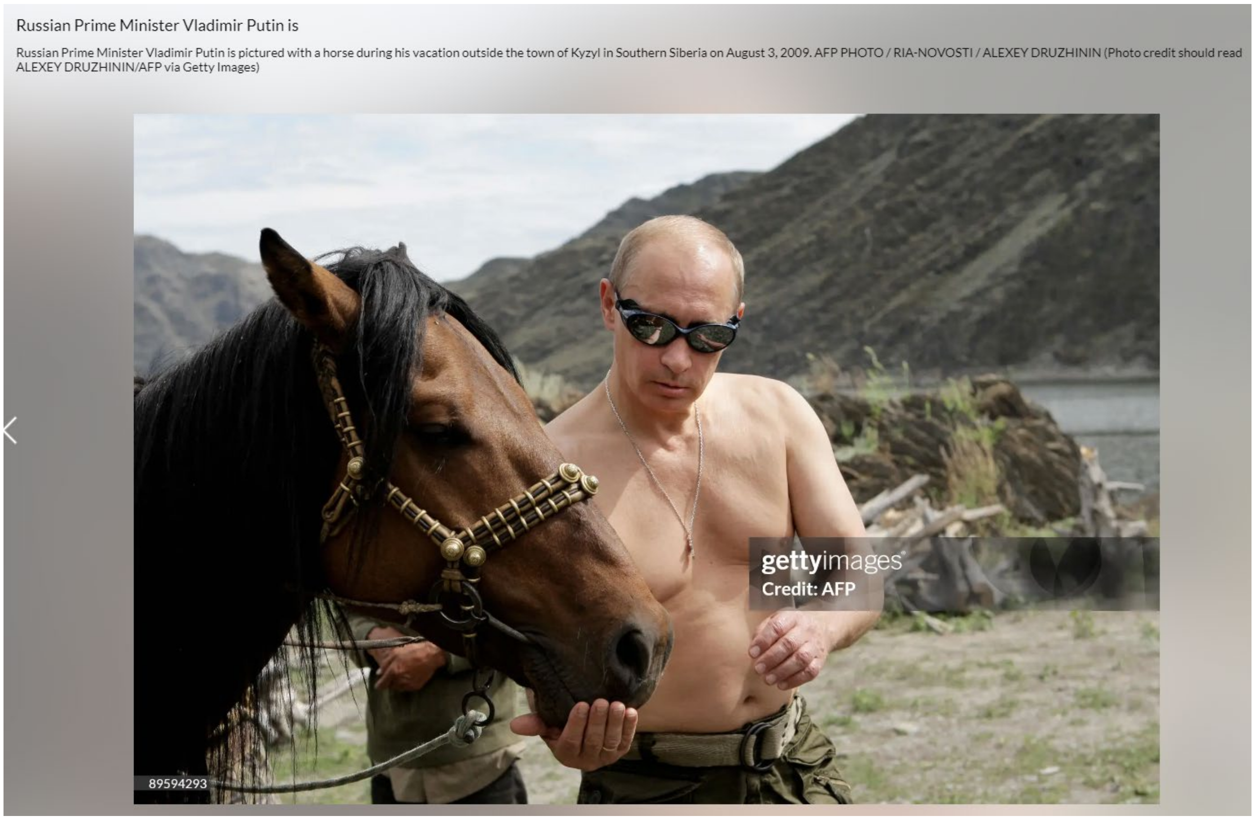
Similarly, if one were to use the text-image pairing set forth below to train a Generative AI Model, the photograph is much more helpful to the training process than a stick figure of a woman and a young girl or a photograph of a woman and a young girl that does not have the rich context of the surroundings and background. And it is much more helpful to inform the model that this is an image of a "[j]oyful young Asian mother and lovely daughter traveling by airplane," that the "little girl is rolling a suitcase at airport terminal while waiting for departure," and that it reflects both "[f]amily travel and vacation" and "[e]mbarking on a new journey with the family," than merely that this is a picture of a "woman and her daughter" or a picture of "a girl rolling a suitcase."



Joyful young Asian mother and lovely little daughter travelling by airplane. Little girl is rolling a suitcase at airport terminal while waiting for departure. Family travel and vacation. Embark on a new journey with the family - stock photo

Joyful young Asian mother and lovely little daughter travelling by airplane. Little girl is rolling a suitcase at airport terminal while waiting for departure. Family travel and vacation. Embark on a new journey with the family

Copying the expressive content of the text-image pairing is a critical driver of the effectiveness of the training. Moreover, it is not just the expressive content of the image and the expressive content of the text, but how those two forms of expression relate to one another. Consider the following text-image pairing for use in training a Generative AI Model:



Russian Prime Minister Vladimir Putin is

Russian Prime Minister Vladimir Putin is pictured with a horse during his vacation outside the town of Kyzyl in Southern Siberia on August 3, 2009. AFP PHOTO / RIA-NOVOSTI / ALEXEY DRUZHININ (Photo credit should read ALEXEY DRUZHININ/AFP via Getty Images)

Informing the model that “Russian Prime Minister Vladimir Putin is pictured with a horse during his vacation outside the town of Kyzyl in Southern Siberia on August 3, 2009” is far more useful than describing this image as “a man with a horse” or “a picture of Vladimir Putin.”

Any contention that training a Generative AI Model does not involve copying of expressive content is simply false.

Sony Corporation of America v. Universal City Studios, Inc. and the “staple article of commerce” doctrine do not excuse infringement arising out of the development and use of Generative AI Models and Generative AI Systems.

A number of commenters contend that Generative AI Models and Generative AI Systems are “staple articles of commerce” and therefore model developers are absolved from liability in light of the U.S. Supreme Court’s decision in *Sony Corporation of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984). These contentions ignore fundamental differences between the nature of copyright infringement claims copyright owners have asserted against developers of Generative AI Models and Generative AI Systems on the one hand and claims asserted against manufacturers of tools considered to be “staple articles of commerce” on the other.

Unlike the training of Generative AI Models and the offering of Generative AI Systems, the manufacture and distribution of the Betamax video recorder at issue in *Sony* did not involve *any* copying of copyrighted works. The only copying at issue occurred *after* the recorders had been sold to consumers, and Sony exercised no control over how consumers chose to use the recorders they owned. The plaintiffs could not assert direct infringement claims against Sony because Sony had not engaged in any copying; they could only assert claims for secondary liability to try to hold Sony accountable for allegedly infringing activities by Betamax users. Adopting the "staple article of commerce" from patent law, the Court held that the Betamax video recorder could be used for "commercially significant non-infringing purposes" and, therefore, Sony was not liable for secondary infringement on the facts presented.[2] The same is true for a variety of other technological tools that can be used to facilitate infringement, but are not themselves infringing, such as the typewriter, the camera, the photocopier, and the printer. In each instance, the manufacturer does not violate any of the exclusive rights afforded to copyright owners in the course of developing or distributing the product, and it does not exercise any control over what consumers do with the product after it has been sold.

Developers of Generative AI Models and Generative AI Systems, in marked contrast, do tread on exclusive rights afforded to copyright owners under Section 106 of the Copyright Act both when they reproduce those works in the course of training models and if they generate and distribute infringing output. That is why, in the copyright infringement cases brought to date against model developers, plaintiffs have sued developers for *direct* infringement. Some plaintiffs may *also* have sued under theories of secondary liability for infringement by users or commercial partners, but the "staple article of commerce" doctrine does not provide a defense to acts of direct infringement. Nor does the doctrine provide a defense to claims of vicarious infringement.[3] The "staple article of commerce" defense only applies to claims of contributory infringement and, even there, model developers are unlikely to qualify for such a defense given the nature of their offerings and the control they maintain when distributing those offerings.

In most instances to date, model developers are not simply injecting their products into the stream of commerce with no knowledge of how their products are used and no ability to

[2] *Sony*, 464 U.S. at 442.

[3] *See, e.g.*, *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1022-23 (9th Cir. 2001) ("*Sony*'s 'staple article of commerce' analysis has no application to Napster's potential liability for vicarious copyright infringement.").

control that use. To the contrary, most developers of Generative AI Systems are selling consumers and businesses *access* to the use of an underlying model, typically in the form of a subscription to one of the services they offer.[4] And this is hardly a passive engagement. Numerous comments submitted on behalf of developers of Generative AI Models and Generative AI Systems touted the steps those entities are taking to ensure the safety of the products and the guardrails they have in place, whether in relation to mitigating the likelihood of delivering infringing outputs or, more commonly, a variety of other harms unrelated to copyright. Numerous courts have found the distinction between selling a product and offering an ongoing service critical for purposes of evaluating whether the "staple article of commerce" doctrine provides a defense to a claim of contributory infringement.[5]

Accordingly, to the extent that some commenters have suggested that the "staple article of commerce" doctrine absolves model developers from potential liability for copyright infringement, they are mistaken.

By selecting copyrighted works for inclusion in training sets, and reproducing such works in the course of training, model developers have engaged in "volitional conduct" for purposes of copyright law.

Some commenters have attempted to suggest that entities that develop and offer Generative AI Systems cannot be liable for copyright infringement because they purportedly do not engage in "volitional conduct." These arguments are specious.

As an initial matter, these arguments do not apply to infringement claims based on unauthorized copies made in the course of training a Generative AI model. Such copying is deliberate and intentional. Developers choose what works they will copy and the sources from

[4] Such subscriptions are typically governed by contractual agreements between the user and the distributor of the service (whether in the form of a "click through" agreement or otherwise) that impose limitations on what the user can do with the service and reserve important rights for the distributor.

[5] *See, e.g.*, *Arista Records LLC v. Usenet.com, Inc.*, 633 F. Supp. 2d 124, 156 (S.D.N.Y.2009) (rejecting defendants invocation of *Sony* as a complete defense and finding non-infringing uses "immaterial" where there was an ongoing relationship with users); *see also Capitol Records, LLC v. ReDigi Inc*, 934 F. Supp. 2d 640, 658-59 (S.D.N.Y. 2013) ("*ReDigi*"), *aff'd*, 910 F.3d 649 (2d Cir. 2018) (finding, in addition to liability for direct infringement for defendant's own acts of reproduction and distribution, liability for contributory infringement where service "provided the site and facilities for their users' infringement") (cleaned up).

which they will obtain those copies. While a plaintiff may need discovery to know specifically which entity or individual was responsible for those choices, there is no room to dispute that whoever was engaged in the actual acts of reproduction did so with requisite volitional conduct for infringement liability, regardless of whether they understood their acts to be unlawful at the time. Copyright infringement is a strict liability tort.

There is ample volitional conduct for liability arising out of claims based on the delivery of infringing outputs as well, even though the user, rather than the developer, selects the prompts. But the prompts alone do not determine the output, as evidenced by the fact that a Generative AI System will typically deliver a variety of different outputs in response to the same set of prompts. And even if they did, in each instance, the output will reflect of how the underlying model was trained, including the choices that the developer made with respect to what works to include the training set, where and how to obtain copies of those works, what type of fine-tuning to provide, the extent of fine-tuning, whether to implement filters of datasets and, if so, which filters to use.[6] The decision to use copyrighted works to train a Generative AI Model and the selection of which works to use distinguish such models from other technological tools in which volitional conduct by the offeror was found lacking, even though users could use those tools to infringe.[7] Model developers are not "passive conduits"; their active engagement in the process by which a user is supplied with an infringing output supplies ample volitional conduct.[8] By selecting particular prompts, users also play a role that influences what content is delivered to them, but that does not absolve offerors of Generative AI Systems from liability for making and delivering an infringing output.[9]

[6] Several commenters (which are also defendants in pending copyright infringement lawsuits) touted that they filter datasets for issues like "safety," "bias," and "quality" or "conduct legal and ethical diligence on the data that we use." The decision not to filter datasets to exclude unlicensed copyrighted works is deliberate.

[7] *Compare Cartoon Network, LP v. CSC Holdings, Inc.*, 536 F.3d 121, 131 (2d Cir. 2008) ("*Cablevision*") *with Fox News Network, LLC v. Tveyes, Inc.*, 883 F.3d 169, 181 (2d Cir. 2018) (distinguishing *Cablevision*) and *ReDigi*, 934 F. Supp. 2d at 656-57 (same).

[8] *See Arista Records*, 633 F. Supp. 2d at 148-49.

[9] *See Am. Broad Co., Inc. v. Aereo, Inc.*, 573 U.S. 431 (2014) (reversing appellate court's determination that provider of service that facilitated subscriber access to Internet streams of copyrighted broadcast

The number of works that a developer may wish to use to train a Generative AI Model does not excuse unauthorized copying.

A number of commenters have complained that licensing the use of copyrighted works in training sets would be either impossible, impractical, or unduly expensive because of the sheer number of works some model developers would like use for training purposes. The scope of infringement in which an infringer would like to engage hardly excuses the infringement.

Voluntarily offering "opt outs" from the unauthorized use of copyrighted materials in training sets does not excuse infringement.

A number of responders who have engaged in the unauthorized use of copyrighted materials to train Generative AI Models and Generative AI Systems recognize that, generally speaking, the creative community vehemently objects to these uses and now claim to want to be sensitive to creator concerns. They suggest that a few half-measures, such as permitting copyright owners to "opt out" of their unauthorized uses, should be sufficient to resolve any objections to their infringing conduct.

Getty Images addressed the numerous reasons why an "opt-out" approach to AI training is not a suitable substitute for affirmative consent in our Initial Comments (at 18-19). But, in any event, forward-looking opt-out regimes are "too little too late" to excuse the infringement to date. While Getty Images of course welcomes reasonable dialogue with model developers on these issues, it is not credible for developers to feign surprise that copyright owners would object to the commercial exploitation of their copyrighted works in this manner, and so even if an opt-out scheme could have passed muster as a matter of copyright law—and it cannot—it should have been offered from the start. Moreover, all of these programs are (at best) voluntary. Other than copyright law, there is nothing stopping any of the commenters who have described such offers from changing their policies tomorrow and refusing to honor "opt out" requests.[10] Toothless commitments such as these should have no bearing on how the

television programming had not engaged in "volitional conduct" and holding provider liable for direct infringement).

[10] It is not even clear whether such offers will be rigorously and scrupulously applied. The Copyright Office, at a minimum, should consider whether and to what extent some model developers might honor requests from copyright owners whose works they do not view as important to the training or for which

Copyright Office, the courts, and Congress consider the interplay of artificial intelligence and copyright law.

These voluntary "opt-out" programs are cold comfort to copyright owners such as Getty Images whose works *already* have been used extensively without its permission. There have been no offers to cease distributing models that were trained on our works without authorization, no offers to unwind the effects of training on such works, and no explanation of how it would even be possible to do so. Absent commitments to train future models and future versions on entirely new sets of inputs that do not include "opt out" works, rather than building on already released models to develop future ones, these "opt out" offers are even less meaningful.

Robots.txt is not an effective way for copyright owners to prevent the unauthorized use of their copyrighted materials to train Generative AI Models.

A number of commenters have suggested that there is little reason to be concerned with unauthorized use of copyrighted materials in training sets for Generative AI Models because copyright owners could indicate a "do not train" preference in their website robot.txt settings. In its Initial Comments (at 19), Getty Images identified several reasons why reliance on such indications of preferences is inadequate, including that (i) it may require copyright owners to block web crawling that they do find desirable, such as (in our case) the indexing of websites by search engines to make them easier for users to find, and (ii) many copyright owners, including Getty Images, license the display of their copyrighted works on third-party websites and do not control what preferences those licensees indicate with respect to web crawling on their own websites.[11] In addition, such instructions would do nothing to address the gobsmacking extent of unauthorized copying that has already taken place.

there are easy substitutes, but then refuse to honor requests from large scale copyright owners with particularly important collections of works, such as Getty Images.

[11] Nor could Getty Images and other copyright licensors reasonably be expected to dictate website settings to their customers as a condition of licensing, especially since those settings are set at a website level, rather than applied to individual pieces of content. In any event, there is no basis to impose the incredibly disruptive burden of trying to negotiate, police, and enforce such conditions on copyright owners in order to protect the exclusive rights copyright law provides, and there is no reason to stifle the licensing that would occur but for the imposition of such a requirement in the event customers are unwilling to accept conditions of that nature.

Other comments submitted in response to the NOI betray an even more fundamental problem with relying on robots.txt instructions as a solution to the intrusion on copyright owners' exclusive rights: those instructions can be ignored. Other comments admit that some model developers are willing to violate website terms of use that expressly prohibit copying for commercial purposes and—worse—are willing to obtain copies of works from known piracy sites to facilitate training Generative AI Models because they think the fair use doctrine will still countenance their unauthorized copying in this context.[12] Because compliance with the Robots Exclusion Protocol is voluntary, a robots.txt instruction would be, at most, a minor speedbump for such unscrupulous actors, no different than the other unmistakably clear indications of copyright owner preferences that they have already willfully ignored. For example, Getty Images already (i) expressly prohibits visitors to its websites from downloading, copying, or re-transmitting any of its content and from using data mining, robots or similar data gathering or extraction methods, (ii) places conspicuous watermarks on each of its images and videos displayed on those websites in order to deter infringement, and (iii) displays conspicuous copyright notices on its websites. Is anyone naïve enough to think that Stability AI would have been deterred by a robots.txt instruction when it decided to scrape 12 million copyrighted images from Getty Images websites from its pre-determined list without permission as part of developing its Stable Diffusion Generative AI Model when it was not deterred by any of the other measures Getty Images had already taken to prevent unauthorized copying?

Developers of Generative AI Models and Generative AI Systems cannot absolve themselves of liability for unauthorized copying with facile but inapt analogies to "human learning."

A number of commenters observe that human beings learn by reading, listening to, or looking at copyrighted materials and often build on what they have learned by creating new works that do not infringe on the source materials from which they learned. In their view, the unauthorized acts of copying and distribution associated with Generative AI Models and Generative AI Systems can be excused with facile analogies to human learning. These analogies

[12] For example, as reported in the *Washington Post*, one data set (Google's C4 data set) used to train prominent large language models included materials obtained from at least 28 different web sites identified by the Office of the U.S. Trade Representative as notorious markets for counterfeiting and piracy. Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart*, WASH. POST, April 19, 2023 (available at https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/).

ignore fundamental differences between training a Generative AI Model and “human learning” that are highly consequential for copyright law.

At the risk of belaboring the obvious, Section 106 provides copyright owners with specific exclusive rights, including, for example, exclusive rights of reproduction, distribution, performance, and display.[13] Copyright law does not provide copyright owners with exclusive rights of learning, reading, looking at, or listening to. While using copyrighted works to train a Generative AI Model involves numerous acts within the scope of Section 106[14], the “human learning” to which these commenters analogize does not implicate any of the exclusive rights belonging to the copyright owner. When humans do tread on a copyright owner’s exclusive right in the course of learning or research, their otherwise infringing acts will sometimes be defensible as a fair use.[15] But they are not necessarily so. Indeed, the scientific publishing and educational publishing industries (among many others) depend on the notion that copyrighted materials cannot be indiscriminately copied and shared without permission from the copyright owner, just because the copying is intended to facilitate learning or research.[16]

Developers of Generative AI Models should be required to prepare, retain, and disclose auditable records of the materials used to train their models and the sources from which those materials were obtained.

In its Initial Comments (at 23-24), Getty Images explained why developers of Generative AI Models should be required to collect, retain, and disclose records regarding the materials used to train their models on the grounds that the training sets that they have elected to use are proprietary. Getty Images also described (at 23-24) the level of specificity that we believe should be required.

[13] 17 U.S.C. § 106.

[14] *see* pp. 3-9 *supra* and Initial Comments at 9-11

[15] 17 U.S.C. §107.

[16] *See, e.g.*, *Princeton Univ. Press v. Michigan Document Servs., Inc.*, 99 F.3d 1381 (6th Cir. 1996); *American Geophysical Union v. Texaco, Inc.*, 60 F.3d 913 (2d Cir. 1995); *Basic Books, Inc. v. Kinko’s Graphics Corp.*, 758 F. Supp. 1522 (S.D.N.Y. 1991).

Some entities that offer Generative AI Models and Generative AI Systems agree that at least some form of transparent, public disclosure of training data should be required.[17] Others, however, oppose required disclosures on the ground that training sets are trade secrets or are otherwise proprietary. The Copyright Office should reject such efforts to exalt whatever purported intellectual property interests model developers may have in the catalog of copyrighted works they have purloined to train their models over the copyright interests in the works on which models were trained. Some commenters have also noted that a disclosure required could tread on confidentiality or privacy interests. Getty Images believes that a disclosure requirement could interfere with *bona fide* privacy interests in only the most exceptional cases. Such cases would be better handled through authorization to redact truly private information (such as social security numbers, checking account numbers, personally identifiable information contained in medical records, and the like) from the required disclosures, rather than by dispensing with a disclosure requirement entirely.

***

We are thankful to the USCO for providing the opportunity to respond to the initial round of comments submitted in response to the NOI. As expressed in our Initial Comments, while we do not think it is necessary to amend existing copyright law to address issues raised by AI, we do believe that transparency standards are necessary to protect copyright interests. We are highly supportive of regulatory efforts to mandate a set of consistent standards applicable to the development and deployment of AI and believe that such standards are key to promoting responsible innovation and global harmonization. We look forward to being part of the solution to ensure that the AI ecosystem prospers, while respecting existing intellectual property rights.

Respectfully submitted,

Getty Images

[17] *See, e.g.*, Hugging Face Response to the Copyright Office Notice of Inquiry on Artificial Intelligence and Copyright at 2, 4, 12.