STATEMENT OF SY DAMLE

BEFORE THE U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON THE JUDICIARY, SUBCOMMITTEE ON COURTS, INTELLECTUAL PROPERTY, AND THE INTERNET

ON

"ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY: PART I — INTEROPERABILITY OF AI AND COPYRIGHT LAW"

MAY 17, 2023

Chairman Issa, Ranking Member Johnson, and Members of the Subcommittee, thank you for inviting me to participate in this hearing on artificial intelligence ("AI") and copyright law. I am a former software engineer, a former General Counsel and Associate Register of Copyrights of the United States Copyright Office, and current partner at the law firm of Latham & Watkins LLP. I speak solely in my personal capacity, and not on behalf of my law firm, any of the firm's clients, or the U.S. Copyright Office. In addition, my testimony is not intended in any way to constitute legal advice and should not be relied on as such.

I. Introduction and Executive Summary

My statement will focus on the copyright implications of AI training and, in particular, whether AI models should be permitted to "learn" from copyrighted works. But first I would like to put these issues into context. The AI tools of the present and near future will impact almost every aspect of the human experience. They will improve our science and medicine. They will make our military more effective. They will make our businesses more efficient and productive. They will transform the way humans learn and work. They will enable anyone to more fully unlock their creative potential. In short, AI has the potential to transform our economy and improve our society as a whole.

But that outcome is not guaranteed. The way we regulate AI will directly determine whether the United States will continue to lead the world in AI development, or whether another country will take up that mantle. Imposing heavy-handed, intellectual-property-based restrictions on AI innovation will hamper the development of AI here in the United States, and likely drive that development to other countries.

At the same time, artists, writers, and other creators have expressed genuine concern that the rapid development of AI—and, in particular so-called "generative" AI models that create text, software code, visual works, music, and other media—will displace human authors. Policymakers should take those concerns seriously. But every new technological development has led to similar fears, and in hindsight, none of those fears came to fruition. For instance, after photography was invented in the mid-1800s, critics dismissed the medium as the "refuge of every would-be painter . . . too ill-endowed or too lazy to complete his studies," and worried that if photography "is allowed to supplement art," it would "corrupt[] [art] altogether."[1] But society embraced the camera as a creative tool, and photography blossomed as an art form that deepened, rather than diminished, the field of human creativity.[2] There is no reason to believe generative AI is any different. Like the camera or the myriad creative tools adopted since, generative AI will be not a replacement for, but an engine of human creativity.

With that context in mind, I want to make two points today.

First, copyright's well-established fair use doctrine is the best way to balance the competing interests in the AI space. Fair use is a flexible doctrine in which our nation's courts explore the specific facts of every case, and determine whether allowing the particular use at issue would further—or obstruct—the Constitutional goal of copyright law, which is to promote the creation and spread of knowledge and learning.[3] Over the course of hundreds of judicial decisions, a simple principle has emerged: that the use of a copyrighted work to learn unprotectable facts and use those facts to create products that do not themselves infringe copyright is quintessential fair use. In general, the training of AI models adheres to that principle. Today's AI leaders have built their innovative products relying on that understanding. Indeed, the uniquely American fair use doctrine is, in large part, why the United States is the epicenter of global AI development.

Some have suggested that generative AI models' ability to replicate artistic styles vitiates any fair use defense. This concern has nothing to do with copyright, which does not, and has never, granted monopolies over artistic or musical styles. And the handful of historical accidents in which courts have overlooked this—including the much-reviled *Blurred Lines* copyright case in which the estate of Marvin Gaye successfully pursued claims against artists Robin Thicke and Pharrell Williams—only demonstrate the importance of this fundamental principle to the health of our creative economy. Other laws may provide more appropriate avenues to redress these concerns.

[1] Charles Baudelaire, *The Salon of 1859 II: The Modern Public and Photography, in* ART IN PARIS 1845–1862: SALONS AND OTHER EXHIBITIONS 152–54 (J. Mayne trans., Phaidon 1965); *see also* Christine Haight Farley, *The Lingering Effects of Copyright's Response to the Invention of Photography*, 65 U. PITT. L. REV. 385, 417–18 (2004) ("[F]or the majority of artists, their first reaction to the invention of photography was outwardly hostile.").

[2] There are countless other examples. Orchestra conductors in the 1930s warned Congress that performing artists had been "rendered helpless and are unable to cope with this vicious and constant repetition" made possible by the advent of recorded performances, which had impacted the demand for live orchestras—only to subsequently embrace recorded music and radio as foundations of the modern music industry. *Revision of Copyright Laws, Hearings Before the House Comm. on Patents*, 74th Cong., 2d Sess. 680 (1936) (statement of Josef Pasternak).

[3] U.S. Const. art. 1, § 8, cl. 8; *Golan v. Holder*, 565 U.S. 302, 324 (2012); *see also* 17 U.S.C. § 107.

That said, it is entirely possible that some "AI" tools may exceed the bounds of fair use. But with the benefit of over 100 years of principle and precedent, our courts are well-equipped to differentiate between fair and infringing uses.

Second, while certain groups are seeking payment for the use of content to train AI models, everyone agrees that it is impossible for AI developers to negotiate and acquire licenses from every rightsholder who owns an copyright interest in the data used to train AI models. So instead, some groups have proposed statutory or collective licensing regimes under which any use of copyright-eligible content to train an AI model would trigger a payment obligation. This would be bad policy, if adopted. Rather than permitting nuanced and case-specific assessments of this new technology, these proposals would eliminate fair use in this area, replacing it with a rigid and inflexible—and most often incorrect—assumption that AI training is infringing.

Moreover, even if it were appropriate to implement a statutory or collective licensing regime, doing it would be far from straightforward. Successfully training an AI model requires using many *billions* of pieces of content, so the scope of any statutory or collective licensing scheme would be many orders of magnitude larger than any similar scheme in the history of American law. Given that scale, any royalty providing meaningful compensation to individual creators could impose an enormous financial burden on AI companies that would either bankrupt them or push all but the largest companies out of the market (or out of the country). Worse still, nearly all the content used to train AI models—including, *e.g.*, anonymous posts on internet forums and review websites—is not only unregistered, but has no identifiable owner. That means the vast, vast majority of royalties would go "unmatched" and therefore unpaid to the original authors. It is worth carefully considering whether it is desirable to adopt a scheme that could cripple AI development—and, by extension, the country's competitive standing—while providing no benefit to the overwhelming majority of the creators whose works are being used.

In short, existing copyright law is more than up to the task of balancing the need for a dynamic domestic AI industry with the rights of creators.

II. Fair Use

The United States is the epicenter of generative AI technology. Almost all the companies responsible for developing the current and future generations of AI—and many of the researchers who developed the underlying technology—are based here. The reason in large part is that United States copyright law, almost uniquely, provides a broad and flexible fair use doctrine.

Foundational copyright cases establish that the use of copyright-eligible content to create non-infringing works is protected fair use, even if the non-infringing works compete with the originals. This principle applies directly to the mine run of today's popular AI models, which extract abstract concepts and patterns from billions of pieces of training data and use those concepts to create new content that significantly differs from, and therefore does not infringe, any individual piece of training data. And while the fair use defense does not necessarily protect all uses of copyright-eligible content to train AI models, the courts are in the best position to delineate fair from infringing uses through the well-worn process of case-specific fair use adjudication.

A. A History of Fair Use and New Technology

Fair use is an "equitable rule of reason" that "permits courts to avoid rigid application of the copyright statute when . . . it would stifle the very creativity which that law is designed to foster."[4] As one court explained, the "ultimate test of fair use . . . is whether the copyright law's goal of promoting the Progress of Science and useful Arts would be better served by allowing the use than by preventing it."[5] The doctrine's roots stretch back to Justice Joseph Story's 1841 opinion in *Folsom v. Marsh,* which laid out a broad framework that Congress ultimately refashioned and codified as Section 107 of the Copyright Act.[6] In the 180 years since Justice Story's opinion, courts have developed the doctrine by applying it to many hundreds of cases.

Over the past few decades, fair use has proved to be an extraordinarily effective and flexible tool for reconciling copyright law with new technology.[7] In 1984, for example, the Supreme Court relied on the fair use doctrine to shield the manufacturers of video tape recorders against novel claims of secondary infringement.[8] This facilitated "the growth of a vast new and unforeseen market for the movie studios in the rental and sale of videos for home viewing" that ultimately became "the largest source of revenue for the U.S. movie industry."[9] Lower courts have similarly

[4] *Stewart v. Abend*, 495 U.S. 207, 236 (1990); *see also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 575 (1994) ("From the infancy of copyright protection, some opportunity for fair use of copyrighted materials has been thought necessary to fulfill copyright's very purpose, to promote the Progress of Science and useful Arts." (cleaned up)).

[5] *Castle Rock Ent'mt, Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 141 (2d Cir. 1998).

[6] *See Folsom v. Marsh*, 9 F. Cas. 342, 348 (C.C.D Mass. 1841) (Story, J.).

[7] Indeed, adapting copyright law to new technology is a core function of the fair use doctrine. H.R. Rep. No. 94–1476 at 65–66 (1976) (courts must "adapt the doctrine [of fair use] to particular situations on a case-by-case basis" and given "rapid technological change"); *see also* Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537, 2602 (2009) ("One of the important functions of fair use is providing a balancing mechanism within copyright law to allow it to address questions posed by new technologies . . . that the legislature could not or did not contemplate."); *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 430 (1984) ("From its beginning, the law of copyright has developed in response to significant changes in technology.").

[8] *Sony*, 464 U.S. at 454–55.

[9] Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 799 (2010); *see also* Fred von Lohmann, *Fair Use As Innovation Policy*, 23 BERKLEY TECH. L. J. 829, 837 (2008) ("[T]he fair use doctrine has been an unsung hero in the tale of America's innovation economy, encouraging investment and creating new markets for technology companies.").

guided the doctrine through various technological developments, using fair use to lay the legal groundwork for revolutionary technologies like internet search,[10] while declining to extend the doctrine to more exploitative technologies, like file sharing and unlicensed media monitoring services.[11] And just recently the Supreme Court applied the doctrine to permit the reuse of software APIs, in part to prevent the copyright monopoly from acting as a "lock limiting the future creativity of new programs."[12] Fair use, in other words, has proven flexible enough to distinguish infringing from non-infringing uses throughout many eras of technological change, consistently balancing the interests of rightsholders with the public's interest in benefitting from technological innovation.

Many uses of copyright-eligible works to train AI models will fit comfortably within the boundaries of fair use. An unbroken line of cases establishes that the use of a copyrighted work to create a non-infringing final product is quintessential fair use.

The first category of relevant cases involves making complete digital copies of copyrighted works to create non-infringing search tools. *Authors Guild, Inc. v. HathiTrust* concerned an online "repository for [] digital copies" of books scanned from university libraries.[13] The repository "allow[ed] the general public to search for particular terms across all digital copies" with those searches yielding results in the form of page numbers and the frequency of the search term on each page.[14] After a group of authors sued the repository for making unauthorized copies of their books (*i.e.*, the scanned book copies), the Second Circuit held that the text-search function was "quintessentially transformative [fair] use," in part because the resulting search results bore "little or no resemblance" to the original scanned works.[15] The year after, the Second Circuit applied that same holding to the "Google Books" service, holding that Google's "creation of complete digital copies of copyrighted [books]" was fair because that copying was in service of a distinct purpose: "identifying books of interest to the searcher."[16]

[10] *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818–22 (9th Cir. 2003) (use of thumbnail images by search engine is fair use); *Field v. Google Inc.*, 412 F. Supp. 2d 1106, 1117–23 (D. Nev. 2006) (caching websites to enable internet search is fair use).

[11] *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1014–19 (9th Cir. 2001) (use of online file-sharing platform not fair use); *Fox News Network, LLC v. Tveyes, Inc.*, 883 F.3d 169, 181 (2d Cir. 2018) (service that "commercially re-distribut[ed]" televised broadcasts "without payment or license" not a fair use).

[12] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1208 (2021).

[13] 755 F.3d 87, 90 (2d Cir. 2014).

[14] *Id.* at 91.

[15] *Id.* at 97.

[16] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 217–18 (2d Cir. 2015).

The second category of relevant cases involves the creation of copies of copyrighted works in the course of technological development of competing products. *Sega Enterprises Ltd. v. Accolade, Inc.* concerned a video game company (Accolade) that "reverse engineered" Sega's game consoles to create and market video games that users could play on their Sega consoles.[17] That process involved copying Sega's copyrighted computer "code contained in commercially available copies of Sega's game cartridges" to study the "interface specifications" for Sega's console.[18] After Sega sued, the Ninth Circuit held that the copying was fair use because it was done for a "legitimate, essentially non-exploitative purpose," *i.e.*, to "study the functional requirements for [] compatibility" and create games that would legitimately compete with Sega in the marketplace.[19] The Court held that the mere fact that Accolade "copied [Sega's] code in order to produce a competing product" did not suggest that the use was not fair.[20] This holding was adopted by the Federal and Eleventh Circuits (and re-affirmed by the Ninth Circuit) in similar cases about the use of copyrighted material to create non-infringing competitive products.[21] The Seventh Circuit adopted the same logic in an opinion by Judge Richard Posner about the verbatim copying of a database.[22] And the Supreme Court recently endorsed *Sega* in support of the holding that fair use precludes the use of a copyright monopoly to impede competition.[23]

These cases—*Hathitrust, Authors Guild, Sega*, and their progeny—are foundational cases in American copyright law. They are taught in every law school in the country and are embraced by a broad consensus of judges, copyright scholars, and industry stakeholders. Indeed, the Copyright Office relied on these cases in declining to recommend further changes to other parts of the

[17] 977 F.2d 1510, 1514–15 (9th Cir. 1992).

[18] *Id.*

[19] *Id.* at 1522–23.

[20] *Id.* at 1522.

[21] *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 603–08 (9th Cir. 2000) (creation of intermediate copy of Sony PlayStation BIOS code to create a PC-based PlayStation emulator was fair use, even though defendant's product would cause Sony to "lose console sales and profits"); *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1539 n.18 (11th Cir. 1996) ("endorsing" the fair use holding in *Sega* in case about reverse-engineering of operating system to compete with copyright owner); *Atari Games Corp. v. Nintendo of Am. Inc.*, 975 F.2d 832, 836–37, 843–44 (Fed. Cir. 1992) (creation of complete copies of computer code to "unlock" a competitor's video game program was fair use).

[22] *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 644–45 (7th Cir. 2003); *see also A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638–40, 645 (4th Cir. 2009) (verbatim copying of copyrighted works to create plagiarism-detection software was fair).

[23] *Oracle*, 141 S. Ct. at 1208 ("An attempt to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression." (quoting *Sega*, 977 F.2d at 1523–24)); *see also id.* at 1198–99 (citing both *Sega* and *Connectix* with approval).

Copyright Act, which the Office found unnecessary because "intermediate copying for purposes of . . . creation of interoperable software is, in most cases, a fair use."[24]

B. Application of Fair Use Principles to AI Training

While these cases unambiguously establish the principle that use of copyright-eligible material to create a non-infringing product is fair, applying this principle to AI training will necessarily be fact specific.

Most of the generative AI models we see today are designed solely to create *new* content. The models are not designed to reproduce copyrightable expression from training data and, in nearly all circumstances, do not do so. Rather, the models derive abstract patterns and relationships—not copyrightable expression—from billions of pieces of training data, and then use those abstract (and uncopyrightable) correlations to create new, non-infringing content.[25]

For instance, in a typical natural language generative AI model (like those underlying AI chatbots), content in the training corpus is broken up into discrete segments. Then, the AI model examines and extracts statistical relationships among those pieces of content, *e.g.*, their frequency, importance, and semantic relationship to each other. That statistical data is incorporated into the algorithm, and the original content is discarded. By doing this across billions and billions of works, the AI model learns facts about the English language as a whole. To take an extremely simple example, the model may learn that the phrase "Today is a" is more likely to be followed by "sunny day" or "Tuesday" than "grey elephant" or "necktie."

The result of this process (assuming that this is a natural language model) is the creation of a complex repository of statistical facts about the relationship between words.[26] Thus, the AI model is not at all a "collage tool" that stores "compressed copies" of works, as some have alleged.[27] Rather, the model derives unprotectable information from the billions of works on which it is trained—much like the tools at issue in *HathiTrust* and *Authors Guild* derived information from

[24] U.S. Copyright Office, Report on Software-Enabled Consumer Products at 57 (2016).

[25] 17 U.S.C. § 102(b) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.").

[26] This process is sometimes called "vectorization" because the goal is to create "word vectors" that "mathematically represent the meaning of a word" so "semantically similar words have similar vectors." Jayesh Bapu Ahire, *Introduction to Word Vectors*, DZone (Jun. 21, 2022), https://dzone.com/articles/introduction-to-word-vectors. For instance "words such as wheel and engine should have similar word vectors to the word car (because of the similarity of their meanings), whereas the word banana should be quite distant." *Id.*

[27] https://stablediffusionlitigation.com/.

scanned copies of books. A language model, for example, learns statistical facts *about the English language as a whole,* while AI models in other domains might learn statistical facts about patterns in imagery, music, software code, and so forth. The models then use those statistical facts to generate new output based on an initial input (*e.g.*, a user's text query). Such a model may also learn unprotectable facts, like the height of the Washington Monument or the year the White House was built. The process, in other words, is designed to generate outputs based on unprotectable facts and abstract concepts, not to recycle or "collage" its training data. To be sure, the output of these models may reflect the same abstract concepts or ideas as works found in the training data, much like two news articles on the same topic might communicate the same unprotectable facts, or like two paintings might depict the same historical landmark. But absent some aberration in the training data or model design (as discussed below), the output will not be a "copy" of (*i.e.*, substantially similar to) any individual work on which the model has been trained.

Under the foundational precedents discussed above, this use of copyrighted works is quintessentially fair—it is a fundamentally a process to extract unprotectable facts about an entire corpus of works, and use those facts to generate original output.

In rare circumstances, quirks in the training data sets used to train AI models can cause outputs that resemble individual pieces of training data.[28] Critically, the ability of AI models to duplicate training data is a bug, not a feature. AI researchers see this "overfitting"—*i.e.*, creation of an output that hews too closely to a single piece of training data—as a problem, and are working on a range of methods for avoiding it.[29] One simple method is to *increase* the amount of training data on which the model is based.[30] (Thus, and ironically, any effort to restrict the availability of training data may have the unintended effect of increasing the frequency of AI models simply copying pre-existing expression.) In any case, under current doctrine, training a model that predominantly creates non-infringing outputs easily qualifies for fair use protection, whether or not the model *can* be used to infringe copyright in rare instances of overfitting (and whether or not those rare individual acts of reproduction would qualify as actionable infringement themselves).[31]

[28] *See, e,g.*, Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, at 4–5 (2023), https://arxiv.org/abs/2301.13188 (testing one popular image generation model and determining that out of 175 million images the researchers generated, only 109 images were duplicates or near-duplicates of the training data).

[29] *Id.* at 14; Henderson et al., *Foundation Models and Fair Use*, at 20–25 (Mar. 29, 2023), https://arxiv.org/pdf/2303.15715.pdf.

[30] Jared Kaplan et al., *Scaling Laws for Neural Language Models*, at 3 (2020), https://arxiv.org/abs/2001.08361 ("the size of the dataset" determines overall "[m]odel performance").

[31] *Sony,* 464 U.S. at 442 ("[T]he sale of copying equipment, like the sale of other articles of commerce, does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses."); *see also iParadigms*, 562

This, however, does not mean that *all* AI models will invariably be able to successfully avail themselves of the fair use doctrine. There are countless ways to design and train an AI model, and not all of them will necessarily stay within the bounds of fair use. For instance, creators of generative models trained on a small amount of data that, as a result, might more reliably replicate that data may find it harder to establish that their uses are fair. But the key point is that our courts are best equipped to make this determination, by marshalling and evaluating case-specific evidence, as they have in many hundreds of fair use cases over the past 180 years.

C. AI's Use of Artistic Styles is Not a Copyright Issue

Several critics of the AI industry have focused on the fact that some generative AI models are able to derive an "artistic style" from training data and replicate that style, on demand, in outputs. I can understand that concern—but this is not a copyright problem. Copyright does not, and has never, permitted an artist to preclude another from mimicking an "artistic style."[32] For that reason, generative AI's ability to replicate individual artistic styles should play no role in the fair use analysis, nor should it motivate Congress to alter the copyright framework.

Courts have repeatedly explained the importance of "styles" remaining free for all to use. In *Williams v. 3DExport*, for example, the court considered copyright claims brought by an artist who "contend[ed] that he invented anime, one of the world's most popular styles of animation."[33] The artist traced the style to a "master dissertation" he wrote in which he described an "art style" featuring "characters with 'round features,' 'big eyes,' and a 'particular nature of [] hair,'" and pointed to a graphic novel he wrote (and registered with the Copyright Office) as an example of that style.[34] In the suit, the artist objected that his style had been "illegally [taken]" and used in

F.3d at 639–40 (explaining "[t]he question of whether a use is transformative does not rise or fall on whether the use perfectly achieves its intended purpose," finding use fair even though system "is not fool-proof").

[32] *Jewelry 10, Inc. v. Elegance Trading Co.*, No. 88-cv-1320, 1991 WL 144151, at *4 (S.D.N.Y. July 20, 1991) ("[A] painter who develops a style or technique, such as the rendition of perspective, impressionism, pointillism, fauve coloring, cubism, abstraction, psychedelic colors, minimalism, etc., cannot prevent others from adopting those ideas in their work."); *see also Douglas v. Osteen*, 317 F. App'x 97, 99 (3d Cir. 2009) ("Furthermore, the use of a particular writing style or literary method is not protected by the Copyright Act."); *Whitehead v. CBS/Viacom, Inc.*, 315 F. Supp. 2d 1, 11 (D.D.C. 2004) ("[S]tyle alone cannot support a copyright claim. While similar writing styles may contribute to similarity between works' total concept and feel, a particular writing style or method of expression standing alone is not protected by the Copyright Act."); *see also* 2 PATRY ON COPYRIGHT § 4:14 ("Particular elements may colloquially be regarded as typical of an author or artist's individual style, but it is only their fixation in a particular work in a particular expression that is eligible for [copyright] protection.").

[33] No. 19-cv-12240, 2020 WL 532418, at *1 (E.D. Mich. Feb. 3, 2020).

[34] *Id.*

new artwork that he found “disturbing,” and sued thirteen defendants.[35] The court did not hesitate to dismiss the claim, holding:

> . . . even if [plaintiff] was the first to think up the anime, he could only have a protectable copyright interest in his specific expression of that idea; he could not lay claim to all anime that ever was or will be produced.[36]

Many other cases reach the same result.[37] And for good reason: extending copyright protection to styles not only violates the express command of Section 102(b) of the Copyright Act, but (as commentators have explained) would lead to an unworkable copyright system:

> If an author or artist claimed broad protection for a style not associated with a particular work and fixation, it would be difficult, if not impossible, to determine the scope of protection. . . . An endless series of derivative works could be created, yet none might be treated as a derivative work since it is the style and not any particular work for which protection would be asserted. Determining substantial similarity between plaintiff's and defendant's “works” would be skewed since plaintiff would not be asserting copyright in a work, but rather in an amorphous style that exists independent of any particular work.[38]

The much-reviled *Blurred Lines* copyright verdict is a perfect example of the chaos that results from extension of copyright to protect styles. In 2015, a jury found that Robin Thicke’s song “Blurred Lines” infringed Marvin Gaye’s 1977 hit song “Got To Give It Up” based on alleged similarities between the two works.[39] That result has been widely criticized by a broad array of industry stakeholders, creatives, and commentators, who have pointed out that “the similarities between the songs are not within the melody, lyrics, or harmony, but rather in the overall sound, groove, and vibe,”[40] and argued that the verdict violates “the principle that ideas cannot be copyrighted, a notion that is essential to free speech and artistic expression.”[41] It has thrown the

[35] *Id.*

[36] *Id.* at *3.

[37] *See* 2 PATRY ON COPYRIGHT § 4:14 (collecting cases).

[38] *Id.*

[39] *See generally Williams v. Gaye*, 895 F.3d 1106 (9th Cir. 2018) (upholding jury verdict on procedural grounds).

[40] Olivia Lattanza, *The Blurred Protection for the Feel or Groove of a Song Under Copyright Law: Examining the Implications of* Williams v. Gaye *on Creativity in Music*, 35 TOURO L. REV. 723, 725 (2019).

[41] Tim Wu, *Why the “Blurred Lines” Copyright Verdict Should Be Thrown Out*, New Yorker (Mar. 12, 2015), https://www.newyorker.com/culture/culture-desk/why-the-blurred-lines-copyright-verdict-should-be-thrown-out; *see also* The Editors of GQ, *Rick Rubin on Pharrell’s “Blurred Lines” Lawsuit*, GQ (Nov. 4, 2019), https://www.gq.com/story/pharrell-and-rick-rubin-blurred-lines-copyright-lawsuit (quoting famed producer Rick Rubin as saying the verdict is “bad for music” because “now, based on that one case,

music industry into mild disarray,[42] and spurred many follow-on lawsuits against, for example, Led Zeppelin and Taylor Swift.[43] Thankfully, many of these follow-on lawsuits have failed,[44] including one recently brought by the estate of a co-writer of a Marvin Gaye song against the artist Ed Sheeran, suggesting that history will view the "Blurred Lines" case as a misguided historical accident.

Notably, prominent rightsholders were among the chorus of voices who denounced the *Blurred Lines* decisions and endorsed the bedrock principle that styles are not protected by copyright. As the Recording Industry Association of America and National Music Publishers' Association warned the Ninth Circuit in an amicus brief urging vacatur of the verdict, "[n]ew generations of musicians would be in constant peril of copyright lawsuits because they had used someone's musical style" if the decision's infringement-by-style rationale were upheld.[45]

Style appropriation, in other words, is not a copyright concern, and it is critical that styles remain freely accessible for public use. As one of this nation's leading jurists has explained, "Intellectual (and artistic) progress is possible only if each author builds on the work of others. No one invents even a tiny fraction of the ideas that make up our cultural heritage . . . Every work uses scraps of thought from thousands of predecessors, far too many to compensate even if the legal system were frictionless, which it isn't."[46] Those who suggest that AI models' ability to replicate individual artistic styles should somehow affect the fair use analysis, or motivate Congress to alter the copyright framework, ignore this fundamental principle. To be sure, generative AI can be used in ways that might implicate other legal frameworks, like unfair competition, trademark, and state rights of publicity and privacy. But the general principle that styles are free for all to use is a

there's a question of what a song is . . . [a] feeling is not something you can copyright").

[42] Jason Palmer, *"Blurred Lines" Means Changing Focus*, 18 VAND. J. OF ENT. & TECH. L. 907, 926–28 (2016) (summarizing the "industry problems" that arose after the *Blurred Lines* verdict, including the payment of "precautionary royalties").

[43] Ben Sisario, *Ed Sheeran Wins Copyright Case Over Marvin Gaye's "Let's Get It On,"* N.Y. Times (May 4, 2023), https://www.nytimes.com/2023/05/04/arts/music/ed-sheeran-marvin-gaye-copyright-trial-verdict.html (commenting on the connection between the *Blurred Lines* case and subsequent similar cases).

[44] *See, e.g.*, *id.*

[45] Brief Amici Curiae of the RIAA and NMPA in *Williams, et al. v. Gaye*, No. 15-56880 (Dkt. 100-2 at 28) (9th Cir., filed Apr. 2018); *see also* Brief Amici Curiae of the RIAA and NMPA in *Skidmore v. Led Zeppelin*, No. 16-56057 (Dkt. 77-2 at 8) (9th Cir., filed Nov. 5, 2018) (urging *en banc* reconsideration of panel opinion vacating jury verdict for defendant, arguing that composers "need copyright law to let them create new music incorporating ideas from the vast cultural library of past musical works").

[46] *Nash v. CBS, Inc.*, 899 F.2d 1537, 1540 (7th Cir. 1990) (Easterbrook, J.).

foundational aspect of federal copyright policy that is essential to our creative economy and the general freedoms of speech and expression.[47]

III. The Feasibility of Licensing AI Training Data

All stakeholders—even the most vocal critics of today's generative AI tools—appear to agree that the technology is "capable of revolutionizing the creative process" and "enhanc[ing] artistic expression."[48] Everyone, in other words, has an interest in making sure that this technology succeeds and does so in a way that enhances, rather than undercuts, our ability to compete on the world stage.

But it is also well understood that mandating bilateral negotiations between AI developers and individual rightsowners will be counterproductive. It will be impossible for legitimate AI developers to negotiate with each and every rightsholder who owns a copyright interest in one of the billions of individual pieces of data the developers' models require. And attempts to build AI models using smaller sets of licensed or public domain material will lead to models that are less effective and, ironically, more likely to (inadvertently) create outputs that simply regurgitate their training data. Moreover, the economic dynamics of any bilateral licensing negotiation will be impossibly skewed: the supply of potential training data is effectively unlimited, meaning that no individual rightsholder will be able to demand more than nominal compensation for the use of its works.

Recognizing these problems, certain rightsholders and industry groups have instead suggested a statutory or collective licensing regime.[49] I believe these proposals are misguided. Imposing a payment requirement for AI training data would cut against the deeply rooted copyright principles discussed above—it would require payment for a process the result of which is the extraction of unprotectable facts about an entire corpus of potentially copyrighted works. And besides being bad policy, such a scheme could raise serious constitutional concerns.[50] And if the concern is that

[47] *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003) (copyright's denial of protection to "ideas" is a "First Amendment accommodation" "built-in" to the Copyright Act).

[48] Press Release, Authors Guild, Creator Groups Meet Lawmakers on AI Issues (Apr. 4, 2023), https://authorsguild.org/news/ag-and-creator-groups-meet-lawmakers-on-ai-issues/.

[49] *See, e.g.*, *id.*

[50] The Supreme Court has stated that the existence of the fair use defense is critical to ensuring that the copyright law does not conflict with the First Amendment. *Eldred*, 537 U.S. at 218–21 (suggesting that any attempt by Congress to "alter[] the traditional contours of copyright protection" could trigger "First Amendment scrutiny," and upholding challenged law against First Amendment challenge due in part to presence of fair use defense). Because a statutory licensing scheme would require payment for uses of copyright-eligible material that the fair use defense would otherwise permit without payment or permission, such a scheme may be subject to a constitutional challenge.

some AI models will not stay within the bounds of established fair use principles, the judiciary's careful and time-tested fair use jurisprudence permits deeper, case-specific consideration, and compensation to creators in appropriate cases.[51] That is far preferable to an across-the-board legislative determination that all AI training is infringing.

Even setting those policy objections aside, any effort to implement a statutory or collective licensing regime would need to grapple with a series of (perhaps insurmountable) practical problems.

The first and most obvious problem with a statutory licensing regime for AI training is scale. AI models can train on any media including, for example, long-form text (*e.g.*, blog posts), short-form text (*e.g.*, Tweets or forum comments), or images (full size or thumbnail). Often, AI developers train their models by pulling this content from the internet.[52] Because virtually all of this content is eligible for copyright protection—regardless of its perceived aesthetic or expressive value[53]—any statutory licensing scheme would need to include virtually the entire internet within its scope. As a result, any licensing regime that triggered a payment obligation to any rightsholder whose work is used to train an AI model would require the administration of that license (and the payment of royalties) for *billions upon billions* of works—many of which are published online with no ownership information.

As precedent, some have pointed to the blanket licensing regime for musical works established by 2018 Music Modernization Act ("MMA").[54] At a high level, that regime created a new mechanism by which "digital music providers" could obtain a "blanket mechanical license" through a designated entity known as the "mechanical licensing collective" or "MLC."[55] The MLC administers the blanket license on behalf of musical work copyright owners by, for example*,* collecting royalties from digital music providers, seeking to identify the owners of the musical works in each song for which royalties are paid (a process known as "matching"), and distributing collected royalties to the appropriate owners.[56] Congress required the MLC to hold any

[51] A related concern is that writing hard-coded legislation in the context of evolving technology is always extremely difficult and carries a substantial risk that subsequent technological developments will render the statutory framework obsolete. *See, e.g.,* Aaron L. Melville, *Note, The Future of the Audio Home Recording Act of 1992: Has It Survived The Millennium Bug?*, 7 B.U. J. SCI. & TECH. L. 372, 383–84 (2001) (addressing the Audio Home Recording Act).

[52] One source for AI training data, for instance, is the repository of web crawl data maintained by Common Crawl—essentially, a copy of the entire world wide web. *See* https://commoncrawl.org/.

[53] *Bleistein v. Donaldson Lithographing Co.*, 188 U.S. 239, 251–52 (1903).

[54] *See* Pub. L. 115–264 at Sec. 102(a), 132 Stat. 3677, *codified at* 17 U.S.C. § 115(d).

[55] 17 U.S.C. § 115(d)(1)(A).

[56] *Id.* § 115(d)(3)(C).

"unmatched" royalties for a certain period of time, and then distribute those funds to "copyright owners identified in the records of the collective" based on their "relative market share."[57]

A statutory licensing regime for AI training would be massively more complicated than the one created by the MMA. Most obviously, the scope of a statutory license covering AI training activities would be orders of magnitude larger than any statutory license ever implemented. The MMA's blanket license, for instance, covers a universe of musical works numbering in the tens of millions.[58] By contrast, because AI models use a vast array of media for training, a statutory licensing scheme for AI training would need to cover, at least, every publicly accessible work on the internet—including every forum comment, online review, social media post, and business website. That universe of works likely numbers in the billions or tens of billions.

In addition, the works covered by this statutory license differ substantially from the works covered by the MMA's blanket license. While it is not easy to identify the musical work owner for each sound recording, the endeavor is relatively straightforward if the MLC has access to the relevant data sources, which are in large part available to the MLC's matching team.[59] Acquiring sufficient data to identify the copyright owner of the billions of images or pieces of text on the internet would be much harder, if not impossible. Many photographs, social media posts, business websites, blog posts, or online reviews on the internet are published anonymously or pseudonymously, or simply with no information to identify the content's author. I cannot estimate the breadth of the issue, but it seems likely that, for nearly all of the content that AI models could use for training purposes, identifying the rightful copyright author or owner will be an unattainable goal.

The U.S. Copyright Office has touched on this issue in its multiple studies of the issue of so-called "orphan works." As the Office explains, "it can be time-consuming, difficult or even impossible to locate the copyright owner" for a given work.[60] That is particularly true when the work is

[57] *Id.* § 115(d)(3)(H), (J).

[58] As of the end of 2021, the MLC's musical works database contained data for 23.8 million musical works. Mechanical Licensing Collective, 2021 Annual Report, Appendix at 4.

[59] Mechanical Licensing Collective, 2021 Annual Report, Appendix at 20 ("The MLC's Matching Team utilizes over 30 public databases and other research sources to support its matching efforts. These sources include various collective rights management organization databases, music credit databases, [digital music provider] websites, content owner websites, and other niche and genre specific sources.").

[60] *See* U.S. Copyright Office, Report on Copyright and Digital Distance Education at 41–43 (1999); *see also* U.S. Copyright Office, Report on Orphan Works at 92 (2006) ("portion of works for which owners [cannot be] located can be significant"); *id.* at 22 (citing a study by Carnegie Mellon University Libraries finding that "for the books in the study, 22% of the publishers could not be found"); U.S. Copyright Office, Report on Orphan Works and Mass Digitization at 2 (2015) (problem of orphan works is "widespread and significant").

published without "information about the author or the owner of copyright in the work."[61] But even if the work is published with such information, subsequent "[c]hanges of [o]wnership"—either due to transfers of ownership or the copyright owner's death, relocation, or corporate dissolution—can render identification of the copyright owner impossible.[62] These problems are compounded by the lack of reliable "information resources" about ownership; as a result, many ownership searches ultimately reach "dead ends."[63]

These problems would be even more severe in the context of works that users post to the internet, often anonymously or pseudonymously. It is almost impossible to imagine how the administrator of a statutory license for AI training could identify the rightful owner of more than a tiny fraction of the works used to train AI models. Any attempt to tackle this problem would require armies of human researchers who (like the members of the MLC's matching team) would have to conduct individualized, work-by-work investigations into a massive and constantly-growing corpus of works covered by the licensing scheme.

As a result, any statutory licensing scheme would lead to massive amounts of "unmatched" royalties that would sit idle in the coffers of the license's administrator. The MLC, which administers a far narrower statutory license covering a universe of works whose owners are far easier to identify, had over $150 million in unmatched royalties in 2021 alone, and there is no indication that any amount of money or resources could eliminate that problem.[64] The amount of "unmatched" royalties under a statutory license for AI training data would no doubt be far larger. Nearly all rightsholders would receive no remuneration whatsoever. It is worth evaluating whether imposing a significant financial burden on the AI industry for such uncertain and attenuated benefits to creators is a desirable result.

A second, related challenge is that any statutory or collective licensing scheme—no matter how carefully designed—would find itself caught between two difficult-to-reconcile policy objectives: (1) to provide meaningful compensation to individual artists and rightsholders, and (2) to ensure that AI companies can continue to thrive in the United States.

A statutory or collective licensing scheme would presumably require any AI developer to pay some fixed or floating rate to compensate the copyright owners for the use of each piece of training data. And because the goal of the proposed collective licensing scheme would be to protect individual "human creators and artists,"[65] the rate paid for the use of any individual works would have to be

[61] U.S. Copyright Office, Report on Orphan Works at 23–24.

[62] *Id.* at 26–29.

[63] *Id*. at 29–34.

[64] Mechanical Licensing Collective, Annual Report 2021, Appendix at 15.

[65] Press Release, Authors Guild, *Creator Groups Meet Lawmakers on AI Issues* (Apr. 4, 2023),

financially significant. A licensing scheme that led to individual creators receiving monthly royalty checks of a few cents from the AI developers who used their works for training purposes would do nothing to protect "human creators" or the "[f]uture of journalism, literature, and the arts."[66]

But AI models require many *billions* of pieces of training data to be effective. As a result, it will be extremely challenging to set a royalty rate that provides meaningful compensation to individual copyright owners without imposing a crippling financial burden on AI developers, who would have to make many billions of rate payments for the works they use. If the royalty rate were set too high, it would either bankrupt the United States AI industry—eliminating our ability to compete on the international stage—or push all but the largest companies out of the market (or out of the country). It would, in other words, be extraordinarily challenging to set a royalty rate that would both compensate individual creators and encourage the growth and development of a domestic AI industry. Developers who are unable to afford the cost of AI development in the United States would surely move their efforts to other countries with more permissible copyright frameworks.[67]

*   *   *   *   *

In closing, I applaud the Subcommittee for its quick engagement on the important and challenging issues surrounding the rapid growth of artificial intelligence technologies. The Subcommittee, and Congress more broadly, has a chance to ensure that the United States continues to lead the world in responsible development of artificial intelligence technologies. While Congress should continue to think carefully about the copyright implications of those technologies, it should also feel confident that the laws that it already passed are well-suited to address whatever issues may arise. Thank you for the opportunity to provide testimony for this hearing, and I look forward to answering your questions.

https://authorsguild.org/news/ag-and-creator-groups-meet-lawmakers-on-ai-issues/.

[66] *Id.*

[67] *See, e.g.*, Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 BERKELEY TECH. L. J. 1407, 1431 (2021) (noting that Japan "was the first country in the world to update its copyright laws" to "demonstrate a national commitment to the flourishing of AI industries," and explaining the country's 2018 Copyright Act implementing a broad exemption for incidental copies and machine leaning); Mark A Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 769 n.147 (2021) (reporting that Singapore is considering an AI exemption much like Japan's).