VIA Electronic Submission by Kate Tummarello on behalf of Engine

October 30, 2023

Shira Perlmutter
Register of Copyrights
U.S. Copyright Office
101 Independence Ave. S.E.
Washington, D.C. 20559-6000

Re: Comments of Engine to the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, Docket No. 2023-6

Director Perlmutter:

Engine is a non-profit technology policy, research, and advocacy organization that bridges the gap between policymakers and startups. Engine works with government and a community of thousands of high-technology, growth-oriented startups across the nation to support a policy environment conducive to technology entrepreneurship. A large part of the startup ecosystem is currently developing, using, or moving towards using artificial intelligence (AI) in their products and services in diverse and innovative ways that benefit their customers and users. Current frameworks around data access and copyright make that innovation possible, and the small, competitive technology companies that make up the startup ecosystem should be front of mind when considering legal and regulatory change that would make it more difficult for developers to build and train AI and integrate it into their products and services. We accordingly appreciate the opportunity to provide comment in response to the questions posed by the Copyright Office in the Notice of Inquiry and Request for Comment.

I. The current application of copyright law is enabling innovation and competition in the startup ecosystem, to the benefit of startups and their users. (Addresses question 1)

Much of the conversation around AI has focused on a handful of specific generative AI services being offered by the largest technology companies or entities working with large technology companies. Fears—and headlines—abound about generative AI systems producing creative works that could theoretically displace human creativity and the humans who make their living in creative industries. But the AI ecosystem is much bigger and more diverse than the companies and the use cases described by those headlines, and considerations of how copyright law applies to both training data used as inputs for AI models and potentially infringing generative AI outputs should not ignore those far-reaching implications.

Startups are already using AI and, more specifically, generative AI, to provide innovative products and services to users, sometimes in subtle and innocuous ways. That includes improving water quality,[1] reducing energy waste,[2] improving equitable access to financial systems,[3] creating better health outcomes,[4] and more. In many cases, startups are using AI behind the scenes to efficiently intake inputs and optimize outcomes for users, but those AI models still need to be trained on immense amounts of training data. Even for those startups that are specifically in the generative AI space, they are often using AI to predict and generate outputs in specific circumstances or for narrower purposes. Take, for instance, ConciergeBot, an AI-powered chatbot for hospitality providers.[5] That AI system is generating outputs to power, e.g., a chatbot answering questions from vacation rental tenants about check-out procedures. There should be little opportunity for the AI to generate an output that would have any real impact on creative industries, but ConciergeBot will have to contend with the regulatory and legal frameworks written with larger, higher-profile companies in mind.

The current copyright framework allows startups on bootstrap budgets to build and scale these AI-powered tools. A different understanding or application of copyright law—for instance, to limit the use of training data sets that include copyrighted work unless the developer can obtain a license from the copyright holder or use only data the developer created herself—would dramatically limit the kinds of companies that can participate and innovate in the AI ecosystem. As we explained in comments to the U.S. Patent and Trademark Office in 2020:[6]

> Big technology companies have many users and troves of in-house data they can use to develop new AI systems. Because they own the necessary content, they would not have to worry about infringement.[7] These large companies also have significant bargaining and purchasing power for acquiring large volumes of content. Startups, on the other hand, who often must look externally for data sources, have to pull-in data from content that might be subject to copyright claims. They need to be able to do this without fear of infringement accusations.

[1] See, e.g., Varuna, https://varuna.city/.
[2] See, e.g., COI Energy, https://www.coienergy.com/.
[3] #StartupsEverywhere Profile: Kenneth Salas, Co-Founder & COO, Camino Financial, Engine (May 20, 2022), https://www.engine.is/news/startupseverywhere-losangeles-ca-caminofinancial.
[4] #StartupsEverywhere Profile: Noelle Acosta, Founder & CEO, Noula Health, Engine (Oct. 28, 2022), https://www.engine.is/news/startupseverywhere-newyork-ny-noulahealth.
[5] See, e.g., #StartupsEverywhere profile: James Silva, Founder & CEO, ConciergeBot (Aug. 6, 2021), https://www.engine.is/news/startupseverywhere-sanfrancisco-ca-conciergebot.
[6] *See Comments of Engine Advocacy in Response to Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation*, Engine (Jan. 10, 2020), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/5e1c9b95bf2cc11b00b9e944/1578933141931/2020.01.10_Comments+to+Docket+PTO+C+2019+0038.pdf.
[7] See, e.g., Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power*, N.Y. TIMES, Sept. 26, 2019.

And, as always, any increase in regulatory and compliance costs will fall disproportionately on startups. As we described earlier this year to the Office of Science and Technology Policy:[8]

> Large and incumbent firms have teams dedicated to navigating the regulatory environment and a larger base of revenue over which to absorb compliance costs. Early-stage startups have small teams that often fill multiple roles at once and few resources to direct away from core startup activities like product development or customer acquisition. If regulatory costs force startups to raise their prices, potential customers will turn to larger companies instead.[9] Unnecessarily high regulatory costs would therefore put AI development out of reach for many startups, decreasing competition and cementing existing market dynamics. These costs might also discourage startups looking to integrate third-party AI into their products and services from doing so, as acquiring these systems will become more expensive, thus harming valuable innovation by preventing startups from improving their offerings.

II. It would not be practical—or feasible—for startups to acquire licenses for any potentially copyrighted material included in vast and diverse training sets, or risk being sued for copyright infringement. (Addresses questions 6, 6.3, 7.2, 8.4, and 9.3)

Outside of the reasons that requiring licenses for the inclusion of copyrighted material in training data would be a misapplication of existing copyright law, as described below, licensing requirements for AI developers to use copyrighted material in training data would be impractical and would dramatically chill innovation and the ability of startups, specifically, to compete in the AI ecosystem.

AI models need significant amounts of training data to draw inferences, create accurate predictions, and produce useful outputs. The source of that data varies dramatically across the startup ecosystem depending on several factors, including the purpose of the AI model. At the same time, all kinds of material can be eligible for copyright protection, and startups should not be responsible for parsing the copyright status of any given material. As we explained in our 2020 comments to the U.S. Patent and Trademark Office:[10]

> Content that is potentially eligible for copyright protection is ubiquitous and varied. Copyright protection can apply to any 'original works of authorship fixed in [a] tangible

[8] *See Comments of Engine Advocacy and the Center for American Entrepreneurship in response to Request for Information on National Priorities for Artificial Intelligence*, Engine and the Center for American Entrepreneurship (July 7, 2023), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/64ad7997962fb07dbb8be067/1689090455946/Engine+CAE+OSTP+AI+Comments.pdf.

[9] *See Hearing on Opportunities and Challenges for Trade Policy in the Digital Economy Before the Subcomm. on Int'l Trade, Customs, and Glob. Competitiveness* (statement of PILOT Inc.) https://engine.is/s/Statement-for-the-record-Ben-Brooks-PILOT.pdf, ("Cumbersome regulatory environments also impact our prospective customers, who respond by reducing the amount of vendors they have. That means they often consolidate their supplier base to work with a few large companies and startups like us lose out on critical business opportunities."); *The State of the Startup Ecosystem*, Engine (April 2021) https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/60819983b7f8be1a2a99972d/1619106194054/The+State+of+the+Startup+Ecosystem.pdf.

[10] Engine, *supra* note 6, at 2, 8.

> medium of expression.'[11] While underlying facts, data, ideas, etc., are not eligible for copyright protection, the broader works that contain underlying facts might have expressive elements that render such works copyrightable.[12] As such, the data used to train, test, and tune AI systems may be taken from content that is potentially eligible for copyright protection. …Everyone who generates any content arguably has some claim to copyright protection (and such content can range from highly expressive paintings to largely-factual academic papers that are at least expressive in part). It would be untenable to require innovators developing new AI technology to assess the copyright status of every piece of content and every data source feeding into an AI system. If developers then had to obtain licenses to any content or data sources, it would magnify the problem enormously.[13] If developers faced those sorts of burdens when compiling datasets, it would slow (and could stall) progress.

Startups have incredibly limited time and resources that would be stretched far beyond capacity if they had to find and negotiate with rightsholders every time a piece of copyrighted material was included in training data and/or defend themselves against copyright infringement claims in court. According to Engine research, the average seed-stage startup—already a successful startup that has received outside funding—has about $55,000 per month to cover all of its expenses, including salaries, equipment, research and development, and customer acquisition.[14] Considering the wide range of data that can be included in training data sets for AI models and the amount of inputs necessary for AI outputs to be useful, it's difficult, if not impossible, to estimate the total cost for a startup. An analysis of hypothetical licensing models for Google Books provides a helpful starting point for evaluating the incredible costs—both in terms of time and money—inherent in licensing large data sets:[15]

> For each book Google [would] have to (1) determine whether the book is in the public domain, (2) determine the identity of the copyright owner(s), (3) locate the copyright owner(s), and (4) negotiate to obtain the permission of the owner(s). … [After removing non-unique books and those published before 1923, there are] about 8.4 million books with some potential copyright constraint. Even if the average clearance cost (the cost of determining the status of the book, finding the relevant copyright owners and negotiating a license) were as little as $200, the total cost of rights clearance before any royalties have been paid would be over a billion dollars. It is easy to imagine that clearance costs could be in the thousands, not merely the hundreds, in which case the total cost of proactively clearing rights on every book could exceed $10 billion. This does not include any royalties paid to authors.

---

[11] 17 U.S.C. § 102(a).
[12] *See* 17 U.S.C. § 102(b).
[13] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. at 1659-61(2009) (describing high transaction costs that would be encountered if Internet search engines had to obtain copyright clearance for all searchable content).
[14] Engine, *supra* note 9, at 17.
[15] Sag, *supra* note 12, at 48-50.

Litigation costs are equally potentially ruinous, especially for startups. Generally, litigation through discovery and judgment can cost hundreds of thousands of dollars in legal fees alone.[16] Copyright infringement lawsuits add on the possibility of incredibly high statutory damages per work infringed. Initial litigation costs plus the potential for statutory damages would disincentivize any but those with the deepest pockets from innovating in the AI space. As we explained:[17]

> The alternative to licensing content is using it without permission and risking infringement litigation. The cost of a single copyright infringement suit, where statutory damages of $150,000 are available as a matter of course, could be ruinous for a startup.[18] But '[b]ecause machine learning datasets can contain hundreds of thousands or millions of works, an award of statutory damages could cripple even a powerful company.'[19] These costs and risks would scare many innovators, companies, and investors away from developing new AI systems.[20]

And even if a startup was willing to take on that initial cost and risk and did prevail with an infringement defense—including if a court found that the developer's use of copyrighted material was protected by fair use—those costs are incredibly difficult to recoup.[21] Taken together, opening up AI developers to licensing requirements and infringement claims in the course of AI training and ingestion would force AI startups to invest significant time and money in determining the copyright status of all data in their training data sets and would create the specter of ruinous and repeated litigation, even if they are likely to win.

III. Considering the use of data to train AI models as an infringing use under copyright law would chill innovation and competition and, at best, require startups to mount a costly fair use defense. (Addresses questions 8, 9, 9.1, 9,4)

For the sake of keeping the AI ecosystem innovative, competitive, and accessible to startups, it is most efficient to determine that the ingestion of copyrighted content as part of a training data set is a lawful, noninfringing use under copyright law, stopping any inquiries into infringement before the question of fair use arises. As we've previously described, an AI model that pulls inferences from training data is not necessarily engaging with the expressive content of copyrighted material.[22]

> [I]f the data is just that—data—and not anything expressive, the entire copyright question is moot because there is no copyrighted material that could even be infringed. For example, a

[16] *Startups, Content Moderation, & Section 230*, Engine (December 2021), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/61b26e51cdb21375a31d312f/1639083602320/Startups%2C+Content+Moderation%2C+and+Section+230+2021.pdf.
[17] Engine, *supra* note 6, at 8.
[18] 17 U.S.C. § 504(c)
[19] Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 80 (2017); see also Daryl Lim, *AI & IP: Innovation & Creativity in an Age of Accelerated Change*, 52 AKRON L. REV. 847-48 (2018).
[20] *Id.* 80-81
[21] *Melville B. Nimmer & David Nimmer, Nimmer On Copyright* §12.02 (2012) at §14.10[D][2][b] ("[M]ost courts deny fees to prevailing defendants when the plaintiffs' claims were not motivated by bad faith.")
[22] Engine, *supra* note 6, at 3.

> facial recognition system may rely on a dataset of tightly cropped images of faces extracted from photographs. If the expressive contents/portions of the photographs are removed when the dataset is created, the data may not even be eligible for copyright protection.[23]

Additionally, the way that AI models interact with data including copyrighted material—"engaging with content like humans do when they read, listen, or view"[24]—falls outside of infringement. Even when, during ingestion, an AI model makes a copy of content within the training data set for the purposes of analyzing it, those copies "are so transitory in nature that they do not even constitute creating a copy as defined in the statute.[25]

If changes in law or legal understanding determine that the ingestion of training data by AI models constitutes a use under copyright law, that use must be considered fair use. As we noted in comments to the U.S. Patent and Trademark Office:[26]

> Numerous courts have applied the fair use factors to similar technology, and have consistently found those to be fair uses. The application of each fair use factor will vary, depending on what problem the AI system is designed to solve, what content it uses, and how that content is gathered and prepared. But, as described below, themes readily emerge and the outcome is consistent.
>
> Factor (1): *Purpose and character of the use.* With this first factor, courts must determine if use of content "merely supersede[s] the object of the originals or instead add[s] a further purpose or different character."[27] … AI's use of the content adds a further purpose or different character. While "[t]ransformative use is most obvious when [a] work is itself transformed[,] in many cases courts have held that the mere recontextualization of a copyrighted work from one expressive context to another is sufficient to sustain a finding of fair use."[28] For example, making a digital copy of a book so that it is easier for people to search within books is transformative.[29] Making thumbnail copies of images to help index and improve access to content on the Internet is also transformative.[30] And archiving student-written term papers within a database for an online plagiarism detection technology is transformative.[31] AI likewise recharacterizes content. …[A] commercial motivation cannot outweigh an otherwise transformative use,[32] so AI developed with a commercial application in mind can still be a fair use. … AI systems do not use the content in datasets to directly profit, and are not

[23] *See*, e.g., Lim, *supra* note 19 (noting that facial recognition databases compiled from news images may not even invoke fair use if the portions of the photos taken is minimal); Sobel, *supra* note 19, at 67-68.
[24] Engine, *supra* note 6, at 4.
[25] Sobel, supra note 19, at 62-63 (citing cases); see also 17 U.S.C. § 101 (defining "copies").
[26] Engine, *supra* note 6, at 5-7.
[27] Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003).
[28] Sag, *supra* note 13, at 1646.
[29] Authors Guild v. Google, Inc., 804 F.3d 202, 216-17 (2d Cir. 2015)
[30] Kelly, 336 F.3d at 818.
[31] A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640 (4th Cir. 2009).
[32] Authors Guild, 804 F.3d at 219.

making a profit off of those individual pieces of content. Instead, each piece of content in an AI dataset is among thousands (or many more) elements being used in a commercial endeavor[33]….

Factor (2): *Nature of the copyrighted work*. The nature of the content used will vary for each AI system. However, this factor rarely plays a significant role—standing alone—in determining fair use.[34] Especially because the use of content to train, tune, and test AI systems is so transformative, even if the content is highly creative, this factor should not tip the scales against a fair use finding.[35]

Factor (3): *Amount and substantiality of portion used*. Here again, the amount of content each AI system uses will vary, but even if an AI system uses entire copyrighted works during the training, tuning, or testing processes, it can still qualify for fair use.[36] What matters is 'the amount and substantiality of what is [] made accessible to a public for which it may serve as a competing substitute.'[37] And in the context of AI ingesting or processing content, the answer is none…

Factor (4): *Effect upon the potential market.* … [A]n AI system's use of a piece of content as part of the training, testing, or tuning process would not harm the creator's ability to sell or license the original content. An AI system does not sell, license, or even make publicly available the underlying original content.

Even if protected by fair use, startups would face significant uncertainty around the potential inclusion of copyrighted material in training data sets as it is fact-specific, and any one finding of fair use in court would not apply blanketly across the entire AI ecosystem.[38] As discussed above, all litigation—but especially copyright litigation, which carries the risk of high statutory damages—is incredibly expensive for startups with limited resources. In a world where the inclusion of copyrighted material in training data requires the ability to mount a fair use defense, policymakers would need to work swiftly and efficiently to create certainty around that protected fair use of training data, including through guidance from the Copyright Office.

IV. Existing models for compulsory and collective licensing are misaligned with the vast universe of copyrighted material that could find its way into AI training data and the

[33] E.g., Kelly, 336 F.3d at 818.
[34] E.g., Authors Guild, 804 F.3d at 220 (citing WILLIAM F. PATRY, PATRY ON FAIR USE § 4.1 (2015)).
[35] *Id.*
[36] *Id.* at 221 ("Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original.").
[37] *Id.* at 222.
[38] Mark A. Lemley, *Dealing with Overlapping Copyrights in the Internet*, 22 U. DAYTON L. REV. 547, 566 (1997) ("[T]he fair use analysis is extremely fact-specific, which means both that it is hard to predict in advance and that it will be expensive to prove.").

actual value each piece of copyrighted content provides to AI models. (Addresses questions 10.2, 10.3 and 13.)

As discussed above, startups should not need licenses to train their AI models on copyrighted materials, both because that should be considered a noninfringing use under the law and, if it were to be considered a use, it would be protected by fair use. On top of that, the creation of licensing requirements and large scale licensing mechanisms—including collective and compulsory licenses—would severely limit the ability of startups to participate in the AI ecosystem, and they would not create opportunities for compensation for the use of copyrighted material in line with the value of the use of that material. As discussed above, AI models need to be trained on large and varied data sets, which may or may not contain copyrighted material to produce high-quality, relevant outputs. The combination of the need for diverse data sets that could contain anything in the universe of expressive material eligible for copyright protection and the indirect—and even diminishing—value of each individual piece of data that an AI model is trained on, means that no existing model for large scale licensing can be easily applied to AI training and development.

Existing collective licensing mechanisms have "been most successful in the context of homogeneous transactions among repeat players with similar preferences,"[39] which does not describe the way AI models interact with copyrighted material or the state of the AI ecosystem. Examples like music performing rights mechanisms—where businesses like nightclubs pay one of a small number of performing rights organizations for the ability to play a catalog of music—deal with a specific kind of customer seeking a license for a frequent and specific kind of use of a specific kind of copyrighted content. That dynamic does not exist in the realm of training AI models, where, depending on the model, all kinds of expressive material that might be copyrighted could be included and, at scale, the value of each individual piece of data in a training data set is unclear (if not negligible). A collecting licensing mechanism would almost certainly only provide partial representation of accessible copyright material, leaving startups to negotiate with rightsholders outside of those collecting licensing organizations on a case-by-case basis, which would be incredibly time consuming and costly. Additionally, the collective licensing mechanisms for music performing rights have been criticized for the impact they have on rightsholders, the public, and innovative technologies,[40] and creating similar mechanisms for AI training data would run the risk of mirroring those impacts on a still-developing ecosystem.

Existing compulsory license models are also not a good fit for AI training data broadly, but especially for startups looking to compete in the AI space. The Copyright Office has called for the sunset of secondary transmission compulsory license requirements, noting in response to a 2019

[39] Robert P. Merges, *Contracting into Liability Rules: Intellectual Property Rights and Collective Rights Organizations*, 84 CAL. L. REV. 1293–94, (1996), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=11497.
[40] Jonathan Band and Brandon Butler, *Some Cautionary Tales About Collective Licensing*, 21 MICH. ST. INT'L L. REV., 687 (2013), https://commons.msu.edu/deposits/objects/hc:35560/datastreams/CONTENT/content.

request from the House Judiciary Committee[41] the many ways in which the video programming compulsory license regimes have distorted the market against new competitors offering video programming using different technology and result in copyrighted material being undervalued and secondary administrative costs for rightsholders.

V. AI developers shouldn't be held liable when users use general purpose AI tools to create infringing outputs. (Addresses question 25)

None of the above is to say that there are not significant copyright questions and potential infringement issues that arise when AI models generate content that mirrors and potentially displaces copyrighted material, which is and should be a separate question from whether the use of copyrighted material in wide-ranging and diverse training data sets is considered copyright infringement. It is undoubtedly true that generative AI will—despite its developer's best intentions and any technical limitations in place—generate content that infringes on something in the vast universe of copyrighted material when directed by a user to do so. As the Copyright Office has done with questions around the copyrightability of AI generated content, the focus on these inquiries should be on the level of human authorship and control.

In instances where a user is directing an AI model to generate content that infringes existing copyrighted material and then uses that generated content in some way that's not protected by fair use, the developer of the AI tool should not face contributory liability. Like the copying equipment at the heart of *Sony*, generative AI "is widely used for legitimate, unobjectionable purposes," and, under that precedent, "need merely be capable of substantial noninfringing uses."[42] Even generative AI models that have a specific purpose—such as a customer service AI chatbot—do not set out to generate infringing content, and the overwhelming majority of uses will not involve generating content that infringes copyrighted material. But if a developer faced any kind of liability and the risk of hefty legal fees and statutory damages if its AI generated copyright infringing content even once, AI developers would face unreasonable risk and expense, significantly chilling startup participation in the ecosystem.

* * *

Engine appreciates the Copyright Office's attention to the intersection of copyright law and artificial intelligence and the chance to provide comment. The needs and perspectives of the nation's startups and technology entrepreneurs should be considered in these conversations, and we are always available to be a resource as the Office continues to develop policy in this space.

[41] *U.S. Copyright Office Analysis And Recommendations Regarding The Section 119 Compulsory License*, U.S. Copyright Office (June 3, 2019), https://www.copyright.gov/laws/hearings/views-concerning-section-119-compulsory-license.pdf.
[42] *See* Sony Corp. v. Universal City Studios, 464 U.S. 417 (1984)