VIA Electronic Submission by Lukas Ruthes Gonçalves on behalf of Engine

December 06, 2023

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

Re: Reply Comments of Engine to USCO's Notice of Inquiry on Artificial Intelligence and Copyright, Docket No. 2023-6 concerning FTC's Comments dated October 30, 2023

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 officials and a community of thousands of high-technology, growth-oriented startups across the nation to support a policy environment conducive to technology entrepreneurship. As stated in our previous comments, many startups are currently developing, using, or moving to integrate 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 changes that would make it more difficult for developers to build and train AI and integrate it into their products and services.

Several commenters addressed issues of innovation and startup competitiveness in their submissions, but some commenters assume concerning positions out of line with technical realities and understanding of copyright that would harm, rather than enhance, the position of smaller AI competitors. The Federal Trade Commission (FTC), writes that increasing regulation in the copyright space—by requiring licenses for training data, for example—would improve competition. This claim is inaccurate and would instead only benefit large entities and large rightsholders, leading to increased market concentration and threatening competition in the development of AI technologies.

I. Startups and emerging firms are at the forefront of the Generative AI revolution, not just large incumbents.

In their comments,[1] the FTC first mentioned that due to the accelerated pace of AI technology's development, many generative tools had been made available to the companies and the public in general. They state that some of the largest companies in the world use machine learning for various purposes such as "process user inputs, identify fraud, and target advertising,

[1] *See*, *COMMENT OF THE UNITED STATES FEDERAL TRADE COMMISSION*, *Docket No. 2023-6*, at 2 (Oct 30, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/p241200_ftc_comment_to_copyright_office.pdf

while AI chatbots and other applications have publicly launched and reached hundreds of millions of people".

What these comments do not consider is that, as mentioned in Engine's previous submission,[2] the AI ecosystem is much bigger and more diverse than only the bigger companies, and considerations of how copyright law applies to generative AI training should take all players into consideration.

Startups are already using AI and, more specifically, generative AI, to provide innovative products and services to users, sometimes in ways that might not even be easily perceived. That includes improving water quality,[3] reducing energy waste,[4] improving equitable access to financial systems,[5] creating better health outcomes,[6] and more.

The FTC then argues that the pace at which AI is being developed and deployed poses potential risks to competition.[7] More specifically, the Commission is concerned that the rising importance of this technology to the economy "may further lock in the market dominance of large incumbent technology firms." These powerful companies would control many of the inputs needed for creating effective AI tools and might have the incentive to "use their control over these inputs to unlawfully entrench their market positions in AI and related markets, including digital content markets."[8]

However, the generative AI revolution is not the exclusive domain of these large incumbent technology firms. In many ways, these firms have been "playing catch up" to smaller companies.[9] This is evidenced by rankings that include emerging market leaders like OpenAI,[10] Stability AI,[11] and Anthropic,[12] as well as several startups who provide services in diverse fields from marketing and creative projects to healthcare and life sciences. This only serves to show that, contrary to what the FTC claimed in their comments, the AI revolution includes new entrants and by startups. Consequently, the solutions proposed by the Commission might actually harm competition instead of fostering it, because the FTC's proposition will heighten barriers to entry and success in AI.

II. Considering the use of copyrighted material in AI training data sets as copyright infringement will undermine, not foster competition.

The FTC, in commenting on matters of liability, suggests that training an AI model with information that includes copyrighted materials could constitute an unfair method of

---

[2] *See*, *Comments of Engine to the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, Docket No. 2023-6*, at 1 (October 30, 2023), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/6541224062cc713bb5eea9d7/1698767424782/Engine+comments+to+CO+on+AI+NOI+10.23.pdf

[3] *See*, e.g., Varuna, https://varuna.city/

[4] *See*, e.g., COI Energy, https://www.coienergy.com/.

[5] #StartupsEverywhere Profile: Kenneth Salas, Co-Founder & COO, Camino Financial, Engine (May 20, 2022), https://www.engine.is/news/startupseverywhere-losangeles-ca-caminofinancial.

[6] #StartupsEverywhere Profile: Noelle Acosta, Founder & CEO, Noula Health, Engine (Oct. 28, 2022), https://www.engine.is/news/startupseverywhere-newyork-ny-noulahealth.

[7] FTC, *supra* note 1, at 4.

[8] *Id.*

[9] See, Mike Masnick, FTC Gets Fair Use Backwards, Claims It's Somehow Anti-Competitive? (Nov. 9, 2023), https://www.techdirt.com/2023/11/09/ftc-gets-fair-use-backwards-claims-its-somehow-anti-competitive/

[10] *See*, e.g, OpenAI, https://openai.com/

[11] *See*, e.g, Stability AI, https://stability.ai/

[12] *See*, e.g, Anthropic, https://www.anthropic.com/

competition.[13] Their submission claims the misuse of materials protected by copyright could be regarded as an unfair practice or unfair method of competition under the FTC Act. Specifically, the Commission says:

> How should liability principles apply to harm caused by AI tools trained on creative work that are used to generate new content? How should liability be apportioned among users, developers of AI tools, and the developers of the training dataset? How should this analysis take into account that training data is often scraped from sources hosting pirated content? Should the availability of disclosures regarding training materials or the lack of such disclosures affect liability? How would such liability principles apply to open-source AI models? These liability questions implicate consumer protection and competition policy. For instance, under certain circumstances, **the use of pirated or misuse of copyrighted materials could be an unfair practice or unfair method of competition under Section 5 of the FTC Act.** [Emphasis added].

Making data training sets available to the greatest number of competitors should be celebrated, not maligned as an unfair method of competition. Enabling startups to train models on publicly available data on the Internet, by creating certainty that doing so is not infringing, will benefit competition by lowering barriers. Conversely, introducing additional barriers by, e.g., requiring licensing in these cases would negatively affect only startups, while large firms would be able to afford fees of the sort and entrench their market power.

As we shared in our initial comments, the current copyright framework allows startups on bootstrap budgets to build and scale these AI-powered tools.[14] 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 afford to seek out and 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. The FTC's intent to regard as an unfair method of competition "training an AI tool on protected expression without the creator's consent"[15] is likely out of line with current understanding of copyright law and would cause an unforeseen financial burden that only large firms would be able to bear. As we explained in comments to the U.S. Patent and Trademark Office in 2020:[16]

> 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.[17] 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

[13] FTC, supra note 1, at 5.
[14] Engine, *supra* note 2, at 2.
[15] FTC, *supra* note 1, at 5.
[16] *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/5e1c9b95bf2cc11b00b9e944/1578933141931/2 020.01.10_Comments+to+Docket+PTO+C+2019+0038.pdf.
[17] *See*, e.g., Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power*, N.Y. TIMES, Sept. 26, 2019.

might be subject to copyright claims. They need to be able to do this without fear of infringement accusations.

And, as is frequently the case, any increase in regulatory and compliance costs will fall disproportionately on startups. These companies have 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.[18] 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.[19] 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 is 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:[20]

> 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.

These numbers are from 2009, and the data to train models are measured in billions—at least three orders of magnitude larger than this example—making it easy to imagine these costs, applied to questions at issue before the Copyright Office, are much higher.

Ultimately, copyright "is a monopoly granted by the government to authors only for the purpose of providing them with an economic incentive to create works for public benefit; and that this monopoly contains important limitations to ensure that the public receives that benefit."[21] Licensing requirements would be impractical and would dramatically chill innovation and the ability of startups, specifically, to compete in the AI ecosystem and diminish benefits to the public.

---

[18] Engine, *supra* note 2, at 4.
[19] *See*, *The State of the Startup Ecosystem, Engine*, at 17 (April 2021) https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/60819983b7f8be1a2a99972d/1619106194054/T he+State+of+the+Startup+Ecosystem.pdf.
[20] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. at 48-50(2009)
[21] Project Disco, *The FTC Veers Wildly Out of Its Lane in Comments on Copyright and Artificial Intelligence* (Nov. 27, 2023), https://www.project-disco.org/intellectual-property/the-ftc-veers-wildly-out-of-its-lane-in-comments-on-copyright-and-ai/

III. The inclusion of copyrighted material in AI training data should not trigger questions about whether a license is required.

If the objective is to keep the AI ecosystem innovative, competitive, and accessible to startups, the most efficient resolution is to determine that the ingestion of copyrighted content as part of a training data set is a lawful and noninfringing under copyright law, stopping any inquiries into infringement before the question of fair use even 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 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]

As the Copia Institute argues in their comments concerning AI training, the software would be reading or otherwise consuming works on behalf of the people developing it, which is an activity not forbidden by copyright law.[24] Such a right, the right to read (or more broadly, the right to receive information and ideas) is found in the First Amendment.[25] In other words, according to Copia,[26] "the developers of an AI system would have the right to read all the works themselves. But that right is not curtailed by the use of tools – including software tools – to help them do that reading."[27]

As Engine highlighted in their previous comments,[28] the way that AI models interact with data including copyrighted material, by reading, listening or hearing, 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."[29]

Comparative law already provides us with an example that could be followed towards considering the training of AI applications a noninfringing practice when it comes to copyright law. In May of 2018, Japan approved a reform of their Copyright Act "that focused on allowing much-needed flexibility and legal certainty for innovators."[30] The objective of this reform was to "promote innovative digital and Artificial Intelligence (AI) services that are emerging or will

---

[22] Engine*, supra* note 18, at 3.
[23] 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).
[24] See, COMMENT OF THE COPIA INSTITUTE In the Matter of Artificial Intelligence and Copyright, Docket No. 2023-6, at 4 (Oct. 30, 2023), https://s3.documentcloud.org/documents/24113231/copiainstitute_copyrightoffice_aistudy.pdf
[25] *See*, *e.g.*, *Board of Education v. Pico*, 457 U.S. 853, 866-67 (1982).
[26] Copia, *supra* note 27, at 4.
[27] *See also, Reno v. ACLU*, 521 U.S. 844, 860 (1997) (finding that the First Amendment applies to the use of computer technology to aid in the exercise of free expression).
[28] Engine, *supra* note 2, at 6.
[29] Sobel, supra note 23, at 62-63 (citing cases); see also 17 U.S.C. § 101 (defining "copies").
[30] European Alliance for Research Excellence, *Japan Amends Its Copyright Legislation to Meet Future Demands in AI And Big Data* (Sep. 03 2018), https://eare.eu/japan-amends-tdm-exception-copyright/

emerge in the future, primarily by removing ambiguity for using copyrighted works for understanding and analysis."[31] Fundamentally, the new article 30-4 of the law lets AI developers use data or information in a form where the copyrighted expression of the work itself is not retained or recreated for future use. Additionally, the Japanese law also underscores that while ingesting copyrighted works for training is not infringing, outputs of AI models can be infringing, which is a separate and distinct question.

Therefore, there are already examples that show that the type of training to build AI models is noninfringing, not covered by copyright law. In any case, 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.

IV. **Conclusion.**

Startups should not need licenses to train their AI models on copyrighted materials, both because that should be considered noninfringing 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 would severely limit the ability of startups to participate in the AI ecosystem, thus increasing the chances of market concentration in the hands of the large incumbents.

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. The FTC's position, in this case, is one that would harm competition and make it more difficult for the greatest number of players to engage in the market in as equal a footing as possible.

[31] *Id.*