Engine

January 10, 2020

Andrei Iancu
Under Secretary of Commerce for Intellectual Property
Director of the United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

VIA EMAIL

Re: Comments of Engine Advocacy in Response to *Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation*, Docket No. PTO-C-2019-0038

Dear Director Iancu:

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 the development of technology entrepreneurship. We appreciate the United States Patent and Trademark Office's ("USPTO") attention to pursuing policies that foster innovation. And we further appreciate the opportunity to submit these comments regarding the impact of artificial intelligence ("AI") technologies on intellectual property ("IP") law and policy, as well as the impact IP law and policy could have on emerging AI technology.

Existing statutory language and case law appropriately permit the use of potentially copyrightable material in training, tuning, and testing AI systems. These comments address the third question raised in the USPTO's request, and explain how the law applies to render such use lawful.

Overall, the relevant legal and policy frameworks are working well. To the extent any changes are pursued, those changes should only codify the legal standards as articulated below and instill increased certainty that these standards will govern the development of AI technology. But if future developments in AI render the current law unduly ambiguous or unclear about the infringement liability associated with AI processing content, then it may be appropriate to amend relevant statutes to confirm that such uses are noninfringing.

I. **Introduction**

Under existing law, it is lawful to ingest content and use it to train, tune, and/or test AI systems. Such uses either involve content that is ineligible for copyright protection, involve uses that are not infringing, or involve fair uses of content.[1] While there may be ways to streamline or codify the law to make it more certain and predictable, existing law is working well.

The USPTO has asked how copyright law does (or should) apply to the process of AI ingesting content and using it as the AI system learns its functions. Ingesting involves bringing data into an AI system, and, for example, filtering, transforming, integrating, and/or validating it as necessary to ensure data quality.[2] And in prior comments, we used the terms training, tuning, and testing to describe the processes through which an AI system learns its functions, and use that same language in these comments.[3] While each disparate approach to and application of AI might call for slightly different treatment under copyright law, there are common features which make ingesting content and using it to train, tune, or test lawful.

AI is already pervasive in our everyday lives, and there are ever (and rapidly) expanding advanced AI technologies and commercial applications on the horizon. AI systems are designed, and will continue to be designed, to solve many different types of problems in reliance on many different types of data. Similarly, content that is potentially eligible for copyright protection[4] is ubiquitous and varied. Copyright protection can apply to *any* "original works of authorship fixed in [a] tangible medium of expression."[5] 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.[6] As such, the data used to train, test, and tune AI systems may be taken from content that is potentially eligible for copyright protection.

An example is useful for understanding why ingesting copyrighted material should be lawful, and why changing the law would be problematic. Internet platforms and service providers are

[1] One possible exception, discussed in Part III below, turns on whether the USPTO considers "ingesting" to include gathering content (i.e., collecting the material that is compiled into the dataset). If data gathering, or collecting, is part of ingesting, and if an AI system's developer creates unauthorized copies of copyrighted material at that point, that could constitute infringement. The infringement analysis would still depend, at least, on whether the gathered content is in fact eligible for copyright protection and whether the system's use was a fair use.

[2] *See, e.g.*, Alistair Croll, *The Feedback Economy*, *in* PLANNING FOR BIG DATA 1, 4 (Edd Dumbill ed., 2012) (defining ingesting and cleaning); SUNILA GOLLAPUDI, PRACTICAL MACHINE LEARNING 70 (describing the data ingestion layer).

[3] *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 4 (Nov. 8, 2019).

[4] For the purposes of these comments, the term *content* is used as shorthand for *content that is potentially eligible for copyright protection*.

[5] 17 U.S.C. § 102(a).

[6] *See* 17 U.S.C. § 102(b).

under increasing pressure to affirmatively identify user-generated copyright infringement. While current technology is (putting it mildly) far from perfect, many companies are trying to develop AI systems to improve automatic detection of potential infringement (or at least narrow the set of potentially infringing material that requires human review).[7] To develop and train those systems, companies need to feed copyrighted material into algorithms—a system that needs to draw lines between infringement and noninfringement can only do that if it knows what infringement looks like. But if using copyrighted material to train, tune, and test such a system is itself infringement, it destroys the incentive (or at least business case) for doing the development work, because the development itself is costly copyright infringement.

Policymakers should not change the status quo. If ingesting content were unlawful, that would open AI startups and innovators to unbearable costs and risks. It would ultimately slow, and could even stall, domestic AI development and along with it American leadership in the field. Overall, while there may be areas of relevant copyright law and policy that are not entirely settled, the existing interpretations described herein are working well and should continue to control.

II. **For datasets that exclude any expressive content, all use is lawful**

All AI systems involve volumes of data at the foundation of their development. And in many cases, that data may be taken from copyrighted works. But if 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.[8] For AI systems that exclusively analyze factual information extracted from content, the copyright inquiry should end there, because that factual information is not copyrightable.

[7] *See, e.g.*, Adam Satariano, *Europe Adopts Tough New Online Copyright Rules Over Tech Industry Protests*, N.Y. TIMES, Mar. 26, 2019; Krista L. Cox, *Automated Copyright Filtering Removes Public Domain Mueller Report From Platform*, ABOVE THE LAW (May 9, 2019, 11:17 AM), https://abovethelaw.com/2019/05/automated-copyright-filtering-removes-public-domain-mueller-report-from-platform/; Evan Engstrom & Nick Feamster, *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools* (Mar. 2017), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf.

[8] *See, e.g.*, Daryl Lim, *AI & IP: Innovation & Creativity in an Age of Accelerated Change*, 52 AKRON L. REV. 813, 850-51 (2018) (noting that facial recognition databases compiled from news images may not even invoke fair use if the portions of the photos taken is minimal); Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 67-68 (2017) (similar).

III. <u>Ingesting data and using it during training, tuning, and testing is not infringing</u>

AI systems are trained, tuned, and tested using volumes of content, and that use is not infringement. That sort of use is most analogous to a person reading a book, listening to a song, or viewing a painting. None of those personal uses are copyright infringement, and AI ingesting and processing content should be treated similarly.

Copyright law has traditionally "left reading, listening, and viewing unconstrained."[9] Not only can people can read books, listen to records, and look at art in museums without running afoul of the law. They can jot down their impressions while they read or recall a movie quote without violating an author's rights.[10] The law excludes myriad similar, technology-based uses from copyright infringement liability. For example, people are allowed to copy digital music files from their computer to an MP3 player.[11] And many of us back-up our hard drives on a periodic basis, making archival copies of any copyrighted material we own. The law and society (either explicitly or effectively) treat these as lawful.[12]

As we have previously noted, "AI technologies perform tasks that conventionally require human intelligence, such as learning, reasoning, and perception."[13] AI technology is performing human-like tasks, and it is engaging with content like humans do when they read, listen, or view. Therefore, when an AI system analyzes content during training, tuning, or testing, just like a lawful personal use, that should be outside the scope of infringement.

The basic development process for an AI system includes ingesting and preparing content as well as a loops of training, testing, and tuning when the content is analyzed and processed.[14] Those uses of content do not involve reproduction, distribution, performance, public display, or creation of derivative works within the meaning of the statute.[15] Admittedly, during ingestion, training, testing, and tuning content "may be copied, emulated, and re-copied thousands of times during

[9] Jessica Litman, *Lawful Personal Use*, 85 TEX. L. REV. 1871, 1882 (2007).

[10] *See, e.g.*, *id.* at 1893 ("It would plainly be unconstitutional to prohibit a person from singing a copyrighted song in the shower or jotting down a copyrighted poem he hears on the radio.") (quoting Justice John Paul Stevens).

[11] Recording Indus. Ass'n of Am. v. Diamond Multimedia Sys., Inc., 180 F.3d 1072, 1079 (9th Cir. 1999) ("the purpose of the Act is to ensure the right of consumers to make analog or digital audio recordings of copyrighted music for their private, noncommercial use") (citations omitted).

[12] *See, e.g.*, Litman, *supra* note 9, at 1895-1898, 1902 (discussing statutory exceptions to copyright infringement, such as making backup copies of computer programs (section 117) and making noncommercial copies of recorded music (section 1008), exceptions carved out in case law, and personal uses that do not fall within an explicit exclusion but nonetheless constitute common personal uses of content that would, uncontroversially, be considered noninfringing).

[13] *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 3 (Nov. 8, 2019).

[14] *Id.* at 4, 17-18.

[15] *See* 17 U.S.C. § 106 (defining a copyright owner's exclusive rights).

the learning process."[16] This could include creating ephemeral copies, that are so transitory in nature that they do not even constitute creating a *copy* as defined in the statute.[17] And depending on the AI method or model used, humans may have no insight into how the AI system is analyzing and processing data, nor what the intermediary content looks like.[18] For these hyper-technical instances of copying or modification, "the spirit of the copyright statute seems to exempt this type of copying."[19] Just like courts "refuse to entertain discovery with respect to early drafts of a noninfringing final work" on the basis that those interim and unpublished drafts are not infringements,[20] the interim status of content during the AI training process should be free from infringement scrutiny.

The one caveat to this assessment of noninfringing uses turns on the definition of "ingesting." If the USPTO understands "ingesting" content to encompass the collection of content in the first instance for inclusion in a dataset, it is possible that developers would make digital copies of content during that collection (or data gathering) process. For some AI applications, and for some developers (particularly those that do not have in-house access to data), it may be necessary to copy content "in order to process [it] as grist for the mill, raw materials that feed [] algorithms."[21] Because, at the very least, that initial copying would qualify as fair use, it is therefore also lawful.[22]

IV. <u>Even if such uses were potentially infringing, they would be lawful fair uses</u>

In the interest of promoting progress and innovation, it would be better to resolve that AI use of content is lawful because it is a noninfringing use. The alternative, fair use, is decided on a case-by-case basis.[23] Proceeding through litigation to establish that a specific use is fair is costly and not dispositive for all future (even similar) uses of data.[24] So while it is a fair use for AI to ingest and process data, it is more efficient to conclude that such uses are not even infringing.[25]

When content is used to train, tune, and/or test an AI system, if ingesting and processing that content were determined to be an infringing use then it would a lawful fair use. 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

[16] Sobel, *supra* note 8, at 62-63.
[17] Sobel, *supra* note 8, at 62-63 (citing cases); *see also* 17 U.S.C. § 101 (defining "copies").
[18] *See, e.g.*, Davide Castelvecchi, *Can We Open the Black Box of AI?*, 538 NATURE 21 (2016).
[19] Sobel, *supra* note 8, at 63.
[20] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1635-36 (2009).
[21] *Id.* at 1608.
[22] *Infra* part IV.
[23] *E.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).
[24] *See, e.g.*, Litman, *supra* note 9, at 1902-03.
[25] *Supra* part III.

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."[26] And in the case of AI, it does not supersede the original content it relies on in training, tuning, and testing. Instead, 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."[27] For example, making a digital copy of a book so that it is easier for people to search within books is transformative.[28] Making thumbnail copies of images to help index and improve access to content on the Internet is also transformative.[29] And archiving student-written term papers within a database for an online plagiarism detection technology is transformative.[30] AI likewise recharacterizes content.

Courts also look at the commercial nature of a use when assessing this first factor. But a commercial motivation cannot outweigh an otherwise transformative use,[31] so AI developed with a commercial application in mind can still be a fair use. Like the use of thumbnail images in a search engine (which is both commercial and transformative), AI systems do not use the content in datasets to directly profit, and are not 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.[32] So even where an AI developer has an overarching commercial purpose, that commercial aspect does not defeat its transformative nature.

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.[33] 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.[34]

[26] Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003).

[27] Sag, *supra* note 20, at 1646.

[28] Authors Guild v. Google, Inc., 804 F.3d 202, 216-17 (2d Cir. 2015).

[29] *Kelly*, 336 F.3d at 818.

[30] A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640 (4th Cir. 2009).

[31] *Authors Guild*, 804 F.3d at 219.

[32] *E.g.*, *Kelly*, 336 F.3d at 818.

[33] *E.g.*, *Authors Guild*, 804 F.3d at 220 (citing WILLIAM F. PATRY, PATRY ON FAIR USE § 4.1 (2015)).

[34] *E.g.*, *Authors Guild*, 804 F.3d at 220.

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.[35] What matters is "the amount and substantiality of what is [] made accessible to a public for which it may serve as a competing substitute."[36] And in the context of AI ingesting or processing content, the answer is *none*. When an AI system is using content during training, testing, or tuning, it does not make anything accessible to the public, certainly no competing substitutes. This third factor also weighs in favor of finding fair use.[37]

Factor (4): Effect upon the potential market. It is unlikely that AI ingesting data will harm the market for the original work, because copyrighted works are developed to entertain, impress, or inform human audiences. And an 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.[38] In any event, because the purpose and character of AI's use of content is highly transformative, that highly transformative nature suggests there will not be market harm.[39]

V. **Allowing AI systems to ingest content and be trained free from copyright liability promotes innovation**

Existing law, as described above, is working well. And there are strong policy reasons against making changes. As one scholar noted, "[h]ow copyright law treats the use of [AI training] datasets will determine whether AI-generated works can reliably develop without a constant threat of litigation."[40] Changing copyright law or policy in a way that creates liability when AI systems ingest content would put innovation at risk.[41]

[35] *Authors Guild*, 804 F.3d 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.")

[36] *Id.* at 222.

[37] *See, e.g.*, *id.* at 221-22.

[38] *See, e.g.*, Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003) (finding use of thumbnail images in search engine did not harm photographer's ability to sell or license full-sized images); A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 643 (4th Cir. 2009) (finding that archive of student-written papers in an online plagiarism detection tool "did not serve as a market substitute or even harm the market value of the works").

[39] *See, e.g.*, *Kelly*, 336 F.3d at 818 (citing *Campbell*, 510 U.S. at 586-87).

[40] Lim, *supra* note 8, at 847-48.

[41] *See, e.g.*, Christian Handke et al., *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research*, in New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust 120 (B. Schmidt & M. Dobreva eds., 2015) (finding that data mining-related research is increasing globally, but for countries in which data mining requires consent from copyright owners, data mining reflects a smaller share of academic output).

Developing AI is a national priority, which means promoting innovation and not creating new hurdles.[42] 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.[43] If developers faced those sorts of burdens when compiling datasets, it would slow (and could stall) progress.

Relatedly, 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.[44] 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."[45] These costs and risks would scare many innovators, companies, and investors away from developing new AI systems.[46]

Lawful use of content can also help prevent bias problems in AI. If it is difficult for developers to navigate the copyright landscape when developing datasets, it could result in them leaving certain content out and creating a non-comprehensive dataset subject to bias. Access to content free from claims of copyright infringement gives developers access to more diverse, less biased content. For example, recently created (and therefore currently copyrighted) works may be less susceptible to inherent or traditional biases, when compared to public domain content that was created when the U.S. was less racially diverse and had more gender disparity.[47] And "[b]ias in AI may be exacerbated by a restrictive fair use doctrine," because if training data "is protected by copyright, those who use [it] do so secretly, preventing biases from being uncovered."[48]

[42] *See, e.g.*, *Artificial Intelligence for the American People*, WHITEHOUSE.GOV, https://www.whitehouse.gov/ai/ (last visited Jan. 7, 2020) (recognizing that "America has long been the global leader in this new era of AI" and describing five pillars for advancing AI in the U.S., including "removing barriers to AI innovation"); National Security Commission on Artificial Intelligence, *Interim Report*, at 1 (Nov. 2019), *available at* https://drive.google.com/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view (describing AI as "integral to the technological revolution that we are now experiencing," and expressing concern that "America's role as the world's leading innovator is threatened").

[43] Sag, *supra* note 20, at 1659-61 (describing high transaction costs that would be encountered if Internet search engines had to obtain copyright clearance for all searchable content).

[44] 17 U.S.C. § 504(c).

[45] Sobel, *supra* note 8, at 80; *see also* Lim, *supra* note 8, at 847-48 (similar).

[46] *See, e.g.*, Sobel, *supra* note 8, at 80-81.

[47] *See, e.g.*, Louise Matsakis, *Copyright Law Makes Artificial Intelligence Bias Worse*, VICE (Oct. 31, 2017, 12:00 PM), https://www.vice.com/en_us/article/59ydmx/copyright-law-artificial-intelligence-bias (citing Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018)).

[48] Lim, *supra* note 8, at 854.

Moreover, bias problems could arise if copyright owners were in a position to dictate when and who can develop AI technology and for which purposes.

Finally, increasing the burden on innovative companies that need access to content and data to develop AI technology disproportionately affects startups and entrenches big incumbents. 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.[49] 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.

AI development is thriving, and the recent explosion in promising AI technology occurred under the current copyright framework. Before making any changes to this framework, policymakers should carefully consider how any changes would impact the current trajectory of ubiquitous, varied, and exciting AI technologies. And policymakers should avoid making any changes that might hamper innovation.[50]

[49] *See, e.g.*, Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power,* N.Y. TIMES, Sept. 26, 2019.

[50] *See, e.g.*, *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 14-15 (Nov. 8, 2019) (addressing how changes to current patent framework might restrict AI innovation).