THE Authors Guild

Before the Library of Congress
Copyright Office

Docket No. 2023–6

Reply Comments of the Authors Guild

Artificial Intelligence and Copyright

December 6, 2023

The Authors Guild thanks the Copyright Office for the opportunity to submit the following reply comments in connection with the Office's study on copyright law and policy issues raised by artificial intelligence ("AI") systems. Given the enormous volume of comments submitted in the first round, our responses do not attempt to comprehensively address the issues discussed. In the comments below, we seek to respond only to some of the most significant legal and factual misstatements advanced by AI developers in defense of their mass-scale infringement of our members' works.

1. *Generative AI Systems Exploit Copyrightable Expression*

Several generative AI companies and their supporters argue that the mass ingestion of works does not implicate copyright owners' interests because, they argue, the companies did not use the copyrighted works they trained on for their expression but instead for the relationships between the words (in the case of textual materials) and the contexts in which they are used. *See, e.g.*, Open AI at 11; Google at 11; Meta at 13; Samuelson, Sprigman & Sag at 14-15; Andreeson Horowitz at 6. This makes no sense as that is exactly what written expression is—words are combined in original ways to create meaning and art, and that expression and meaning is exactly what generative AI copies. OpenAI, for example, contends that "when undergoing pre-training, a model is not interested in the expressive aspects of individual copyrighted works." OpenAI at 11. Instead, it explains, the copying is done so that the model can "learn[] how words fit together grammatically, but also how words work together to form higher-level ideas, and

ultimately how sequences of words form structured thoughts or pose coherent problems." *Id.* at 5.

This is a bit like saying someone who plagiarizes a piece of music is interested in the notes the composer chose and the way she organized them to create an original piece of music, but is not interested in the melody and rhythm. What these commenters describe is precisely what copyrightable expression is—the particular selection and arrangement of words the author has chosen to convey their ideas. *See, e.g.*, *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 348 (1991) ("Others may copy the underlying facts from the publication, but not the precise words used to present them."). And, far from not being "interested" in that expression, they are copying it because it is valuable to them: the authors' choices with respect to syntax, tone, jargon, and innumerable other creative elements are what allow the AI model to generate higher-quality outputs, which of course can then be used to compete with and displace the originals in the marketplace. This cannot be characterized as a non-expressive use.

Expert testimony cited with approval by Professors Samuelson, Sprigman, and Sag (at 8) reinforces this conclusion. Computer scientist Christopher Callison-Burch explained to the House Judiciary Committee's Subcommittee on Courts, Intellectual Property, and the Internet that, in addition to language rules and factual information, AI systems learn the following from pre-training:

> Ideas and opinions: AI systems learn about different perspectives, opinions, and ideas expressed in their training data. This enables them to understand and generate text that reflects diverse viewpoints, although it may also lead to the propagation of controversial or biased opinions.
>
> Limited common sense reasoning skills: Pre-trained AI systems gain some capacity for common sense reasoning, which allows them to understand basic cause-and-effect relationships, infer missing information, and make simple deductions. However, this ability is limited and often falls short when compared to human reasoning.[1]

[1] Written Testimony of Christopher Callison-Burch at 10, House Jud. Comm., Subcom. on Courts, Intell. Prop., and the Internet, Hearing on Artificial Intelligence and Intellectual Property: Part I–Interoperability of AI and Copyright Law, May 17, 2023,

Thus, AI systems are not simply extracting information *about* the ingested works. They are consuming, and learning from, the works' expressive content.

Meta, for example, asserts (at 13) that LLMs are just like an author who learns language by "reading hundreds or thousands of existing books, consciously or subconsciously identifying patterns and concepts from those books." Notwithstanding that comparing computers to humans is a false equivalency which obscures the fact that machines can memorize, mimic, and mass-produce books at scale in a way that no human ever could, ingesting the expressive content of books in that manner can hardly be called a non-expressive use. And, of course, Meta fails to mention that the books and other copyrighted works used to teach humans are typically paid for—either by the readers themselves, or by the library, educational institution, publication, or other entity that makes them available.

Moreover, use of the LLMs clearly demonstrates that the AI memorized texts it was trained on. For instance, prompts using ChatGPT, Anthropic's Claude AI, Meta's Llama, Google's Bard, and others have resulted in outputs that included identical passages from many works queried, paraphrased the works accurately, and provided character names and attributes, settings, and story lines from text queried about. It is nonsense to argue that the AI did not copy expression from works it ingested.

2. *AI Companies Overstate the Complexities of Licensing Works for AI Training*

Another common refrain among comments of AI companies and their supporters is that because AI training involves the scraping and ingestion of millions of works from the open internet, the licensing of such materials is impossible and that the use must therefore be excused. *See, e.g.*, LCA at 1; OpenAI at 12-13; Andreeson Horowitz at 10. As an initial matter, and as discussed in our first-round comments, the scale of the companies' infringement does not help them under the fair use analysis. It would turn copyright law on its head to hold that a party can avoid liability so long as its infringements are too numerous to account for. Conducting illegal activity at a massive level is never excused simply because the purpose of that activity requires the mass use.

---

https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/callison-burch-testimony-sm.pdf (numbering omitted).

Moreover, this argument mistakenly assumes that the fair use calculus—and thus the need for a license—is the same with respect to every work ingested by the AI systems. Fair use by its nature requires a case-specific inquiry, including analysis of the nature of the work under factor two and of the actual or potential market effect for the work under factor four. The analysis surely would differ as between, for example, a published novel and social media posts written without any reasonable expectation of commercial benefit. *See, e.g.*, *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 930 (2d Cir. 1994) (courts consider "traditional, reasonable, or likely to be developed markets" under fourth factor). Indeed, AI developers have admitted that they need books in training because of the high quality of writing and richness of details. *See, e.g.*, Yukun Zhu et al., *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* 1 (2015), *available at* https://arxiv.org/pdf/1506.06724.pdf ("Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story.").

The Office should reject the AI companies' effort to cast licensing as an all-or-nothing proposition. The fact that ingestion of certain lower-value materials may be deemed fair use in some circumstances does not obviate the need for licensing of any and all other works.

In any event, as discussed in our initial comments, the licensing of books and other copyrighted works for AI training is eminently feasible. The Authors Guild and others are currently building such systems based on well-established collective licensing structures in place around the world. We would welcome the AI companies' partnership in this effort. If they had asked permission first, we could have delivered before they used our members' works without permission or compensation. Arguments that it is too expensive do not justify the use. AI companies are spending millions and even billions on development and computing power. Why should the authors' contribution be free for the taking when generative AI is nothing without the works it is trained on? This is especially true for books that were stolen from pirate websites and for books, journalism, and other works where uses of the system can and will replace the human-authored works if left unrestricted.

3. Reverse Engineering Case Law Is Inapposite

The comments of several AI companies and their supporters rely heavily on two Ninth Circuit cases, *Sony Computer Entertainment Inc. v. Connectix Corp.*, 203 F.3d 596 (2000), and *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1993), which found fair use where the copying of computer code was necessary to access the code's functional elements for purposes of creating interoperable products. *See, e.g.*, Google at 11; OpenAI at 11 n.44; Meta at 13 n.47; Samuelson, Sprigman & Sag at 11-12, 16 n.50. The companies argue that AI training is similar to that type of reverse engineering because it "is a computational process of deconstructing existing works for the purpose of modeling mathematically how language works." Google at 11. As an initial matter, the companies are shifting their analogies to suit their purposes: while previously comparing AI ingestion to a human reader consuming works to learn from their substantive content, they now describe it as nothing more than computational analysis. For the reasons noted above, that description ignores the central importance of the works' protected expression in allowing the model to generate commercially desirable outputs.

In any event, the mass ingestion of creative works is far different from the intermediate copying in *Sony* and *Sega*. Both those cases turned on the unique nature of computer software, which "lies at a distance from the core [of copyright] because it contains unprotected aspects that cannot be examined without copying." *Sony*, 203 F.3d at 603 (citing *Sega*, 977 F.2d at 1526). The courts concluded that allowing a software developer to prevent copying necessary for reverse engineering would be anticompetitive in that it would effectively give it "a lawful monopoly on the functional concepts in its software" without meeting the requirements for obtaining a patent. *Id.* That concern is not present in the AI context, where companies are copying works at the heart of copyright's protection in order to exploit their expressive elements.

4. The Source of the Ingested Works Is Relevant to Fair Use

Professors Samuelson, Sprigman, and Sag in their comments raise the question (at 24-25) whether fair use takes into account the source of the copies used to train LLMs. They note, as we did in our initial comments, that AI systems frequently have been built using infringing copies of works scraped from notorious pirate websites like Library Genesis and Sci-Hub. The professors recognize that "it may

be deemed harmful or unfair for commercial users to bypass the market for access to train their LLMs without a compelling reason" and that "[s]uch conduct arguably undermines the economic incentives that copyright is designed to create." Samuelson, Sprigman & Sag at 24. At the same time, they contend that "there are strong arguments to be made that copying from an infringing source may still be fair use" because "[t]reating an otherwise fair use as unfair because it was made from an infringing source would lead a court to deny the public access to the products of secondary uses that fair use is designed to encourage." *Id.* at 25 (quoting Michael Carroll, Copyright and the Progress of Science, 53 U.C. Davis L. Rev. 893, 955 (2019)).

Whatever may be the significance of good faith in other contexts, the knowing use of pirated copies in the service of a commercial enterprise should weigh against fair use. Section 107 makes clear that the user's intentions matter, as it expressly looks to "whether such use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1). Thus, the inquiry is not limited to whether the activity promotes public access to the products of secondary uses. If it were, it would make no difference whether the user is pursuing a commercial purpose; all that would matter is that the use increases public access. Looking to increase public access without recognizing the copyright owner's exclusive rights is in direct conflict with the purpose of copyright law as it undermines the incentives to create and disseminate works. Needless to say, the law recognizes that the potential harm to copyright owners' interests is greater to the extent users can exploit works for commercial gain in the guise of promoting a public interest. *See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 143 S. Ct. 1258, 1277 (2023) ("If an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying.").

By the same reasoning, it is entirely proper for a court to consider whether an AI company knew or should have known that its training materials were sourced from pirated copies. In particular, courts should consider the consequences if such conduct were to become widespread. *See Harper & Row, Publrs. v. Nation Enters.*, 471 U.S. 539, 568 (1985) ("[T]o negate fair use one need only show that if the challenged use 'should become widespread, it would adversely affect the potential market for the copyrighted work.'" (citation and emphasis omitted)).

Giving multi-billion-dollar technology companies free rein to exploit knowingly pirated works for commercial gain would inevitably encourage even more piracy, leading to a cycle of infringement that would have devastating consequences for authors.