Statement of

Mary Rasenberger
Chief Executive Officer
The Authors Guild

before the
SENATE COMMITTEE ON THE JUDICIARY
SUBCOMMITTEE ON INTELLECTUAL PROPERTY

July 12, 2023

Dear Chair Coons, Ranking Member Tillis, and members of the Committee:

Thank you for the opportunity to submit this written testimony on behalf of the Authors Guild.

The Authors Guild is a national non-profit association of almost 14,000 professional writers. Since its founding in 1912, the Guild has promoted the rights and professional interests of authors in various areas, including copyright, freedom of expression, and taxation. It counts as members leading historians, biographers, academicians, journalists, and other writers of nonfiction and fiction whose works have appeared in the most influential and well-respected publications in every field. The Guild has a fundamental interest in ensuring that works of authorship and the rights of authors are protected, and that the hard work and talents of our nation's authors are rewarded so that they can keep writing, as intended by the Framers of the Constitution. The Guild believes that it is crucial for our culture and the future of democracy to ensure that our literature and arts remain vibrant and diverse.

The Challenges Generative AI Brings to the Writing Profession

Text-generating AI technologies provide useful tools that many writers are already adopting to assist them in their writing process, and their use will quickly grow. Generative AI is being used today to help writers brainstorm, research, organize their work, overcome writers' block, and think through creative problems. A survey we conducted of authors in May to gauge their views on generative AI found that 23% of the respondents used generative AI technology in their writing process, with 47% saying they use it as a grammar tool, 29% for brainstorming plot ideas and characters, 14% to structure or organize drafts, and 26% to assist in marketing. Only around 7% of writers who employ generative AI said they use it to generate the text of their work. And of this 7%, only 1.4% said they used generative AI to produce half or more of their work, while 89% said less than 10% of their work comprises AI-generated material.[1]

But even as writers are adapting to generative AI technology, they remain seriously concerned about its impact on the writing profession—as they should be. 69% of the respondents to our survey said that they consider generative AI to be a threat to their profession. An overwhelming 90% said they believe that writers should be compensated for the use of their work in "training" AI. In the last two and a half weeks, over 9000 writers and supporters have signed the Authors Guild's open letter to the CEOs of the leading AI companies calling on them to respect copyright and compensate writers fairly for the use of their works. Signatories to the letter include some of America's most well-known and well-regarded voices, such as James Patterson, Ron Chernow, Jennifer Egan, Michael Chabon, Nora Roberts, Jodi Picoult, Celeste

[1] Survey Reveals 90 Percent of Writers Believe Authors Should Be Compensated for the Use of Their Books in Training Generative AI https://authorsguild.org/news/ai-survey-90-percent-of-writers-believe-authors-should-be-compensated-for-ai-training-use/

Ng, Louise Erdrich, Suzanne Collins, Margaret Atwood, Viet Thanh Nguyen, Roxane Gay, Min Lee, Jonathan Franzen, George Saunders, David Baldacci, and many more.[2]

Use of Works Without Permission or Compensation is Inherently Unfair

Writers' works were used to so-called "train" generative AI to write, and well-written texts like books, short stories and articles written by professional writers have been copied and incorporated into generative AI to allow it to write well. To put it another way – if it was not for these well-crafted texts written by professional writers generative AI would not write well, and would be far less useful. These works were used to train the foundational large language models (LLMs) without consent, compensation, or credit. And AI systems will continue to need high-quality written to be able to respond cogently and reliably to user inputs. [3]

Where AI companies like to say that their machines simply "read" the texts they are trained on, that is inaccurate anthropomorphizing. Rather, they copy the texts into the software itself, and then they reproduce them again and again, when prompted. What's more, the works contained in the training datasets are often downloaded from pirate sites and sources. The books datasets used to train these models (and several others) were scraped and downloaded from pirate sources and "tokenized"[4] into training data without the copyright owner's consent or

---

[2] Authors Guild Open Letter to Generative AI Leaders, https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders

[3] This is because using the AI's own output in further training results in a degenerative process called model collapse, "where generated data end up polluting the training set of the next generation of models."
See, Ilia Shumailov, et al, The Curse of Recursion: Training on Generated Data Makes Models Forget, available at https://arxiv.org/abs/2305.17493 ("To make sure that learning is sustained over a long time period, one needs to make sure that access to the original data source is preserved and that additional data not generated by LLMs remain available over time. The need to distinguish data generated by LLMs from other data raises questions around the provenance of content that is crawled from the Internet: it is unclear how content generated by LLMs can be tracked at scale.")

[4] "Tokenizing" refers to the process of breaking down text into smaller units, known as "tokens," which can be words, characters, or subwords depending on training needs. Once a work is tokenized, each token is typically mapped to a unique integer or vector (in case of word embeddings), enabling the machine learning model to perform numerical computations on textual data.

compensation, in some cases by engineers working for the companies themselves and in other cases by third parties. For example, according to Washington Post's reporting, a dataset organized by Google's engineers called C4 contained books downloaded from b-ok.org,[5] a mirror of notorious pirate network Z-Library, which was indicted with assistance from the Authors Guild and is being prosecuted by the Department of Justice for criminal copyright infringement.[6] Another commonly used dataset containing almost 200,000 books called "Books3"[7]—which was used in training by Meta and some suspect by OpenAI in training GPT-3.5 and GPT-4—was created using books downloaded from Bibliotik, a pirate ebook torrent site. Despite this, Meta and other AI companies continue to say that their models are trained on "publicly available" data.[8] Even though generative AI systems ingest a wide range of text-based content, books, in particular, have a special significance to the AI's ability to generate high-quality responses. According to University of Toronto and MIT researchers who compiled one of the first books datasets for machine learning, which was subsequently was used in training early versions of GPT, Google's BERT, and Amazon's Bort and thirty other models:[9] "Books provide…very rich,

---

[5] Schaul, et al., Inside the secret list of websites that make AI like ChatGPT sound smart, Washington Post, April 19, 2023, available at https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[6] The Authors Guild assisted the FBI and the Department of Justice in the indictment of Z-Library principals and domain seizures. https://authorsguild.org/news/federal-law-enforcement-indicts-z-library-operators-with-ag-assistance/

[7] According to the Dataset card on HuggingFace, a site that hosts and provides access to machine learning resources and datasets: "This dataset contains all of bibliotik in plain .txt form, aka 197,000 books processed in exactly the same way as did for bookcorpusopen (a.k.a. books1). seems to be similar to OpenAI's mysterious "books2" dataset referenced in their papers. Unfortunately OpenAI will not give details, so we know very little about any differences. People suspect it's "all of libgen", but it's purely conjecture." https://huggingface.co/datasets/the_pile_books3

[8] In its paper on training the Llama model, Meta reveals that 4.5% of the training data comprised books obtained from Project Gutenberg and The Pile, calling the latter a "a publicly available dataset for training large language models." Tourvon, et al., LLaMA: Open and Efficient Foundation Language Models https://www.arxiv-vanity.com/papers/2302.13971/#S2.SS1.SSS0.Px5.p1

[9] Zhu et al, Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, available at, "https://arxiv.org/abs/1506.06724; see also Richard Lea, "Google swallows 11,000 novels to improve AI's conversation," The Guardian, September 28, 2016, available at, https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation

descriptive text that conveys both fine-grained visual details (how people or scenes look like) as well as high-level semantics (what people think and feel, and how their states evolve through a story).” From the early days of machine learning to the latest ground-breaking large language models, books have made up a significant portion of the training data for text-generative AI technologies: for e.g., 13% of the training data for Google PaLM, 16% of GPT-3’s training data (OpenAI has not disclosed the training data for GPT models beyond GPT3), and 4.5% of Meta’s Llama training data.[10]

The copying and use of millions of copyrighted works from the internet or illegally compiled databases without permission presents a very different scenario from that presented in prior “fair use” cases such as *Authors Guild v. Google* where the use claimed by Google was to create a database to make books searchable (and only snippets readable). Here, many millions of copyrighted works have been copied, and made a part of programs that are capable of creating material that could compete in the market with the copied works. The Supreme Court’s recent decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (598 U.S. ___, 2023) reminds courts that in any fair use analysis, the degree of transformation (and not simply whether there is any transformative use) must be weighed against the commercialism of the use, as well as the other factors. Nevertheless, courts could still find fair use in some cases, given its subjective nature and the fact that, the fair use doctrine is still ill-suited for mass use cases such

---

(quoting anonymous Google engineer: “We could have used many different sets of data for this kind of training, and we have used many different ones for different research projects….[I]t was particularly useful to have language that frequently repeated the same ideas, so the model could learn many ways to say the same thing – the language, phrasing and grammar in fiction books tends to be much more varied and rich than in most nonfiction books.”)

[10] See for e.g., Peter Schoppert, March 11, 2023, The books used to train LLMs, https://aicopyright.substack.com; see also Alan D. Thompson, What’s in my AI?: A Comprehensive Analysis of A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher, available at https://s10251.pcdn.co/pdf/2022-Alan-D-Thompson-Whats-in-my-AI-Rev-0.pdf; https://www.businessinsider.com/chatbot-training-data-chatgpt-gpt4-books-sci-fi-artificial-intelligence-2023-5; Google’s PALM paper; Open AI’s GPT3 paper, https://arxiv.org/pdf/2005.14165.pdf; Meta’s Llama training data. https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md

as here where damage to the value of each particular work might be hard to establish. It is worth noting that in some cases, generative AI developers have privately licensed specific training material, an indication of its value. But most writers and other creators have been left out in the cold, their work expropriated without limit.

Considering the value of generative AI technologies and how much they owe to professional writers, while at the same time threatening the very existence of the profession, this is grossly unfair.

Generative AI Outputs Compete with Human Authored Works and Create Risk of Market Dilution

The generative AI systems were created to allow users to generate text—based on what the machines ingested—and as such to compete with writers and to displace them. As nascent as the technology is, we are already seeing vast numbers of AI created works attempting to displace human created works. AI systems can be prompted to reproduce actual text from the works they have ingested,[11] and to creative derivative works such as summaries, story sequels, and other content based on the original works.[12] We are seeing chatbots of famous authors, such as the Dan Brown chatbot, which simulates "the writing style and flair of the world-renowned author," and for $4.99, allows users "to explore the secrets of the Illuminati, decipher hidden codes, and

[11] For instance, University of Chicago computer science professor Ben Zhao was able to prompt ChatGPT to reproduce the lines from "Harry Potter" in sequence, suggesting that the process could be used to reproduce the entire book. https://twitter.com/ravenben/status/1670667820470870017

[12] The plaintiffs in a recent lawsuit against OpenAI allege among other things that ChatGPT "summaries" of their works infringe their derivative work rights. Silverman v. OpenAI, Inc., 4:23-cv-03416, (N.D. Cal.) ("When ChatGPT was prompted to summarize books written by each of the Plaintiffs, it generated very accurate summaries. These summaries are attached as Exhibit B. The summaries get some details wrong. This is expected, since a large language model mixes together expressive material derived from many sources. Still, the rest of the summaries are accurate, which means that ChatGPT retains knowledge of particular works in the training dataset and is able to output similar textual content. At no point did ChatGPT reproduce any of the copyright management information Plaintiffs included with their published works.")

discover the mysteries that lie hidden within the pages of Dan Brown's novels."[13] We are also seeing AI generated books that are remarkably similar to upcoming pre-order books; [14] and scammers using AI-generated books to exploit "page read" payouts from Amazon.[15]

Moreover, companies are already replacing copy, marketing and web writers with AI, and news publications are using AI to write articles—albeit of low quality for now, but that too could change. Our members have reported losing copywriting work and clients due to companies switching to generative AI, and that they are being asked to edit or develop AI-generated drafts in lieu of their writing services. One member who wrote business-to-business marketing and web content for corporate clients reported losing 75% of their work. These experiences track news reporting on how companies are overwhelmingly switching to generative AI to produce marketing and other content.[16]

The same trend can be seen in web content, where online publishers are increasingly abandoning human contributions for AI-generated articles sometimes published under the bylines of made-up authors.[17] We are also starting to see well-known online publications experiment

---

[13] Dan Brown Chatbot, https://socialdraft.com/products/dan-brown-chatbot (last accessed, July 13, 2023)

[14] See for e.g., Will Oremus, The Washington Post, May 5, 2023, available at https://www.washingtonpost.com/technology/2023/05/05/ai-spam-websites-books-chatgpt/ ("The book's publisher, a Mumbai-based education technology firm called inKstall, listed dozens of books on Amazon on similarly technical topics, each with a different author, an unusual set of disclaimers and matching five-star Amazon reviews from the same handful of India-based reviewers.")

[15] In one recent egregious case, book scammers were able to exploit Amazon's content integrity algorithms to upload a large volume of books into Kindle Unlimited in order to generate revenue from "page reads." While Amazon detected and quickly removed the books, it wasn't until they'd come to dominate the top-100 bestsellers in the book category. See, Jules Roscoe, AI-Generated Books of Nonsense Are All Over Amazon's Bestseller Lists, VICE, June 28, 2023, available at: https://www.vice.com/en/article/v7b774/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists

[16] Oremus, supra note 3 ("Semrush, a leading digital marketing firm, recently surveyed its customers about their use of automated tools. Of the 894 who responded, 761 said they've at least experimented with some form of generative AI to produce online content, while 370 said they now use it to help generate most if not all of their new content, according to Semrush Chief Strategy Officer Eugene Levin.")

[17] Id. (discussing how Ingenio, an online publisher behind popular sites horoscope.com and astrology.com is using generative AI to quickly and dramatically scale up its content offerings). The article further notes: "In a separate report this week, the news credibility rating company NewsGuard identified 49 news websites across seven

with using generative AI. Earlier this month, popular tech news site Gizmodo—which is owned by GO Media along with other popular online publications Deadspin, The Root, Jezebel and The Onion—stirred controversy for an error-filled AI-generated article it published soon after announcing to staff its plans to use generative AI to produce content, and provoked a strong rebuke from the staff union and a statement from Writers Guild of America (East) demanding "an immediate end of AI-generated articles" on the company's properties.[18]

In the context of book publishing, the speed and negligible costs of using generative AI systems to produce books that compete with human-authored books they are based on could inundate the markets, particularly for the types of popular books that are easier to write with AI, such as genre fiction, self-help, and children's books. Books in these categories often make larger profits, allowing publishers and writers to invest in serious nonfiction and literary fiction, the types of books that present new ideas and create discourse and the exchange of ideas so crucial for democracy. Without much-needed guardrails, publishers will feel the need to compete in a race to the bottom to use AI to generate works, and rely on it for narration in audiobooks, creating translations, and to perform other tasks in book production.[19] Research, reporting and the kind of writing that involves critical thinking will become rare and all of us will lose out.

---

languages that appeared to be mostly or entirely AI-generated. The sites sport names like Biz Breaking News, Market News Reports, and bestbudgetUSA.com; some employ fake author profiles and publish hundreds of articles a day, the company said. Some of the news stories are fabricated, but many are simply AI-crafted summaries of real stories trending on other outlets."

[18] Erik Pedersen, WGA East "Demands Immediate End" To AI-Generated Articles On G/O Media Sites, Deadline.com, July 12, 2023, *available at* https://deadline.com/2023/07/writers-guild-demands-no-ai-stories-g-o-media-sites-1235435700/ (last accessed July 17, 2023).

[19] The Authors Guild is attempting to mitigate these scenarios through new contract clauses that give authors control over use of generative AI technologies in the context of their works, including a right to prevent publishers from licensing their works for training purposes. *AG Introduces New Publishing Agreement Clauses Concerning AI,* https://authorsguild.org/news/ag-introduces-new-publishing-agreement-clauses-concerning-ai/

Copyright Incentives Will Suffer if Guardrails are not Placed Around Generative AI Development and Use

The damage to the copyright incentives that will ensue if the law and practice are not reformed soon is not just a problem for writers and other creators—it is not just about job retraining—but a crisis we all face as a society. Human authorship simply cannot be effectively replaced by generative AI, which is inherently limited. AI-generated text and other content is always derivative in the sense that it is wholly based on what has come before, rehashed, and remixed from what the AI has been fed. AI does not think, add reflection or analysis, emotion, or any new context. Only humans can produce real literature and meaningful writing that reflects the ethos of our times.

Copyright is the "engine of free expression" because it creates a market economy for creative work that allows human writers and other creators to write what they think, to represent the world as they see it, not to write or paint pictures that corporate interests, government or patrons want. It is only by giving the individual the right to free expression in a market economy that we get new ideas that lead us forward, help us meet challenges, and resolve our differences.

As such, the tremendous benefits of these technologies, be it to writers or the public, will pale in comparison to the grave harm that generative AI can cause to the writing and other creative professions, and to the copyright incentives themselves. To prevent this from happening, and to ensure that a sufficient and diverse number of writers and other creators can still make a living practicing the crafts they have spent most of their lives honing, we need appropriate robust rules around the development and use generative AI technologies. The cost of inaction to our culture and democratic institutions is far too great.

In sum, the challenges that we see are two-fold:

- Authors need to be compensated for the use of their work by AI or their incomes will be whittled down to the point where professional writers would not be able to write for a living, and we will be left with nothing but hobbyists.
- The market for books, journalism, and other literary and text works needs to be protected from being flooded with AI-generated material so that it remains economically viable for authors.

Authors Guild's Principles and Proposals

The Authors Guild recommends that the following core principles be adopted into law to mitigate the threats to the writing profession:

1. **Consent**

AI companies should be required to obtain permission for the use of writers' works in generative AI. Because writers are so numerous, it appears that some form of collective licensing is necessary as a practical matter. We envision voluntary, free-market collective licensing where permissions and rates are negotiated between a collective management organization and the AI companies. Congress should consider clarifying that AI "training" is an exercise of the author's exclusive right to reproduce their work or create a sui generis right.

2. **Compensation**

Authors should have the right to be compensated for the use of their works in generative AI. For past use, AI developers need to pay for all works of professional creators that they used. Going forward, AI developers should be expressly required to seek permission from the rightsholders. Where the rightsholders are a multitude of individual creators or small businesses, Congress should consider enabling voluntary, free-market collective licensing solutions whereby

a collective management organization (CMO) would license out rights on behalf of the authors (and other creators and small businesses) on a non-exclusive and industry-by-industry basis. Larger corporate copyright owners would be welcome to join as well but could also engage in direct licensing should they prefer. The CMO would negotiate fees with the various AI companies and then distribute those payments to copyright owners who have registered with the CMO, setting aside funds for those who have not yet registered but might in the future. These licenses could cover past uses of copyrighted works in AI systems, as well as future uses.

Copyright owners who authorize the CMO to license their works would receive a distribution based on algorithms that would take various factors into account, such as the number of works published, the type or length of those works (the criteria would vary by industry), and possibly any available sales data. The board of the CMO would be responsible for authorizing distributions and the board or membership (which includes all other rightsowners who sign up) would approve the factors for allocation. A certain amount would be set aside for those who have not yet registered but do so later. The amount set aside would be calculated based on the estimated number of eligible creators and the number who have signed onto the license.

For the past training of what are called the "foundational" models – like OpenAI's GPT, Google's BERT, and Meta's LLaMa – the CMO would seek compensation on an annual basis going forward for as long as those foundational models are in use. Training AI on pre-existing works is not a one-time event – the works are constantly being used to continue to train the AI. As such, payment should continue as long as the AI model is in use.

Antitrust law is currently a barrier to forming CMOs that set rates on behalf of their members. Without some form of recognized exception from the antitrust laws, both the

rightsholders and users who enter into negotiations collectively are potentially exposed to antitrust lawsuits. Legislation may be necessary to avoid U.S. antitrust law violations arising from collective negotiations and agreement on terms.

3. **Transparency**

AI developers should be required to disclose what works they use to "train" their AI. Writers and other copyright owners deserve to know if and how their works have been used in AI systems. Apart from copyright considerations, robust disclosure requirements are also necessary to ensure safety of the models, and prevent use of sensitive, harmful, or illegally harvested data in the training. This requirement will also further encourage AI developers to work with copyright owners to license works for AI uses, instead of relying on datasets created with pirated copies of the works, as has been the case so far.

4. **Use in outputs**

Outputs that copy the style of a particular author, or that reproduce or are derivative works of an author's work, should not be permitted without the author's consent and if permission is provided, the author should be fairly compensated. These outputs, while clearly taken from a particular human creator, may not rise to copyright infringement under current U.S. copyright law, which requires that the expression in an infringing work be "substantially similar" to that of the original work. When these outputs are sold in the marketplace in competition with an author's or artist's own work, however, they harm the market for the original work, amounting to an uncompensated taking of the author's or artist's expression and raising issues of authenticity and unfair competition. A well-articulated federal right of publicity law would be helpful in that it would provide a cause of action when author's names or other indicia of their

identities are used but would not necessarily protect use of a creator's style – which can now be easily and perfectly mimicked without the addition of another author's voice. Congress should consider adding a new economic right that protects clearly identifiable styles of creators, whether under right of publicity, copyright, or a sui generis law. In addition to licensing works for training AI, CMOs could also license rights, collect, and distribute the fees on behalf of the writers who wish to permit the use of their works, styles, or other indicia of personality in outputs, and be compensated, providing the writers means for earning additional income.

5. **Label AI-generated content**

Anyone who makes AI-generated content available to the public, including authors, publishers, platforms, and marketplaces, should be required to label it as AI-generated, or to otherwise identify when a significant portion of a written work has been generated by AI. This will provide transparency to consumers and avoid their being misled into purchasing AI generated content (generally of much lower quality) when they mean to purchase human creative work.

6. **No copyright for AI - generated outputs**

Pursuant to U.S. case law and Copyright Office policy, the copyright law has long been understood to protect only human authorship. This means that material generated using AI should not receive copyright protection unless and to the extent there's also sufficient, demonstrable original human expression in the work; and any AI-generated elements or portions of the work should be excluded from copyright protection.

We oppose efforts to deem AI generated content protectible under copyright, or the creation of a limited "sui generis" right to protect AI-generated material. If AI-generated works

were entitled to the same protection as human-created works, it would incentivize the use of AI to generate content that mimics human-authored works in place of hiring human creators, and it would give AI outputs artificial leverage in the marketplace, inevitably crowding and diluting the marketplace to the point that copyright incentives no longer function as intended. Few human creators will be able to earn enough to sustain a profession, and the human quality of work produced by professionals—those who have talent and have trained in their careers for many years—will disappear.

The Guild's full policy recommendations are attached as Appendix A.

Final Remarks

In the last fifteen or so years, writers have faced unprecedented hurdles to their ability to earn from their craft. Driven by monopsony conditions in the market for books and journalism, authors have experienced a 40% decline in income. The median writing-related income for full-time writers for 2022 was a mere $23,330 according to the Authors Guild most recent income survey with over 5700 respondents.[20]

When we factor in generative AI to the mix of technological and economic disruptions on the markets from which writers derive income, their professional outlook quickly goes from dire to devastating. It won't take much more loss of income to break the proverbial camel's back. These technologies would not exist without the hundreds of thousands of books, and millions of other copyrighted textual works, obtained in many cases from pirate websites and sources. It is grand theft in the extreme and should not be permitted without the permission of the copyright

[20] Results to be published imminently at www.authorsguild.org.

holder. It is only fair that these creators and those who have invested in them be paid for the use of their work in AI systems.

In conclusion, we respectfully ask Congress to consider our proposals or others to ensure that the copyright incentives can be retained for human authors well into the age of AI and as such to preserve the economy for human-created literature and arts.

Appendix A

Authors Guild Policy Proposals Regarding the Development and Use of Generative AI

The Authors Guild believes that it is crucial for our culture and the future of democracy to ensure that our human-created literature and arts remain vibrant and diverse. Generative artificial intelligence—computer programming that can develop new content from huge volumes of existing material—is about to have a significant impact on human creators and the future of our arts. It is imperative that we approach this issue with a full understanding of the impact AI technologies of today and tomorrow will have on the writing profession and the arts more broadly, and with respect for human creators and copyright.

To protect the future of journalism, literature, and the arts, we must develop sensible policies and regulations governing the development and use of generative AI. Three specific issues are paramount, and the need to deal with them is urgent.

Issue 1: Generative AI Uses Human-Authored Works to Mimic those Works Without Consent, Compensation, and Credit

Developers of artificial intelligences like GPT have copied millions of copyrighted works from the internet or illegally compiled databases without permission, relying on claims of fair use under U.S. copyright law to do so. The works are not only copied man times in the course of compiling the databases and the continuous training but are embedded in very the fabric of the AI programs.

This is a very different scenario from that presented in prior court cases such as *Authors Guild v. Google* where the use claimed by Google was to create a database to make books searchable (and only snippets readable). Here, many millions of copyrighted works – almost the entire corpus of human creative works that are available online, including from some pirate website – have been copied to create material that competes in the market with the copied works. These Generative AI technologies would not exist if not for this taking; indeed, the copyrighted creative works are part of the fabric of the workings of generative AI, and they are intended to be used to mimic and regurgitate the language, stories, style, and expression copied. It is grand theft in the extreme and should not be permitted without the permission of the copyright holder.
Reproducing copyrighted works to "train" generative AI should not be considered "fair use." However, some recent fair use cases have adopted an exaggerated version of the test laid out by the Supreme Court in *Campbell v. Acuff Rose* (510 U.S. 569, 1994) that favors finding fair use for background copying where the use is deemed "transformative," and the output is non-infringing. While the Supreme Court's recent decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (598 U.S. ___, 2023) reminds lower courts to consider the degree of transformation against the commercialism of the use, as well as the other factors, the fair use doctrine is still ill-suited for mass use cases such as here where damage to the value of each particular work might be hard to establish. As such, we are asking Congress to help prevent the

evisceration of the creative professions before this egregious, massive taking becomes de facto, uncompensated, and uncredited.

It is worth noting that in some cases, generative AI developers have privately licensed some specific training material, an indication of its value. But most writers and other creators have been left out in the cold, their work expropriated without limit. It is only fair that these creators and those who have invested in them be paid for the use of their work in AI systems.

Proposal: AI developers should obtain permission for the use of copyrighted works in generative AI; this can be achieved though collective licensing in the marketplace

Free Market Collective Licenses Need to be Established Where Necessary

It is not efficient to require AI developers to track down hundreds of thousands if not millions of individual creators to obtain licenses, and they have demonstrated an unwillingness to do so. Large-scale licensing schemes will need to be created so that AI developers can license the rights of individual creators in bulk from a single or several entities. Collective licensing is an established concept and an appropriate one for marketplaces where individual creators (rather than corporate copyright owners) often retain the copyright, and so are numerous and may be hard to locate and obtain license from. We recommend establishing free market collective licenses rather than statutory ones where rates are set by the government.

Free market based collective licenses, are already available in the U.S. by voluntary agreement in many forms in the fields of musical compositions and sound recordings; and they are widely used in Europe and much of the rest of the world for most creative industries, including books and journalism. Collective licenses are an effective means of paying creators and publishers where obtaining a license from every individual creator and publisher creates market inefficiencies. For instance, in much of Europe, libraries and universities pay into collective licenses for photocopying and other specified uses. In the U.S., the Artists Rights Society licenses artwork for various uses; and for many years now, the Authors Registry, the Authors Coalition of America, and more recently the American Society for Collective Rights Licensing (ASCRL), have distributed royalties received from foreign collective licenses to U.S. authors.

Licensing human creation for AI training will not solve all the issues that AI will present to writers and other creative professionals, but it will put some money back into the pockets of creators and their distributors and at least partially compensate them for their efforts, so that many might remain in the creative professions. It is a step in the right direction towards respecting human creativity.

How Collective Licensing Would Work

Collective licensing organizations could be established, or existing ones could be expanded, to offer bulk licenses for certain kinds of works (e.g., text, images) to AI developers to allow those works to be used for training generative AI systems. Each collecting organization would collect and distribute the fees to participating copyright owners. What works are licensed and who qualifies to receive distributions, as well as formulas for who gets what percentage, can be

worked out by the collecting organization and its members. For example, a collective license could be created for the use of a database of books to train AI. The copyright owners who participate would share in the revenue collected from developers who use books to train AI. A mechanism would need to be developed to determine how to apportion payments. This is already done for many other uses throughout the world.

Proposed Legislation: ***Legislation may be necessary to avoid U.S. antitrust law violations arising from collective negotiations and agreement on terms on behalf of members.*** Without some kind of recognized exception from the antitrust laws, both the rightsholders and users who enter into negotiations collectively could be exposed to antitrust lawsuits. Most collective licenses currently set out in the Copyright Act that allow for directly negotiations include a proviso "notwithstanding the antitrust laws." (See, e.g., §§ 17 U.S.C. Secs. 112, 114, 118.)

Aside from the antitrust issues, legislation is not necessary to form an opt-in, voluntary collective licensing organization that negotiates and enters into licensing arrangements on behalf of its members.

Legislation could also provide for arbitration, or for Copyright Royalty Board (CRB) or another entity to mediate rates and terms, should the copyright owners and AI developers fail to reach an agreement by a certain date or on behalf of parties who were not party to negotiations for an agreement for the voluntary collective license each year.[21]

Extended Collective Licensing

For the mass, indiscriminate training of AI that has already taken place, where the AI companies cannot necessarily even identity all works that the AI was trained on, they need to obtain ex post facto blanket licenses that would cover all of the works – and pay for them. Extended collective licensing (ECL) would assist with this. It is a type of collective rights licensing where qualifying collective management organizations (CMOs) can negotiate licenses in the marketplace for a specific type of use on behalf of a specific class of copyright owners, whether or not they are

[21]The Copyright Act contains several provisions that permit negotiations between owners and users of a certain class of works, "notwithstanding any provision of the antitrust laws," and if the parties fail to reach a voluntary agreement, for the CRB is to conduct proceedings to determine rates and terms. See, e.g., §§ 17 U.S.C. Secs. 116, 118, 119. Other countries use various forms of meditation to assist when there is a failure to reach agreement. See, e.g., Denmark's Consolidated Act on Copyright 2014 (Consolidated Act No. 1144 of October 23rd, 2014), Section 52 ("(1) In the absence of any result of negotiations on the making of agreements as mentioned in section 13(1), section 14, section 16 b, section 17(4), section 24 a and section 30 a, each party may demand mediation. (2) Demands for mediation shall be addressed to the Minister for Culture. The request may be made if one of the parties has broken off the negotiations or rejected a request for negotiations, or if the negotiations do not appear to lead to any result. (3) The mediation shall be made by a mediator to be appointed by the Minister for Culture. The mediation negotiations shall be based on the parties' proposal for a solution, if any. The mediator may propose to the parties to have the dispute settled by arbitration and may participate in the appointment of arbitrators. (4) The mediator may make proposals for the solution of the dispute and may demand that such a proposal be submitted to the competent bodies of the parties for adoption or rejection within a time-limit fixed by the mediator. The mediator shall notify the Minister for Culture of the outcome of the mediation..."), located at **https://www.wipo.int/wipolex/en/text/546839.**

existing members of the organization, but there must be an effective mechanism for non-members to opt out of the licenses. Legislation is necessary to authorize these types of licenses (which otherwise would have to be on an opt-in basis) and are intended for mass use, where users cannot negotiate directly with all individual copyright holders due to their sheer numbers. This is particularly appropriate for uses where many of the rightsowners are individual creators and they are numerous. The ECL legislation authorizes collective management organizations that meet certain criteria (and an agency may be appointed to formally approve the organizations) to negotiate blanket licenses on behalf of the entire class on an opt-out basis.[22]

ECLs for AI training could be subject to authorization by the U.S. Copyright Office. The CMO would be required to show that it represents a broad group of impacted rightsholders, that its membership consents to an ECL, and that it adheres to sufficient standards of transparency, accountability, and good governance. Once authorized, a CMO would be entitled to negotiate royalty rates and terms with AI developers on behalf of the class.

In 2011, the Copyright Office looked at the potential for ECLs for mass digitization and issued a Notice of Intent to obtain public comment on the proposal.[23] The Office concluded that it was premature to create ECLs in 2017 due to a lack of interest among certain stakeholders who did not see a need at the time. New AI technologies have changed that perspective, however, and many individual copyright owners are interested in such solutions today –as their works are already being used with impunity to train generative AI to produce material that competes with human creation.

Proposed Legislation: ***Legislation would be necessary to authorize ECLs and to permit an opt-out regime rather than opt-in. The ECLs should be authorized to provide licenses for use of particular classes of works in generative AI on an industry-by-industry basis, where the right holders are particularly numerous, such as where individual creators hold the right and, as such, it is particularly inefficient to provide licenses on an opt-in basis (i.e., where each rights holder must be identified, contacted, and permission obtained on an individual basis). Such an ECL could apply to past use only provided that the law is clear that permission need be obtained going forward.***

The Copyright Office would be given the authority to authorize organizations who represent a substantial number of the rightsholders of a particular class of works (e.g., literary works) to enter into agreements with the users—the generative AI companies—for the extended collective license. Rightsholders in those classes would have the right and ability to opt out of such licenses. The Copyright Office would issue regulations to ensure that robust notice was provided

[22] For examples of ECL's, see Swedish Copyright Act, Chapter 3a: Lag (1960:729), English translation located at https://www.wipo.int/wipolex/en/text/532409; Denmark's Consolidate Act on Copyright 2014 (Consolidated Act No. 1144 of October 23rd, 2014), Section 50-52 , English translation **located at https://www.wipo.int/wipolex/en/text/546839**; DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, Article 12; **located at https://eur-lex.europa.eu/eli/dir/2019/790/oj**.

[23] **https://www.copyright.gov/policy/massdigitization/**

to the covered class of authors of their right to opt-out and that the procedures for doing so are simple and readily available.

Issue 2: Generative AI Has Been Trained on Human-Authored Works to Mimic those Works Without Consent, Compensation, and Credit

Many generative AI systems can be prompted to produce outputs similar to other works or in the style of a certain author or artist or to allow a particular author's or artist's works to be incorporated into outputs. These outputs, while clearly taken from a particular human creator, may not rise to copyright infringement under current U.S. copyright law, which requires that the expression in an infringing work be "substantially similar" to that of the original work. When these outputs are sold in the marketplace in competition with an author's or artist's own work, however, they harm the market for the original work, amounting to an uncompensated taking of the author's or artist's expression and raising issues of authenticity and unfair competition. The right of publicity and unfair competition laws can assist in these cases but will not always apply to or redress this kind of unfair taking.

Proposal: AI developers should require permission and pay compensation for "in the style of" works

Where human-authored works are incorporated in any given AI output, permission should be obtained from the human creator and compensation paid, if desired. We understand that technologies that would enable tracking the use of data from input to a particular output are not yet fully developed but are feasible. It should also be feasible to track and pay creators when their names or the titles of their works are used in prompts.

Proposed Legislation: ***Legislation that requires permission for substantial uses of works in AI outputs would incentivize the development and implementation of such technologies, and it would enable copyright owners to be paid for the use of their works in AI generated outputs.***

This could be framed as **an additional exclusive right under copyright or as a *sui generis* right.**

A federal right of publicity would also assist in protecting creators' rights to use of their identity in AI outputs.

Issue 3: Market Dilution Due to AI-Generated Works Competing with Human Works

Generative AI can produce works exponentially faster and cheaper than the human-authored works they are based on; humans won't be able to compete with the volume of AI-generated works that flood the market.

Proposal: AI-generated works should be clearly labeled

A labelling-requirement for AI generated material—with enforcement provisions that give it teeth—would protect consumers from being misled into purchasing or consuming AI-generated

content that they assumed was human—created. This would help reduce incentives to dump large quantities of low-quality AI generated content into online and other marketplaces. It would also protect consumers against fake imagery and videos that are passed off as authentic news, furthering Congress' goal of preventing AI-generated disinformation and scams from proliferating media and markets.

Proposed Legislation: ***Legislation could require authors, publishers, platforms, and marketplaces to label AI-generated works and otherwise identify when a significant portion of content has been generated by AI. The legislation would also provide for private causes of action and public enforcement, and could be made part of an omnibus AI law.***

Proposal: AI-generated works should not receive copyright protection

Pursuant to U.S. case law and Copyright Office policy, the copyright law has long been understood to protect only human authorship. This means that material generated using AI should not receive copyright protection unless and to the extent there's also sufficient, demonstrable original human expression in the work; and any AI-generated elements or portions of the work should be excluded from copyright protection.

Nevertheless, some argue that works generated solely by AI should be copyrightable (and patentable) because the Copyright Act does not specify that authorship must be human. For instance, Stephen Thaler sued in the District Court for the District of Columbia to appeal the Copyright Office's refusal to register a visual work that he claimed was generated solely by AI. He brought a similar case against the USPTO for denying a claim of inventorship by AI, and he has filed a petition for a writ of certiorari with the Supreme Court in that case.

If AI-generated works were entitled to the same protection as human-created works, it would incentivize the use of AI to generate content that mimics human-authored works in place of hiring human creators, and it would give AI outputs artificial leverage in the marketplace, inevitably crowding and diluting the marketplace to the point that copyright incentives no longer function as intended. Few human creators will be able to earn enough to sustain a profession, and the human quality of work produced by professionals—those who have talent and have trained in their careers for many years—will disappear.

AI systems do not need incentives to generate new works, nor are AI-generated works original in the sense of "original authorship" required under the Copyright Act. They are merely derivative of the works the AI was trained on, and lack any new meaning or expression, unlike human-created collages and other derivative works. Humans necessarily put some of themselves, their thoughts, emotions, experience, and personalities into the works they create; an original work of authorship must contain that spark of human intellect. AI technologies as known today are not capable of adding such sparks of creativity. They can mimic human creativity, but only by regurgitating what they have been trained on. They do nothing to promote the "Progress of the arts and sciences" – the very basis of copyright law under the Constitution.

Proposed Legislation: ***No legislation is required now. If, however, courts find that AI authorship is copyrightable, legislation will be required to clarify that Congress did not intend for non-human authorship to be included in section 102 of the Copyright Act.***

Proposal: AI companies should be required to disclose training data

AI companies should be required publicly disclose training data of their models to ensure safety of the models, and prevent use of sensitive, harmful, or illegally harvest data in the training. This requirement will also further encourage AI developers to work with copyright owners to license works for AI uses, instead of relying on datasets created with pirated copies of the works, as has been the case so far.

Proposed Legislation: ***Legislation would be necessary to require training data disclosures, and can be made a part of an omnibus AI law.***

In sum:

Generative artificial intelligence raises significant issues for human writers, artists, and other creators—and the public that enjoys the fruits of their labors. These issues must begin to be addressed now and continue to be monitored in the future so that our knowledge, arts, and culture can continue to grow and thrive. The Authors Guild stands ready to provide our expertise in this vital aspect of our cultural heritage.