THE Authors Guild

Before the Library of Congress
Copyright Office

Docket No. 2023–6

Comments of the Authors Guild

Artificial Intelligence and Copyright

October 30, 2023

The Authors Guild thanks the Copyright Office for the opportunity to submit the following comments in response to the Notice of Inquiry published in the Federal Register on August 30, 2023, pertaining to the Office's study of the copyright law and policy issues raised by artificial intelligence ("AI") systems.

The Authors Guild is a national non-profit association of over 14,000 professional, published writers of all genres. 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 our 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.

Introductory Remarks

AI systems capable of generating text (large language models or LLMs) pose a serious risk to the writing profession. The development and deployment of these technologies without adequate guardrails not only threatens our members' ability to earn from their works, but the vitality and diversity of our literary culture. Through these comments, we hope to inform the Office of the exact nature and dimensions of these risks, and suggest recommendations for their mitigation. While we use some terms such as "training" and "learning" that seem to anthropomorphize AI machines, we do so only for clarity and since these terms have become commonplace. That said, we want to be clear in our objection to the false equivalence between human and machine learning. Some proponents of AI attempt to elide over questions of copyright and creativity by suggesting that a machines "read" like humans and that a computer "learning" to write is no different from a human learning to become a writer. This disingenuous comparison ignores how differently LLMs function than the human brain, as well as the talent, dedication, and the lived experiences and personality that human writers put into their work and craft. What's more, such comparisons also ignore the reality that AI systems can produce mimicries of human works *at scale* much faster and cheaper. AI systems generate text by making predictions based on the relationships between words and sentences, whereas the human creative process springs from the mysteries of experience, embodied and imagined. It is this quality of human writing and voice that shapes culture and societies, and the reason why we need to ensure that our laws and policies continue to support and incentivize human writing.

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

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. Writers are using generative AI for brainstorming, research, organizing their work, overcoming writers' block, and thinking through creative problems. But any potential benefits of these technologies to writers and the writing profession are far outweighed by the risks created by their development and use without appropriate guardrails. According to our May 2023 survey, 67% of writers believe generative AI to be a threat to the writing profession, and 70% believe that publishers will begin using AI to generate books in whole or part while replacing human authors.[1]

Already we are seeing that generative AI is being used to generate low-quality ebooks, impersonate authors, and displace human-authored books on marketplaces like Amazon. We are also seeing opportunities for freelance journalism, writing, and writing services dry up due to publications and companies switching to generative AI. The current versions of text-generating AI platforms available to the public can perform some writing tasks well, but they cannot yet on their own produce long-form texts well enough to completely substitute for human-authored books and long articles, especially in the literary genres. However, it is only a matter of time before these technologies become more sophisticated and displace these works as well. Our response to question 2 provides more detail on exactly how these markets are being affected.

In addition to posing an immediate, existential threat to the writing professions, the unregulated use of these technologies also raises troubling questions about copyright incentives, guaranteed by the U.S. Constitution, and the public interest they serve. We need to safeguard the incentives that fuel the creation of a rich and diverse creative culture and markets.

Our comments below are limited to generative AI technologies, and specifically LLMs and the applications and platforms built on them, and the impact of these technologies on literary works.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

The increased use and distribution of AI-generated material raises serious issues for the publishing industry and the writing profession.

[1] https://authorsguild.org/news/ai-survey-90-percent-of-writers-believe-authors-should-be-compensated-for-ai-training-use/

a. Impact on book marketplaces

In the book market, there is a serious risk of market dilution from machine-generated works. Generative AI systems are already being used to produce low-quality ebooks that attempt to compete with authors' works. We have seen examples of AI-generated books dominating Amazon's best-seller lists in certain categories, resulting in fewer copies sold by human authors who write in those categories. We are also seeing so-called publishers using AI to generate books on closely related or very similar topics that pop up for sale almost immediately after the human author's book is listed for pre-order, thus pre-empting the sale of the human authored works. Another entity generates short biographies of individuals with upcoming memoirs, comprised of text that reads as though it is paraphrased from a mashup of Wikipedia entries and web copy. The books appeared on Amazon next to the authors' actual book listings in an attempt to steal sales. There has also been a growth in the number of unauthorized AI-generated "summaries" of books that attempt to steal market share from the actual book by getting placed next to it in Amazon listings. In a recent case, author Jane Friedman found on Amazon "a cache of garbage books" written under her name on subjects she is known for. The books were clearly AI-generated from an AI system trained on her work.

Generative AI is also being used to create unauthorized derivative works, such as a developer who used ChatGPT to write the concluding books in George R.R. Martin's *A Song of Ice and Fire* series,[2] and chatbots, like the Dan Brown Chatbot,[3] which allows users to enter the work of an author's books and "converse with the author." All of these uses take money directly out of authors' and publishers' hands.

While so far, the AI-generated books are very low quality and generally disappoint the reader, they are so cheap and fast to produce, that they will nevertheless flood the markets. Anyone anywhere in the world can post on Kindle and other online book marketplaces, and because it is so easy to generate books using AI, they will do so even to capture a small number of sales. Moreover, the quality will improve quickly, and if AI-generated books are not labelled as such, readers will buy them in certain categories, displacing sales of human written books. Amazon, the dominant online book marketplace by far, has taken down the very low-quality AI-generated books and those that violate the Lanham Act when requested, but soon it will be overwhelmed, and it will become harder to keep up. Amazon does not prohibit AI-generated books from being sold as long as they do not violate any of its terms. Some types of books are easier to generate with AI, such as genre fiction, self-help and children's books, and as better book-generating AI applications become available, those markets undoubtedly will quickly become flooded with AI-generated books, suppressing and making it harder to find human-written books.

The Authors Guild asked Amazon to require that AI-generated books be labeled as such, and on September 7, 2023, they adopted a policy that requires book posters to disclose to Amazon if a book is AI-generated (as opposed to AI-assisted).[4] At this time, Amazon is keeping the

[2] Rebekah Valentine, Someone Used ChatGPT to Finish the Game of Thrones Book Series, IGN, https://www.ign.com/articles/someone-used-chatgpt-to-finish-the-game-of-thrones-book-series
[3] Dan Brown Chatbot, https://socialdraft.com/products/dan-brown-chatbot
[4] Content Guidelines, https://kdp.amazon.com/en_US/help/topic/G200672390#aicontent

information about AI-generated disclosures internal. We hope and expect that the information will be made available to customers to prevent consumers from unwittingly buying AI-generated books but imagine that some of those posting AI-generated books will not be truthful. Amazon also adopted a policy limiting the number of books any "publishers" can post on a particular day.[5] These measures are helpful, but will do little to stem the tidal wave of AI-generated books that will certainly come with the next iteration of LLMs (e.g., GPT-5) and improved writing applications built on GPT (e.g., Jasper and Sudowrite), as well as with the release of improved versions of other major LLMs.

Some believe that literary fiction and nonfiction are safe because AI will never be able to generate books of literary quality (and we agree as explained below); however, it is important to remember that mass-market highly commercial books, including genre fiction, self-help and other types of works that are more easily replaced by AI, help sustain the overall marketplace. As a result, when the markets for mass-market commercial books become flooded with AI-generated books, it will impact the publishers' bottom line and consequently their ability to invest in literary fiction and nonfiction.

b. Impact on market for freelance journalism and professional writing services

Freelance journalists and professional writers of web and marketing content are reporting losing work at an alarming rate as a result of clients switching to AI. An Authors Guild member who writes marketing and web content reported losing 75% of their work, and a content writer featured in a *Washington Post* story about ChatGPT's impact on writers stated that he had lost half of his annual income.[6]

Most book authors sustain their profession with other writing-related work, including freelance journalism and writing web and marketing content. Loss of freelance work that supplements book income will hamper writers from focusing on their book projects. As a result, we will have fewer talented writers who would be able to devote much of their professional lives to writing—meaning fewer great books and less diversity in what gets published.

c. Overall harm to author incomes and diversity in publishing

In the last fifteen or so years, writers have faced unprecedented hurdles in earning a living, and it is not going to take that much loss in income to make it impossible for writers to continue practicing their craft. Between 2009 and 2018, authors' median incomes dropped 42%.[7] Due to

[5] Ella Creamer, Amazon restricts authors from self-publishing more than three books a day after AI concerns, The Guardian (Sept. 20, 2023), available at https://www.theguardian.com/books/2023/sep/20/amazon-restricts-authors-from-self-publishing-more-than-three-books-a-day-after-ai-concerns

[6] Pranshu Verma and Gerritt de Vynck, ChatGPT took their jobs. Now they walk dogs and fix air conditioners, Washington Post (June 2, 2023), available at https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/

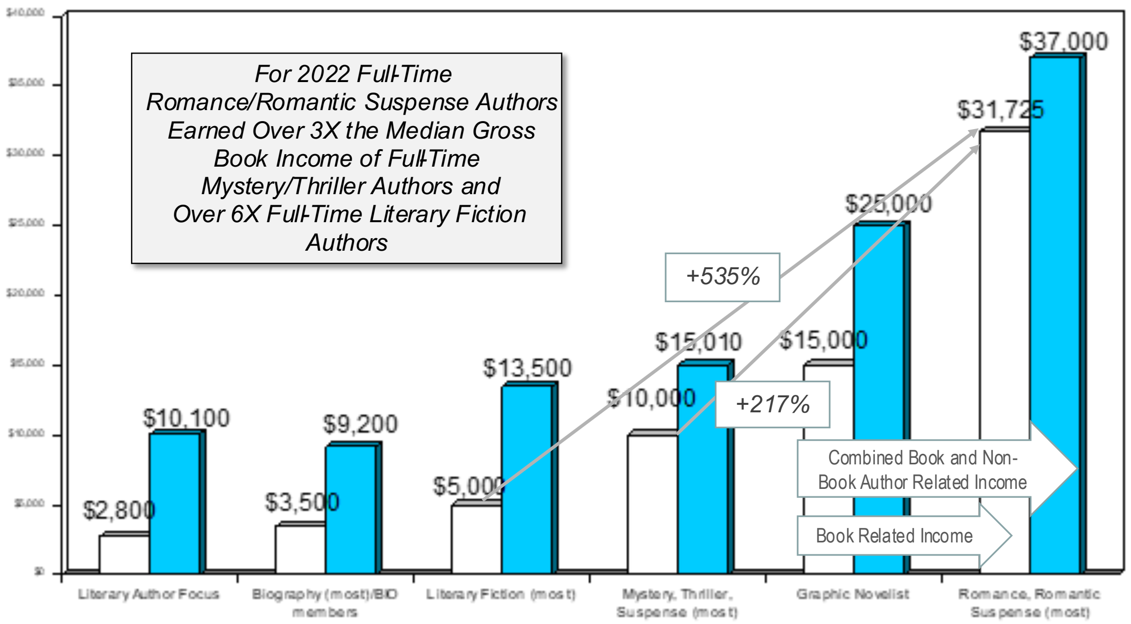
[7] Authors Guild Survey Shows Drastic 42 Percent Decline in Authors Earnings in Last Decade, https://authorsguild.org/news/authors-guild-survey-shows-drastic-42-percent-decline-in-authors-earnings-in-last-decade/

already inefficient market conditions, even marginal dilution of the marketplace with AI books, coupled with the loss of freelance jobs, could be potentially devastating. Our most recent authors' earnings survey found that the median writing-related income for *full-time authors* in 2022 was just over $20,000, with only half of that from books.[8] The median income for commercial (non-academic) traditionally published full-time authors from all writing related activities—which includes speaking, journalism, teaching, editing, and coaching—is $25,000 and only $12,000 from books. Looking at all authors, including those who reported writing part-time, the median book income was $2000 in 2022, and the median income from books plus other writing related work was $5000. Literary authors, even those who write full-time, earned even less from their books. Full-time literary fiction writers earned a median of $13,500 from all writing-related activities and just $5000 from their books. The survey found that writers must engage in other work to earn sufficient money to stay in the profession. That means that many are dependent on income from journalism and content writing that we already see drying up. When we factor in the additional income losses that generative AI will add to the mix of technological and economic disruptions in the markets from which writers derive income, the ability of most authors to stay in the profession quickly goes from dire to devastating.

The chart below shows book and total writing related median Incomes for full-time authors in 2022 by certain genres.

[8] Key Takeaways from the Authors Guild's 2023 Author Income Survey, https://authorsguild.org/news/key-takeaways-from-2023-author-income-survey/

Median Income 2022: Gross Book + NonBook Related Income* Category Published Most- Full-Time Authors



Base 422 118 142 299 41** 373

Of note – Romance/Romantic Suspense authors had the highest % of book related earnings v. combined book + non-book author related earnings (86%) suggesting a far greater author focus on book results.

* Excludes non-zero earning authors **CAUTION: very low base

The Authors Guild © 2023

4

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

Because generative AI is still a new technology, there haven't been any studies to assess the economic effects of generative AI on writers or publishers. We imagine it will be hard to measure the losses on book income as a direct effect of generative AI for a couple years. We hope to complete a survey sometime next year to get updated figures on any losses in journalism, web content, business, copy and other writing.

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?

Efforts are currently underway in the EU, the UK, and other countries to develop frameworks around generative AI, but so far, no legislative or regulatory approach has been fully implemented. International consistency in AI governance will be extremely important to prevent forum-shopping and off-shoring of AI-related activities prohibited under one country's copyright laws to more favorable jurisdictions.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

Yes, we believe that generative AI technologies can be used to exploit the value of copyrighted works in entirely new ways and that the copyright law must cover those forms of exploitation if it is to remain meaningful and continue to fuel robust and diverse ecosystems for human creativity.

1. *Use of copyrighted works to train generative AI must licensed*

First, generative AI copies and ingests massive quantities of preexisting works and it is completely dependent on those preexisting works. It would not exist without them. A great deal of the value of any large language model (LLM) derives from those pre-existing works. The major text generating LLMs, such as Open AI's GPT, Google's LAMDA, and Meta's Llama, and Anthropic's Claude are "trained" on hundreds of thousands of books and millions of copyrighted articles, and websites. So far, they have mainly done this without credit, compensation, or consent, arguing that it is a fair use of the copyrighted works.

The LLM knows nothing and would not be able to generate coherent text without the works that it has been "trained" on; and the quality of an LLM depends greatly on what works it has been trained on. The ability of LLMs to generate text that is grammatically and stylistically well written, and acceptable to todays' readers (linguistically and also in terms of biases and concepts) is dependent on having ingested relatively recent high-quality texts—meaning books, stories, articles, and blogs, often those with the highest value in the marketplace. The phrase "garbage in, garbage out" refers to this phenomenon, and so if AI developers want their LLMs to have good outputs, they need good inputs. From the early days of LLM development, AI researchers and developers understood that they needed books to train LLMs to have decent outputs because books generally have well written, thought out and articulated texts and information.

The leading LLMs today were "trained" on datasets of approximately 200,000 books, plus datasets that include millions of news articles and books scraped from the internet. All the book datasets that we are aware of were copied from pirate ebook sites, and almost all other books that were scraped up by the creators of the datasets compiled from web crawls (e.g., Common Crawl, which has 60 billion domains scraped over 12 years) were also copied from pirate sites.

Open AI has kept what its current GPT models, GPT-3.5 or 4, were trained on completely secret, except that it is generally known that the amount and quality of its so-called "training data" (what we refer to as copyrighted texts) used is vastly greater and superior to that of prior models.

When Open AI released an earlier model, GPT-3, in 2020, they did release some high-level information on what it used to train it:[9]

GPT-3 training data[1]:9

Dataset	# tokens	Proportion within training
Common Crawl 60 billion domains over 12 years Includes BBC, *The New York Times*, Reddit, the full text of online books, and more	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Books1 is known to be the “Books Corpus” data set that includes 11,038 fiction books (around 74M sentences and 1G words) of 16 different sub-genres scraped from indie publishing site Smashwords.[10] But Open AI never revealed the composition of the Books2 dataset, which alienated the sharing and open access culture in which most generative AI researchers resided, and which Open AI had come out of. As a result, an independent AI researcher, Shawn Presser, decided to create something similar to Open AI’s Books2 for use by open-source developers; he did it by downloading around 200,0000 books from a pirate torrent tracker called Books 3 and named the resulting dataset “Books3” as a play on Open AI’s references to Books1 and Books2.[11]

Other companies have been more transparent about their training data, and many have used the Pile, which includes the Books3 dataset. The Pile is a conglomerate dataset composed of 22 smaller datasets that a group of independent researchers (who later became the not-for-profit EleutherAI) compiled and released on January 1, 2021.[12] Meta AI researchers and developers revealed that they used the Pile, including Books3, to train Meta’s LLaMA.[13] Researchers have also traced use of pirate books in datasets used by Google, including book text from B-ook.cc, a

[9] GPT-3, https://en.wikipedia.org/wiki/GPT-3

[10] “Book Corpus” was compiled by University of Toronto researcher in an early machine learning experiment to align book text with movie screenshots. See Zhu, et al. https://arxiv.org/abs/1506.06724

[11] Kate Knibbs, The Battle Over Books3 Could Change AI Forever, Wired (Sept. 4, 2023, https://www.wired.com/story/battle-over-books3/

[12] Datasheet for The Pile, https://arxiv.org/pdf/2201.07311.pdf

[13] Touvron, et al, LLaMA: Open and Efficient Foundation Language Models, https://arxiv.org/abs/2302.13971

mirror site for the major pirate site Z-Library, which was taken down by the U.S. government (with assistance from the Authors Guild) for criminal copyright piracy.[14]

Why have LLM developers relied on pirate copies of ebooks? Because the only places to get a trove of ebook texts without permission from the copyright owners (or licensing Google's Google Books collection) are from pirate websites. The developers understood that they needed large numbers of books, but they did not want to pay to license them. They easily could have gone to publishers and authors to request licenses and paid for the use but chose not to. One important thing to note about ebook pirate sites is that the books they contain tend to be the most commercial ones—the ones that readers are likely to be looking for. So, it is precisely the books that authors and publishers rely on to earn money that end up getting pirated and then ingested into LLMs.

This practice must be brought to an end. A good deal of the value of any major general LLM is derived from books, and those same LLMs are being used to put the writers out of work. Not only is it self-evidently unfair, but if unrestrained, will put many human writers out of work and devalue our literature and culture.

As the Copyright Office is aware, several copyright infringement lawsuits have already been brought against LLMs that include claims of infringement for the ingest of books and other texts.[15] The cases seek to clarify the applicability of the fair use doctrine to use of copyrighted works in training generative AI cases, and the outcome of the cases will test the flexibility of existing copyright law to address the issues raised by generative AI while ensuring that the vital incentives for human authorship remain.

Each of the cases against LLMs have been brought on a class action basis on behalf of authors (because individual authors, even in groups, usually cannot afford to bring these suits as they are likely to be expensive), which means it could take some time before decisions are reached on the merits. **If the courts find fair use or leave openings for the unauthorized use of commercial literary texts for AI training purposes, we will need legislation in order to preserve the copyright incentives for writers.** The Authors Guild hopes that we will not have to see too great a loss of human-written literary culture before authors and publishers receive the necessary protections.

2. *Individual creators need the ability to license collectively*

Regulatory and/or statutory intervention is urgently needed to enable creators to negotiate licensing terms with AI developers collectively. AI companies have expressed an unwillingness

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

[15] Authors Guild v. OpenAI Inc., 1:23-cv-08292, (S.D.N.Y., Cmpl. Filed Sept. 19, 2023); Silverman et al. v. OpenAI et al., Case No. 3:23-cv-03416 (N.D. Cal., Cmpl. filed July 7, 2023); Kadrey, Silverman, & Golden v. Meta Platforms, Case No. 3:23-cv-03417 (N.D. Cal., Cmpl. filed July 7, 2023); Tremblay & Awad v. OpenAI et al., Case No. 3:23-cv-03223 (N.D. Cal., Cmpl. filed June 28, 2023); J.L., C.B., K.S., P.M., N.G., R.F., J.D. and G.R., individually, and on behalf of all others similarly situated v. Alphabet Inc., Google Deepmind, and Google LLC, Case No. 23-cv-03440 (N.D. Cal., Cmpl. Filed July 11, 2023))

to acquire rights on an author-by-author basis. We understand that some AI companies have recently approached certain publishers to license their catalogues, but the publishers in many cases do not possess the rights to license out works for AI training or the contract is unclear. Moreover, rights to many out-of-print books (which includes many books that are more than a decade old) have reverted to the authors, and the majority of books published today are self-published where the author retains the rights, and this self-publishing market continues to grow.[16]

This means that to properly license books, as well as many freelance literary texts where the writer retained rights, AI companies will need a means of licensing directly from the authors. Since it is highly inefficient for them to do so on a case-by-case basis, they are unlikely to. If AI training on literary works is to be licensed, then it is imperative that we establish a collective license of some sort—a means of providing blanket or dataset licenses to AI developers. **This may require an exemption from antitrust law for authors and creators to engage in collective licensing**, and is an important part of our legislative proposals, which are attached to our comments. An antitrust exemption will allow authors (and other creators) to negotiate and enter free-market licensing arrangements with AI developers on an industry-by-industry basis. If we cannot obtain an antitrust exemption, then we may need legislative authorization for extended collective licensing to ensure that the use of copyrighted works by AI developers is authorized and compensated.

3. *Generative AI has created a new form of exploitation of authors' works that is not clearly covered by current law—copying recognizable style, voice, or body of work*

LLMs excel at capturing authors' style, voice, or body of work. LLMs' algorithms are trained to generate text by recognizing patterns and relationships in the language they are fed, and it turns out this gives them an uncanny ability to detect and impersonate writing style and voice. When humans write or create, we naturally put our own personality into what we are creating. Every writer has some style or voice, but we don't often think about it except when it comes to writers or publications known for strong and consistent voice (e.g., Ernest Hemingway, the New Yorker). Humans can mimic others' styles and sometimes do——but then they always also unavoidably add their own voice. It takes a lot of training and talent to be able to effectively mirror another's style—and since writing is usually about expressing one's own voice in any event, there is almost no incentive to try to mimic another author's style for profit (except for parodies). It is easier and preferable to write in one's own voice. As such, protecting style or voice has not been a concern or considered an exploitable right. Generative AI changes this reality.

For example, one of our Council members, Mary Bly, a best-selling romance author and Shakespeare scholar, did an experiment with ChatGPT in September. She asked it to write the first three lines of a historical romance novel, which it did. She then asked it to write the text in the style of three different known historical romance authors (including her own pen name Eloisa

[16] See e.g., https://www.publishersweekly.com/pw/by-topic/industry-news/bookselling/article/91298-romance-books-were-hot-in-2022.html
https://www.nytimes.com/2023/08/12/nyregion/ripped-bodice-brooklyn-romance-book-store.html
https://www.npr.org/2023/06/02/1179850128/even-as-overall-book-sales-are-declining-romance-novels-are-on-the-rise

James), as well as her daughter's (not a romance writer). ChatGPT responded with paragraphs that captured each author's style, and for her daughter the response was: "I'm sorry, but I'm not familiar with the writing style of [X], as she may not be a widely recognized author as of my last knowledge update in September 2021. If you could provide more context or details about her writing style, I'd be happy to try to emulate it for you."

Suddenly, we see people using generative AI to generate texts in the style of authors. Authors have likened this to identity theft, and its potential to rob authors of future income is a tremendous concern. Pulitzer Prize winning author and Authors Guild Council member Min Jin Lee recently wrote on X (Twitter): "Al companies stole my work, time, and creativity. They stole my stories. They stole a part of me." We hear this heart-wrenching sentiment over and over from authors, who, after years of developing their unique voice and style, are finding AI appropriating a part of their personality and mimicries of their work being sold in the market.
Bestselling authors in particular are concerned about the fact that now anyone can generate books in their voice that will be sold to compete with their books. One bestselling author mentioned that they were about to retire and are very worried that people will publish future books in their series without permission and those books will compete with their earlier books, leading to a loss of sales. Chatbots today allow anyone today to write a new Lee Child or Dan Brown story, or the "next book" in Suzanne Collins *Hunger Games* series, or the next *Pendergast* adventure from the bestselling series by Doug Preston and Lincoln Child. As noted above, we have already seen someone write the last two novels in George R.R. Martin's *A Song of Ice and Fire* (*Game of Thrones*) series.

A writer's recognizable style or voice is not itself protected under the current copyright laws. Some instances of this type of appropriation may be actionable under copyright, where the AI-generated book is substantially similar to one of the author's works—especially if the AI book uses the author's characters, plots and worlds. But these cases will be rare, and often too murky for an individual author to want to risk litigation. Copying style in and of itself is not a cognizable harm under any other law. Current right of publicity does not cover it—unless we can extend the notion of a person 's identity to their style or voice—nor will unfair competition laws necessarily apply. This made sense in the context of prior technologies, but just as the printing press necessitated protections against copying (initially in the form of monopolies granted to printers over their publications), we will need to find a way to prevent authors' body of work or recognizable style from being exploited by others without permission. Indeed, even CEOs of AI companies, including Sam Altman of Open AI, have echoed the concern about the unfair appropriation of authors' and creators' unique styles and voices.[17]

Creating protections against theft of style or voice without the addition of another author's voice, or against the taking of an overall body of work (as opposed to considering infringement on a work-by-work basis) will not be simple. Any such protections have to be very narrowly defined to prevent imposing on free speech and future expression, but it will be necessary to find a way to protect these attributes of creative works if we are to retain robust copyright incentives.

[17] Transcript: Senate Judiciary Subcommittee Hearing on Oversight of AI, https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/

Enacting true moral rights under U.S. law would help.[18] Other ideas include a federal unfair competition law that would protect readily identifiable style or voice in the way that trade dress or descriptive brand names can acquire protection if they acquire "secondary meaning." Alternatively, a new form of *sui generis* protection could be added to the copyright statute.

We also support the enactment of a well-articulated federal right of publicity law to provide a cause of action when authors' names or other indicia of their identities are used.

Training

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?
6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

AI platforms capable of generating text rely on large language models (LLMs), trained on a vast amount of text data to generate human-like text. The model contains millions or billions of parameters that capture statistical relationships between words and phrases in the training data. This allows the model to predict probable sequences of words and generate coherent, natural-sounding text. Text used in training LLMs is collected and ingested by: (1) web-crawling and of websites; or (2) downloading copies of works in bulk datasets.

LLMs are trained on text datasets ranging from gigabytes to terabytes in size. As described in response to question 2, these datasets include text of books, news corpora, online articles, and journal papers, billions of web pages crawled from across the internet, social media posts from platforms like Reddit, Twitter, and forums, product reviews, instruction manuals, e-mails, and other text content. The largest LLMs are trained on hundreds of billions of words from the web and books and other sources.[19]

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 subsequently was used in training early versions of GPT, Google's BERT, and Amazon's Bort and thirty other models: "[b]ooks provide…very rich, 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

[18] Comments of the Authors Guild, Study on the Moral Rights of Attribution
and Integrity, available at https://authorsguild.org/app/uploads/2017/04/Authors-Guild_Moral-Rights-NOI_3.30.17.pdf

[19] Alan Thompson, What's in my AI? available at https://s10251.pcdn.co/pdf/2022-Alan-D-Thompson-Whats-in-my-AI-Rev-0b.pdf

story)."[20] From the early days of machine learning to the latest groundbreaking LLMs, books have comprised 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 (OpenAI has not disclosed the training data for GPT models beyond GPT3), and 4.5% of Meta's Llama model.

As noted in the response to question 2, many if not most of the major LLM developers, including OpenAI, Google, and Meta, trained their models using datasets of copyright-protected works that were scraped and ingested from all over the internet, including copying of pirated books that appear on rogue websites and by circumventing firewalls on sites in order to access copyrighted material on subscription-based websites. Researchers and users have created training datasets using copyrighted material for LLMs, and there are many such datasets available online. Hugging Face, for instance, provides a platform for users to share datasets. EleutherAI's "Pile dataset" is a compilation of datasets that many LLMs have used.[21]

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

Licensing is commencing slowly in the LLM world, so far at minimal scale, and only recently, following a spate of lawsuits against AI developers. For instance, in July 2023, the Associated Press signed a deal with OpenAI to license content as training data for the latter's AI models.[22] We understand that some STM publishers may also have licensing deals in place, especially for fine tuning special use applications. At this time, there are no known licensing deals between book publishers and AI developers, but we know that some companies have started to reach out to trade (commercial) publishers to request licenses and some trade publishers are considering them. As noted above, however, publishers in most cases will need to obtain consent from individual authors in order to engage in such licensing. Other solutions are also being considered, including one by the Authors Guild, that would create a licensing platform for authors and publishers to license books for training.

Many authors are very concerned about generative AI being used to replace them and so will not jump on opportunities to license their work for training. To encourage authors to license their works, AI companies will need to implement restrictions on end uses or find a way to pay authors for end uses that incorporate an author's work or style, etc. For instance, AI companies could disallow prompts that use an author's name or titles of their works. Chatbots and other applications built on the LLMs could also be prevented from responding to requests to copy style or voice. However, current measures can be circumvented by the wily and willing, and it is inevitable that a cat and mouse game will ensue where the AI companies will always be one step behind those determined to get around filters.

[20] 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
[21] See supra note 7.
[22] Nitasha Tiku, Newspapers want payment for articles used to power ChatGPT, Wash. Post (Oct. 20, 2023), available at https://www.washingtonpost.com/technology/2023/10/20/artificial-intelligence-battle-online-data/

To date, there is no consensus on how much authors and publishers should be compensated. For licensing to take place in an open free market (without statutory licensing), there need to be willing sellers and buyers. That means the negotiated fees must be substantial enough for authors and publishers to find it worthwhile to participate but cannot be unreasonably expensive for AI companies. Because so many books and other works are required to train an LLM at least today, it might make sense to split fees into two payments: a reasonable, negotiated ingest fee that all authors or copyright owners would receive, as well as fees to authors for output uses that that rely substantially on their work.

6.3. To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

Public domain works and books are used in training, but they cannot substitute for high-quality, recent copyrighted works. The exact extent to which public domain works are used depends on the model, and making such a determination depends on whether or not the developer has made training data information available. The Pile, an 800GB publicly available dataset, contains 10.88GB of books from Project Gutenberg (and 100.96GB of pirated books in the Books3 data-subset)[23] and has been used to train several models, including Meta's LlaMa.[24] Datasets used by Google and OpenAI may also contain some public domain books, but it cannot be ascertained due to lack of detailed information.

There are some obvious reasons why public domain books do not provide ideal training material for large language models. Public domain books are typically older works, which means they might not capture contemporary language, slang, or topics, and moreover, they might contain societal norms and values of their time, including outdated or offensive views. They also contain outdated knowledge. Indeed, AI developers like Open AI have admitted that removing copyrighted works from the training datasets would "lead to significant reductions in model quality."[25]

We are aware that some AI developers are hiring writers to write texts to train AI.[26] Others in the AI community have acknowledged that there might be a need for this. There is a phenomenon called "model collapse" that occurs when an LLM is trained on AI-generated content. The outputs quickly start degenerating into nonsense.[27] AI developers understand that they will always need a fresh supply of human written content to keep LLMs functioning and so have started to consider this new form of employment for writers.

[23] https://arxiv.org/pdf/2101.00027.pdf

[24] https://arxiv.org/pdf/2302.13971.pdf

[25] https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf

[26] Why Silicon Valley's biggest AI developers are hiring poets, https://restofworld.org/2023/ai-developers-fiction-poetry-scale-ai-appen/

[27] Shumailov, The Curse of Recursion: Training on Generated Data Makes Models Forget, https://arxiv.org/abs/2305.17493

It would make far more sense, however, to preserve the human writing ecosystem under copyright law so that writers write for the consuming public rather than for AI machines. The former gives us the diverse and rich literary culture we have today that reflects our human experience, emotions, ideas, hopes, and dreams, whereas the latter will lead us to a world of algorithmically synthesized mimicries of books, literature, and art.

In any event, as discussed in our response to question 6, LLMs require a staggering and diverse volume of text content to learn inferences and generate quality responses; even if some of this training data was specially commissioned or created, it would likely only form a marginal portion of the dataset.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

We don't have information about the training protocols of AI companies as to perseveration of training data but given that they used so many works without permission or respect for the creators or copyright owners, we do not expect that they would have established practices for deleting the datasets. Moreover, as noted above, many datasets such as Books3 are available on third party sites, and already have been downloaded by many thousands of users.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in:
7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

Copyright-protected works and other training materials may be reproduced several times during the training process. The first important point here to note is that training LLMs, at this stage, requires ingestions of complete works. Copyright-protected works and other training materials are reproduced when AI developers download them or download archives containing the raw files of the works (as opposed to pre-processed datasets). Another copy of these materials is created when the developers tokenize[28] and process the works into a training dataset, which is essentially a large text file containing the entire text of the works and other training materials. Even if an external third party is doing the pre-processing, a copy of the training materials still occurs within the control of the AI developer when they download the tokenized dataset. As we understand, however, data preparing, processing, and tokenizing typically take place within the control of the AI developer as this process has bearing on the training and performance of the model. Copying of copyright-protected works into training datasets and the datasets themselves are infringements of the exclusive right of reproduction and derivative work rights of the copyright owners. At this stage, training necessitates making fixed copies of copyrighted works, and these copies are not transient or chunked, unlike those that occur during "buffering."

[28] "Tokenizing" is a process where the input text is broken down into smaller units, called tokens. These tokens can be words, parts of words, or even individual characters, depending on the tokenization method used.

7.2. How are inferences gained from the training process stored or represented within an AI model?

We don't have a deep enough understanding of the training process to answer this question. That said, we understand from our research and conversations with AI researchers that the notion that machines only "read" the training materials during training is inaccurate, and that many of the supposed "inferences" about particular copyrighted works are in fact regurgitations of the work from the model's memory.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?

We have heard from some AI developers that unlearning inferences may be possible but has not occurred at scale. One recently introduced "unlearning" technique that replaces idiosyncratic expressions with generic counterparts in response to target tokens has been gaining notice, but we are not aware of its economic feasibility or feasibility for wide-scale adoption.[29]

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Yes, often it is. LLMs can be prompted to respond to specific information about a work and to produce quotations from a work. We cannot say whether this technique works for all training materials, and suspect that it's not useful when attempting to identify generic text training material; but as far as books and other original copyrighted works are concerned, our experiments have shown that LLMs like GPT are able to generate high-fidelity responses to queries specific to particular copyrighted works, including by reproducing exact lines (and successive lines from a particular text), summarizing portions of the text such as chapters in detail, and generating compelling outlines and text for hypothetical sequels of works. Many authors have undertaken similar experiments to interrogate GPT's memorization of their books and found the results to be too starkly accurate to be possible other than through ingestion of the complete work or substantial portions of the work. AI researchers have also conducted experiments into GPT's memorization of copyrighted books, with results indicating that GPT models have memorized the text of popular books such as *Harry Potter and The Sorcerer's Stone, 1984, Fifty Shades of Grey, The Hunger Games, The Da Vinci Code,* and *A Game of Thrones.*[30]

Moreover, prompts that pull from only a few works will often simply paraphrase the original sources. For instance, when we asked ChatGPT to write a fundraising letter for the Authors Guild Foundation as a test, the result was a decent fundraising letter where the fundraising language may (or may not) have been culled from many sources, but the way the Foundation and its work was described was paraphrased directly from our website. This also occurs, for instance,

[29] Who's Harry Potter? Approximate Unlearning in LLMs, https://arxiv.org/pdf/2310.02238.pdf
[30] Speak, Memory An Archaeology of Books Known to ChatGPT/GPT-4, https://arxiv.org/pdf/2305.00118v2.pdf

in outputs to prompts for biographical information about people who have Wikipedia entries but for whom there is not a lot of other biographical information available online or in books: ChatGPT will paraphrase the Wikipedia site and maybe fill in other details, obtained from elsewhere or made up.

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.

Fair use is determined on a case-by-case basis, and therefore any fair use analysis will depend on the particular facts at issue. The ingestion of copyrighted works for purposes of training a generative AI system to mimic those works is unlikely to ever be fair use under current case law. Such activity involves a direct exploitation of the works' creative expression, and thus it is far different from cases in which copying works for purposes other than accessing their expressive content was held to be "transformative." Moreover, it substantially harms the market for the ingested works and their derivatives.

1. *Ingestion for training generative AI systems is unlikely to be transformative*

As discussed below in response to question 8.1, the Supreme Court recently clarified the analysis under the "purpose and character of the use" factor of fair use in *Andy Warhol Foundation for the Visual Arts Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023). In particular, the Court made clear that the "transformativeness" of a secondary use is not dispositive under that factor but must be weighed against other considerations, including commerciality and the justification for the copying. The decision also emphasized that the degree of transformativeness needs to be considered and that transformativeness is not an on-off switch. Copying full verbatim text from copyrighted works for purposes of creating a generative AI system is unlikely to constitute a transformative use but is highly commercial.

AI companies have argued that ingestion for training purposes is analogous to the large-scale copying held to be transformative in cases like *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014), and *AV ex rel. Vanderhye v. iParadigms, LLC*, *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630 (4th Cir. 2009). But those cases all involved uses that did not exploit the works' creative content. For example, in *Google*, the court held that the first factor favored fair use because Google's mass digitization of books to create a full-text searchable database involved a transformative purpose. The purpose of that copying, the court found, was "to make available significant information *about those books*." 804 F.3d at 217. The books' expressive content was immaterial to how well the index functioned as a search tool. Likewise, in *iParadigms*, the student papers were archived not for the quality of their expression, but simply to detect future instances of plagiarism. *See* 562 F.3d at 640. By contrast, AI companies seek out published books for ingestion precisely *because of* their expressive content, as high-quality, professionally authored works are vital to enabling an LLM to produce outputs that mimic human language, story structure, character development, and themes.

Nor is ingestion comparable to the copying held to be transformative in *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). Those cases arose in the distinct context of computer code, and the courts found that the copying was necessary to access the code's functional aspects for purposes of creating interoperable products. *See Sony*, 203 F.3d at 603-04; *Sega*, 977 F.2d at 1527-28. As noted, AI companies do not copy works simply to access uncopyrightable elements like grammar rules or historical facts. Rather, the works' expressive elements are what is needed for the companies to create a more commercially desirable product—one that can generate outputs that compete with the very works used to build the system. No court has found such direct exploitation of copyrighted expression to be transformative.

2. *Ingestion for AI training harms the market for and value of copyrighted works*

The fourth statutory factor is unlikely to favor use where works are ingested to train a generative AI system. Such activity harms the market for the copied works in at least two ways. First, unauthorized copying for AI training undercuts copyright owners' ability to license their works for that purpose. As discussed below, copyright owners in various sectors have already begun developing mechanisms for licensing works for AI training, and established collective licensing systems around the world can provide useful models for that market. Collective licenses have been used in the music industry for decades, for television programming, by the Copyright Clearance Center for text licenses, and by other entities. For over two decades, the Authors Registry and the Authors Coalition of America have distributed royalties received from foreign collective licenses to U.S. authors and visual artists. More recently, the American Society for Collective Rights Licensing (ASCRL) has distributed royalties received from foreign collective licenses to U.S. authors, artists, photographers, and other rights holders. Mass copying by AI companies undermines the efforts already underway to adapt these successful models to allow licensing by copyright owners who wish to make their works available for AI training.

Second, AI-generated outputs that mimic or are otherwise based on the ingested works undermine market demand for those works when they are introduced into the marketplace to compete with the originals. Open AI's ChatGPT is already being used to generate books that mimic human authors' work, as described in our response to question 2. Examples include the recent attempts to generate volumes 6 and 7 of George R.R. Martin's *Game of Thrones* series *A Song of Ice and Fire*, as well as the numerous AI-generated books that have been posted on Amazon that attempt to pass themselves off as human-generated and seek to profit off a human author's hard-earned reputation. The harm will only be compounded as the creation and dissemination of such materials become more 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.") (internal quotation marks, citation, and emphasis omitted).

8.1. In light of the Supreme Court's recent decisions in Google v. Oracle America and Andy Warhol Foundation v. Goldsmith, how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor?

Warhol makes clear that the fair use analysis must be made in reference to the particular use at issue. The "use" is defined according to which exclusive rights of the copyright owner are implicated. 143 S. Ct. at 1277 ("The fair use provision, and the first factor in particular, requires an analysis of the specific 'use' of a copyrighted work that is alleged to be 'an infringement.'" (quoting 17 U.S.C. 107)). The ingestion of copyrighted works in connection with developing an AI platform implicates the exclusive right of reproduction under section 106(1). Therefore, the first factor analysis will look to the purpose and character of those acts of copying.

While fair use is determined on a case-by-case basis, as a general matter the mass copying of copyrighted works for the purpose of developing a commercial generative AI platform is unlikely to satisfy the "purpose and character" factor. Under that factor, courts consider, among other things, whether the secondary use is commercial in nature and whether the user has sufficient justification for the copying. *Id.* at 1276-77. The highly commercial nature of the lucrative public-facing generative AI platforms offered by companies such as OpenAI and Google weighs heavily against fair use. That factor is unlikely to be offset by any purported "need" for the copying. As the Court in *Warhol* made clear, the fact that an author's original expression may be useful to the secondary user is not sufficient justification. *See id.* at 1286 ("Copying might have been helpful to convey a new meaning or message. It often is. But that does not suffice under the first factor. . .. As Judge Leval has explained, '[a] secondary author is not necessarily at liberty to make wholesale takings of the original author's expression merely because of how well the original author's expression would convey the secondary author's different message.'") (citation omitted)). Thus, to the extent an AI company copies works because it believes the quality of their expression will enable it to produce a superior platform in the marketplace, the first factor is unlikely to favor fair use.

Moreover, even if ingestion of works is deemed transformative in some circumstances, *Warhol* makes clear that such a finding is not determinative—either to the fair use analysis as a whole or even to the first factor. The Court emphasized that "'transformativeness' is a matter of degree" and that "the commercial character of a secondary use should be weighed against the extent to which the use is transformative or otherwise justified." *Id.* at 1275, 1279 n.13. Thus, a court would still need to consider the commercial nature of the use, whether the use is likely to provide a substitute for the original work or derivatives of it, and whether there is a legitimate need for the copying.

Google v. Oracle America does not change this analysis. That case arose in the context of computer programs, which, as the Court noted, "differ from books, films, and many other 'literary works' in that such programs almost always serve functional purposes." 141 S. Ct. 1183, 1198 (2021). The Court made clear that the distinctive nature of computer code was highly relevant to the first factor analysis. *See id.* at 1203 ("Google copied portions of the Sun Java API . . . in part for the same reason that Sun created those portions, namely, to enable programmers to call up implementing programs that would accomplish particular tasks. But since virtually any unauthorized use of a copyrighted computer program (say, for teaching or research) would do the same, to stop here would severely limit the scope of fair use in the functional context of computer programs."). The case thus has limited application in the context of works closer to the heart of copyright's protection for creative expression. As the Court later explained in *Warhol*, *Google* "did not hold that any secondary use that is innovative, in some sense, or that a judge or

Justice considers to be creative progress consistent with the constitutional objective of copyright, is thereby transformative." 143 S. Ct. at 1283 n.18.

As to whether different stages of training raise different considerations under the first factor, the analysis should, as always, focus on the purpose and character of the specific use. To the extent that ingestion of works is undertaken solely for noncommercial purposes, the first factor balance may be struck differently than it would be in the context of a commercial AI enterprise. Where, however, pre-training or fine-tuning are conducted for purposes of advancing a broader commercial purpose, the first factor will weigh against fair use. An AI company cannot shield itself from the fair use analysis simply by breaking up the ingestion process into multiple components.

8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

Since these activities involve the making and/or the public distribution of copies, they are infringing unless an exception applies. With respect to fair use, the analysis will rest in large part on how the court weighs the commerciality of the use. Creating datasets from copyrighted materials is not transformative. Training datasets of books for instance simply included all of the unformatted text of the books, and tokenizing the text does not transform it in any way. The court would also look at whether the collection and distribution are done for commercial purposes. An entity in the business of preparing datasets and marketing them to AI companies for training would be engaging in unauthorized commercial use of copyrighted works, such that the first factor is unlikely to favor fair use. Conversely, an entity that prepares datasets purely for research or other nonprofit purposes would have a stronger claim under that factor.

The second factor would weigh against fair use where the works are highly creative and closer to the heart of copyright. As such, training on books for instance would weigh against fair use, whereas perhaps the use of functional and standard code would weigh in favor of fair use. The third factor would almost always weigh against fair use when the entirety of the works is taken. Under the fourth factor, a court would need to consider whether the collection and distribution of copyrighted materials for AI training threatens to usurp copyright owners' opportunity to license their works for that purpose. Such market harm can arise even where the entity is not pursuing a commercial purpose.

In addition, doctrines of secondary liability may be applicable in the case of an entity that induces, encourages, or profits from another party's AI training activities that directly infringe copyright. *See MGM Studios, Inc. v. Grokster, Ltd.*, 545 U.S. 913 (2005).

8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

A court would need to apply the fair use analysis to each stage of the process that involves an unauthorized use of copyrighted works. If the training activities are done purely for noncommercial purposes, and without any expectation of commercial benefit, the training entity may be able to argue that the first factor favors fair use as to those activities. As noted, however, a court would still need to consider the other factors, including whether and to what extent the training activities may affect the copyright owners' markets.

If the training party makes the resulting models or datasets available to a commercial entity, the transfer will in some manner involve the reproductions and/or distribution of copies of the ingested works. Again, the training party's purpose in doing so will be a key consideration. To the extent the training party makes its datasets or models available with the intention or expectation of deriving a benefit from their commercial exploitation by other parties, the first factor is unlikely to favor fair use. Evidence that the training party is funded by for-profit AI developers would be highly relevant in proving such intention. Additionally, as noted above, such a party could be liable under secondary liability doctrines to the extent it induces, encourages, or profits from the commercial entity's infringement.

Moreover, the fact that a dataset was created for noncommercial research use should not impact how a court analyzes the commercial entity's use of the dataset under fair use.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

As detailed above, the leading AI platforms—including ChatGPT, Bard, and MetaAI—are trained on billions of copyrighted works obtained by crawling and scraping the entire open internet. These works include hundreds of thousands of books copied from one or more notorious pirate repositories such as LibGen, Z-Library, Sci-Hub, and Bibliotik.

The number of materials copied without permission does not affect the fair use analysis. There is no "volume" exception to copyright, and a party cannot avoid liability by claiming that the scale of its infringement is too vast to calculate. Some AI companies have argued that licensing of works for training is impossible given the number of rightsholders affected. As discussed above, that objection is unfounded, as it ignores the numerous examples of successful licensing models in the United States and around the world. But regardless, the volume of rights implicated by a use has never been recognized as a relevant consideration under fair use. While courts consider the justification for the copying under the first factor and the reasonableness of the amount copied under the third, those inquiries go to the fairness of copying *particular works*, not to the purported need to copy works on a mass scale. To conclude otherwise would have the perverse result of privileging large-scale infringers over all others.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

Fair use is a defense to infringement, and so the fourth-factor analysis looks to the market for or value of the work or works alleged to be infringed. Though current U.S. copyright law may not always protect against the creation of AI-generated outputs that are similar to, or in the style of, an author's works but are not substantially similar to any particular work of that author, such outputs in many cases will nevertheless harm the potential market for or the value of the work, as well as the public's interest in incentivizing the creation of new works.

When these outputs are sold in the marketplace in competition with an author's or artist's own work, 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.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

Copyright law is opt-in, meaning that unauthorized use of a copyrighted work is prohibited (absent an exception or statutory license) unless the rightsholder has affirmatively consented beforehand. That should continue to be the law in the AI context, with the possible exception of an extended collective licensing (ECL) mechanism for past training uses, as outlined below.

For the mass ingestion of copyrighted works that has already taken place, it may not currently be possible as a practical matter for rightsholders to retroactively opt out of training. To the best of our knowledge, technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained, though such technologies are in development. *See, e.g.*, Ronen Eldan & Mark Russinovich, *Who's Harry Potter?: Approximate Unlearning in LLMs* (Oct. 2023) (proposing a "novel technique for unlearning a subset of the training data from a LLM, without having to retrain it from scratch"), https://arxiv.org/pdf/2310.02238. As discussed below, an ECL system would give rightsholders the opportunity to receive compensation for this prior unauthorized use. A rightsholder who does not wish to participate in ECL would have the opportunity to opt out of the licensing system, which would preserve their right to sue for the unauthorized ingestion that has taken place. It would likely not, however, result in removal of their works from the AI model given these current technological limitations.

Some AI companies have adopted practices of recognizing opt-outs especially when scraping content from the internet. (See response to question 9.2 below.) These practices pertain to collecting content for future training, and do not address any past training. In all events, these practices do not insulate them from liability since copyright law is opt-in, and not applying or exercising an opt-out should not factor into a decision on infringement or fair use. In any event, the opt-out mechanisms provided by LLM companies to date only provide the opportunity to opt out by those who control the web domains. So, an author whose book is on a pirate site has no opportunity to implement one of the "do not crawl for training purposes" tags, nor does the writer whose blog appears on a third-party platform have any ability to "opt out." They are wholly reliant on the platform or publisher to do so, likely per unnegotiable click-through agreements required for use of the service.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?

Consent should be required for all such uses, not just commercial uses. As noted, unauthorized use of copyrighted works for AI training can harm the licensing market for such materials even where the use is noncommercial in nature. To the extent a party argues that certain noncommercial training activities are fair use, the noncommercial nature of those uses will be factored into the first-factor analysis. Thus, existing law can already account for noncommercial training activities.

9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses?

For the mass, indiscriminate training of AI that has already taken place, the AI companies may not be able to identify all works that the AI was trained on, such as copyrighted texts in the Common Crawl or other web crawl dataset. AI companies as such may need blanket licenses that would cover all of the potential works that were used. Legislation permitting ECL would assist with this. ECL would allow qualifying organizations that represent a large number of a particular class of creators to negotiate licenses in the marketplace for a specific type of use on behalf of the entire class of copyright owners, whether or not they are current members of the organization.

There must, however, be an effective, robust, and easy to use mechanism for non-members to opt out of any such licensing system. The Copyright Office could be given authority to issue regulations governing notice and opt-out procedures. Notice methods, similar to those used in class-action lawsuits, should be used to ensure that all authors covered by the license are informed and have the opportunity to opt out. Opting out should be made very simple, such as through a publicly posted and easily located online form.

There are some web-based opt-out tools currently in use by AI companies, but these have significant limitations. For example, OpenAI has said that its web crawler, GPTBot, will not crawl websites when the site owner uses robots.txt to disallow access.[31] Similarly, Google has a product token called Google-Extended that "web publishers can use to manage whether their sites help improve Bard and Vertex AI generative APIs, including future generations of models that power those products."[32]

Because these tools are controlled by web publishers, as noted above, they are not effective in the many cases where the works at issue are owned by a third parties. For instance, authors often publish on third-party platforms and websites, such as Medium or Substack, that they do not own or control. Nor can they prevent the ingestion of the vast number of works on pirate websites, as noted above. For opt-out tools to provide any meaningful, practical purpose for authors, they

[31] See https://platform.openai.com/docs/gptbot.

[32] See https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers#google-extended.

must allow for copyright owners to themselves tag (or direct the tagging) of specific works to be excluded from training no matter where distributed or posted. We understand that such technology is being developed, but we do not know if it is affordable for individual creators.

But, again, we do not see opt-outs as a path forward. It is unfair to those who do not want their works used for training, and it provides no mechanism for compensation. The Authors Guild strongly believes that use of copyrighted works for training requires consent and, when requested, compensation.

Meta has a form with which users can request to delete personal information from third-party sources used to train its generative AI models. However, Meta has said that the form is not an opt-out tool and that they "don't currently offer a feature for people to opt-out of their information from our products and services being used to train our AI models." Kate Knibbs, *Artists Allege Meta's AI Data Deletion Request Process Is a 'Fake PR Stunt,'* Wired (Oct. 26, 2023), https://www.wired.com/story/meta-artificial-intelligence-data-deletion/ (quoting Meta spokesperson). It appears that the form is used only to initiate a review under any data protection laws applicable in the user's jurisdiction. *Id.*

9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

For ECL, legislation will be necessary to authorize a CMO to negotiate on behalf of non-members. The Copyright Office could be given the authority to authorize organizations who meet certain criteria to enter into agreements with the users—the generative AI companies—for the extended collective license. The requirements might include, for instance, demonstrating that the organization 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 an entire class of creator and copyright owner members without government rate-setting or oversight responsibilities.

For the licensing of works for future AI training activities, advance consent should be obtained through voluntary collective licensing, and it is entirely feasible. We discuss this framework below in response to question 10.2.

9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

If an AI company fails to honor a copyright owner's opt-out request and uses their work in AI training, the rightsholder would be entitled to pursue the same remedies for infringement as are available to copyright owners generally. Failure to honor such a request would be evidence of willful infringement for purposes of calculating statutory damages. These existing copyright remedies are appropriate to address such unauthorized uses and would not preclude other causes of action where applicable.

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

Authors should retain the right to control and be paid for the use of their work in AI even when they assign the copyright or sign work made for hire agreements. While in theory individual authors can include provisions to that effect in their agreements with publishers, in practice such agreements are typically not freely negotiated. That is because few authors have much bargaining power and even best sellers who are in the strongest negotiation position only have any leverage around the edges. Most terms are provided by publishers as standard, on a take it or leave it basis, and authors and their agents find it impossible to get changes to such terms. Unlike creators who are employees, professional freelance creators, as independent contractors, do not have the right to bargain collectively, which would give them such leverage. The Authors Guild has long advocated for changes to labor and antitrust laws to give professional freelance creators the ability to engage in collective negotiations and other concerted economic activities. Without collective bargaining rights, the law must step in to ensure that individual creators have the right to prevent their work from being used for training purposes and to share in licensing revenue even when they have signed work made for hire or assignment agreements.

In addition, we support legislation that would provide for greater and clearer protection of authors' moral rights under the Berne Convention in the United States.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained?

Collective licensing has proven an effective means of providing licenses on behalf of a very large number of individual creators for specific uses in numerous creative sectors. Here, collective licensing could solve the problem of how to license a mass number of works to AI developers for AI training on behalf of individual creators and small business on an industry-by-industry basis, as discussed below. Collective management organizations (CMOs) that represent writers, for instance, would obtain mandates from writers who chose to be included that allow the CMO to license the creators' works (whether all or only those indicated). The creator would identify what uses they were willing to allow the CMO to license out on their behalf. For instance, an author could indicate what kinds of prompt and output uses would be permitted of their work (see response to question 6.2) and what kind of additional compensation they would receive if technically feasible to track.

AI companies would request licenses from the CMO for certain uses and the CMO could provide a ready dataset or a blanket license to works in their catalog that meet the parameters of the requested use. While it might be feasible to have the AI company pay fees set by each individual copyright owner, it would likely be far easier to manage if the CMO set rates and copyright owners agreed to the standard fees for particular types of works and uses.

10.1. Is direct voluntary licensing feasible in some or all creative sectors?

Yes. There already are instances of rightsholders licensing their works for commercial AI uses, and others are in the process of developing such licenses. Our proposed collective licensing framework would not preclude direct licensing by individual rightsholders. Collective licenses would be offered in the free market on a non-exclusive basis to licensees—meaning that copyright owners can always directly license and that there can be more than one organization offering a collective licensing solution.

10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses?

Collective licensing is an established concept and an effective means of paying creators and publishers where licensing creates market inefficiencies. In the field of literary works, the Authors Registry and the Authors Coalition of America have for many years distributed royalties received from foreign collective licenses to U.S. authors. For AI licensing, one or more new CMOs could be established or an existing one—e.g., the Authors Registry or the Copyright Clearance Center (CCC)—could be augmented for purposes of negotiating licenses with AI companies and distributing amounts collected to writers.

For AI licensing, one option is to have each CMO negotiate fees with the various AI companies and then distribute those payments to the creators and other copyright owners who have registered with the CMOs, setting aside funds for those who have not yet registered but might in the future. We envision offering licenses for past uses of copyrighted works in AI systems, as well as future uses.

What stands in the way of collective licensing is that antitrust laws impose risks to forming CMOs that set rates on behalf of their members. As such, the CMOs and their members might be exposed to public and private antitrust action. We therefore are seeking legislation clarifying that these specific AI-use collective licenses will not violate U.S. antitrust laws. This could take the form of legislation that would allow certain creators (for instance, in the text and image sectors) to collectively bargain with AI companies for a specified period for purposes of negotiating licensing fees.

10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed?

No. We do not see a need for compulsory licensing for AI. As the Copyright Office has long recognized, compulsory licenses “conflict with the fundamental principle that authors should enjoy exclusive rights to their creative works, including for the purpose of controlling the terms of public dissemination.”[33] Therefore, such licenses are appropriate “only in circumstances of genuine market failure and only for as long as necessary to achieve a specific goal.”[34] As discussed, there is no indication that AI licensing markets have failed or are likely to do so. Rather, there are numerous examples of successful collective licensing models that can be adapted for AI licensing.

10.4. Is an extended collective licensing scheme a feasible or desirable approach?

As discussed above in response to questions 9.2 and 9.3, extended collective licensing is an appropriate way to address the mass, indiscriminate ingestion of copyrighted works for AI training that has already taken place. Going forward, licensing likely can be accomplished through voluntary collective licensing on an industry-by-industry basis for professionally created and commercially published and distributed works. That is not necessarily true for the vast amount of text on the open web that was written and owned by members of the general public who have not in any way attempted to monetize the posted texts.

If courts or Congress determine that fair use or implied licenses do not apply to the use of publicly available open web content, such as the websites of organizations intended for informational purposes (and where the text is not intended to be monetized), or platforms on which people post text (and do not assign or license the copyrights in the text per the terms of service, an area that is in need of some regulation), then AI companies will need the ability to license from a vast number of non-professional creators. Where CMOs and professional organizations have means and knowledge to reach most creators in their respective fields to obtain express permission, that will be much harder to do when dealing with the public at large. An ECL may be an appropriate way to compensate members of the general public, as opposed to professional creators, on a forward-looking basis.

Many writers and other professional creators have precarious livelihoods, and to maintain their careers must carefully manage and control what, when, how, and how much use of their work is allowed under what circumstances. Moreover, it is crucial to recognize their moral rights in their work. As such, an ECL is appropriate on a going forward basis only where opt-in, affirmative licensing is not manageable. For now, we believe that opt-in licensing for professional writers will provide sufficient training materials to allow the continued development of generative AI. If rights are needed from the general public, however, an ECL would be an appropriate and manageable means of providing such licenses.

10.5. Should licensing regimes vary based on the type of work at issue?

[33] U.S. Copyright Office, *Legal Issues in Mass Digitization: A Preliminary Analysis and Discussion Document*, at 38 (2011), https://copyright.gov/docs/massdigitization/USCOMassDigitization_October2011.pdf.

[34] *Id.*

Yes. As discussed above, rightsholders in particular industries are in the best position to determine the licensing structures, royalty calculation methods, and distribution processes appropriate for that sector, just as is true for other types of copyright licensing.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model, and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

The legal, technical, and practical issues associated with licensing are discussed above in response to questions 9 and 10. With respect to who should be responsible for securing licenses, as discussed above in response to questions 8.2 and 8.3, any party that engages in a use of copyrighted works that implicates one or more exclusive rights is required to obtain a license unless an exception applies. The exclusive rights implicated, and the application of the fair use factors may differ at different stages of the process. Further, as *Warhol* confirmed, the fact that one use in a series of uses may be deemed fair (in that case, Andy Warhol's creation of the Prince silkscreens) has no bearing on the analysis with respect to the other uses. Each use must be analyzed separately.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

Some existing technologies and others in development make it possible to identify the degree to which a particular work contributes to a particular output from a generative AI system. Current technologies like GPT-Zero's AI detection model and Personal Digital Spaces' similarity analysis technology are able to detect where an output has borrowed from a particular work; however, these technologies are not able to trace the use of a work's parameters through the inference to output stage. We understand that more precise ways to trace a work's influence are currently being researched and developed.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

Today's foundational LLMs are valued at tens of billions of dollars, and their use already in these very early years generates large amounts of revenue for the companies that own them;[35] as a result, we don't see foresee licensing requirements being a cost barrier at least for developers of large, commercially used models. Licensing the copyrighted materials to train their LLMs may be expensive—and indeed it should be given the enormous part of the value of any LLM that is attributable to professionally created texts and the risk to the value of creative works from LLMs. And so, expense cannot be treated as an excuse for LLMs not to be trained on licensed material

[35] Cade Metz, OpenAI in Talks for Deal That Would Value Company at $80 Billion, N.Y. Times (Oct. 20, 2023), available at https://www.nytimes.com/2023/10/20/technology/openai-artifical-intelligence-value.html

or a reason to value LLMs over creators. We must find a way for the creative professions to remain robust and co-exist with LLMs and other generative AI.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

Nothing at this time.

Transparency & Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

Yes. We believe that commercial AI developers who make generative AI models commercially available should be required to disclose the dataset and/or works used to train their models. This will ensure safety of the models, and prevent use of sensitive, harmful, or illegally harvested data in the training. The requirement will also further encourage AI developers to work with copyright owners to license works for AI uses and allow copyright owners to know when their works have been used.

We don't think it that is necessary to impose similar obligations on creators of training datasets as long as there are robust requirements on the AI developers to ensure that the content in the datasets, they are using is ascertainable and does not violate copyright, privacy, and other laws. AI developers typically process and clean datasets even if they have not compiled them, and disclosure requirements will compel them to remove illegal data from the training datasets.

15.1. What level of specificity should be required?

AI companies should be required to make publicly available or provide access to a complete list of all copyrighted works used in the training datasets, including the URL from where the data was obtained, and any copyright management information. This requirement should cover all models that were developed using copyrighted works without the permission of the copyright owners. For models that will be developed in the future, the works should be licensed, in which case the disclosure and transparency requirements may be relaxed so that the AI company can just point to the datasets, or the licensors of content used to train.

15.2. To whom should disclosures be made?

These disclosures should be public and part of the "model" card and training information that AI developers customarily release. We understand that some AI have developers have expressed reservations about disclosure of training information because of "trade secrets;" however, we don't see how disclosing a list of copyrighted works without any further details about weights or parameters jeopardizes the developers' trade secrets.

15.3. What obligations, if any, should be placed on developers of AI systems that

incorporate models from third parties?

Developers that incorporate models from third parties should be required to conduct due diligence before using the models and be required to only use models for which the detailed data subject to disclosure requirements is available. Developers of AI systems should also be obliged to secure rights to or ensure that rights have been obtained for any copyrighted material.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

We don't think the costs of compliance with a record-keeping and transparency obligations will be prohibitive for AI developers. If the datasets were compiled in-house, then the developers already know where and what data is included. And if the datasets are obtained from third parties the AI developers can require that the information be provided to them.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

We believe that copyrighted works should be licensed for training future models. For models that have already been created, we would like to see disclosure of the training data, notification to the copyright owners of the use, and an ECL for text-based works, so that authors who wish to be compensated for past and current uses can be compensated.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

No response.

Copyrightability

18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the "author" of material produced by the system? If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?

We generally agree with the guidance issued by the Copyright Office on March 16, 2023, regarding the registration criteria for works containing material generated by AI. We also agree with the detailed analysis of the "human authorship" in the guidance, and its application to the question of copyrightability. The Copyright Office, in line with longstanding case law and the D.C. District Court's recent decision in *Thaler v. Perlmutter*, has adopted the position that copyright eligibility requires human authorship. Accordingly, any material that lacks human authorship should not receive copyright protection.

Human creators can and do use generative AI to as a tool to assist in creation, and in many of those cases, the resulting work will be copyrightable authorship of the creator. For instance, if an author prompts an LLM to revise a story that she wrote to change it from the third to the first person, and the AI output substantially reproduces her story, then authorship should inhere in the work as a whole, though any AI-generated portions that the author adopts in the final work would not be protected.

We think it is reasonable and consistent with copyright law to grant copyright in output that are generated by an author using models or embeddings for models that an author has trained (or authorized to be trained) on their own work and expression and if the output is substantially similar to the author's original expression.

As the Copyrights Office explains in the Kashtanova letter, the determining factors in the analysis of whether a work created using AI can be copyrightable are control and predictability.[36] While these factors look at the "process" by which a work is created, they derive from the theory of authorship as articulated in case law,[37] and elaborated upon by copyright experts like Jane Ginsburg: that authorship requires "conception and execution."[38] For instance, as in the Kashtanova matter, where the output of an AI system is unpredictable, it may be difficult to show that there was sufficient control and consequently a sufficient closeness between "conception and execution."

We believe that the Copyright Office's holistic approach of considering the interrelated factors of control, predictability, conception and execution factors is correct and practical way to analyze questions of copyrightability in the AI context, as opposed to a bright-line rule or standard. This approach gives the Office flexibility to evaluate new use-cases of AI while adhering closely to the human authorship requirement. If a human uses AI tools to create a work that she has conceived and she controls the execution of the work, then the work as a whole is human authorship deserving of copyright (and any AI-generated elements of the work are excluded for protection just as any facts, scene à faire, or public domain elements would be). In cases where there is little or no evidence of creative control or direction, there should be no copyright. Even when there is human direction and control, the AI-generated portions, if separable, are not copyrightable and should be disclaimed in an application for registration of a claim to copyright.

For example, we do not believe that providing iterative series of text commands or prompts alone is sufficient to claim authorship in output from a text-generating AI model. An AI system can generate an innumerable, perhaps infinite, number of outputs from a given prompt. The user

[36] The processing of text instruction to images created from training data makes the output unpredictable. However, in a situation where an artist has trained a model on their own art and associated certain text strings to predictably produce results derived from the artists' own training data, the result may be different.

[37] See for e.g., *Burrow-Giles Lithographic Co. v. Sarony*, 111 U.S. 53, 58 (1884) ("An author . . . is 'he to whom anything owes its origin; originator; maker; one who completes a work of science or literature.' . . . writings in that clause is meant the literary and other productions of those authors… by which the ideas in the mind of the author are given visible expression.").

[38] Jane C. Ginsburg & Luke A. Budiardjo, Authors and Machines, 34 BERKELEY TECH. L. J. 343 (2019), available at: https://scholarship.law.columbia.edu/faculty_scholarship/2323.

cannot reasonably conceive of and predict the outcome of any particular prompt, no matter how long. It is a little like spinning a roulette wheel with infinite possibilities. No one should own copyright in those outputs. And when a user spins the wheel dozens of times until they land on an output they like, it does not give the user any more right to claim ownership of that one output. The doctrine of authorship by adoption, which some argue for, does not cover generative AI outputs. It allows for some unplanned elements to enter a work and be adopted by the creator, but it is a matter of degree and does not give copyright protection to random outcomes.

There may, however, be cases where the prompts are so directive and detailed, and the author directs the AI to produce many iterations and refines outputs with each subsequent prompt until they are able to get the machine to produce what they conceived,[39] that the human author did conceive and execute the final adopted output. In such cases, the human author should be entitled to copyright protection. That scenario is likely to be rare, as most creators will find it more efficient and easier to control the execution of a work through other means.

The more common case is where a human author directs an LLM to create an output and then rewrites, edits, and rearranges it to such a degree that she makes it her own and little of the original AI-generated text remains. Some writers use generative AI in this way. The Chatbot helps them brainstorm and may even give them text to start with, but they fully revise and edit the outputs and put it in their own voice, so that very little of the AI-generated material remains. In that case, authorship will inhere in the final output, and any AI-generated text that remains in the final would be disclaimed.

In sum, the degree of human authorship in works created by humans using AI tools generally will have to be evaluated on a case-by-case basis. The Copyright Office cannot be expected to conduct that evaluation and so must rely on the honesty of applicants to disclaim AI-generated outputs and elements. Decisions about what exactly is or is not copyrightable in a given work can be made by courts in infringement cases under existing copyright doctrines, and AI-generated elements can be filtered out prior to comparing the allegedly infringing work to the partially AI-generated work in a substantial similarity test, just as courts filter out other non-protectable elements.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?

We do not think any revisions are necessary to the Copyright Act to clarify the human authorship requirement. Even though the human authorship requirement is not specifically stated in the Copyright Act, it is implied, and a long line of case law establishes human authorship as an incontestable requirement for copyrightability.

[39] We want to emphasize that our comments pertain to "iterative" prompting, and not a single prompt. A single prompt even if several pages long could result in an innumerable number of outputs, meaning the prompter has no real control over the output. The person crating the prompt however has a copyright in the prompt assuming it has sufficient original expression.

With respect to works where human and AI-generated material is commingled, existing jurisprudence in several areas of copyright law can assist courts and litigants in determining what is copyrightable or not and what is authored by whom in a given work. The courts have proven adept at sorting through these issues. Areas of the law that will instruct courts in how to determine what is copyrightable in an AI-assisted human-created work or human-assisted AI-generated material include, among others: (i) cases involving questions of authorship and originality in other contexts, (ii) derivative work cases where courts had to decipher what aspects of the derivate work the derivate work copyright owner possessed, as compared to those owned by the owner of the original works it was based on, (iii) joint work cases where the issue of whether a secondary creator contributed a sufficient amount to rise to the level of an author,(iv) the "lion's share" line of cases addressing whether the computer program developer or user was the author of software outputs, and (v) infringement cases that filter out non-copyrightable material prior to comparing works for substantial similarity.

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?
20.1. If you believe protection is desirable, should it be a form of copyright or a separate sui generis right? If the latter, in what respects should protection for AI-generated material differ from copyright?
21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material? Would such protection "promote the progress of science and useful arts"? If so, how?

We oppose granting legal protection, whether through copyright or a separate sui generis right, to AI-generated material because doing so will dilute the market for human-created works and because it does not serve the goals of copyright or the needs of society.

First, developers of AI systems and technologies do not need additional incentives beyond those they already have in the form of patent, copyright, and trade secret protection for their technologies. Generative AI technology development until now has occurred without copyright incentives on outputs, and we see no rationale for providing copyright or sui generis protection for AI outputs to encourage further innovation in that area. LLMs do not earn income today from selling outputs; third party users do. Most LLMs do or will earn back their investments through licenses to third party businesses and researchers, or through subscription services. Open AI reports that already—less than one year after the release of ChatGPT—it is earning $1.3 billion a year in revenue.[40] In addition, LLM owners recover value from their users who, in the course of using the LLM or associated chatbot, help fine-tune their systems.

Second, the amount of funds that most AI companies already have received and that investors are willing to contribute to further development dramatically eclipses the funds available to support

[40] OpenAI's Revenue Crossed $1.3 Billion Annualized Rate, CEO Tells Staff, The Information, https://www.theinformation.com/articles/openais-revenue-crossed-1-3-billion-annualized-rate-ceo-tells-staff

the creative professions. As such, it would be absurd to give preference to the AI companies over the creative professions, and thereby trade in our future culture so that the AI companies can make even more money. Imagine what writers and artists could do with just .001% of all the money pouring into generative AI development!

Nor do the users of generative AI systems need additional legal incentives to generate content. It is extremely easy and cheap to generate content using these systems. If AI-generated works were entitled to the same protection as human-created works, it would give their producers (who in the book field for now are mostly scam-like operations that appear to have no writing or publishing background, but are just trying to find a way to mislead people to make a few bucks) unfair leverage in the marketplace and would further incentivize the distribution of AI-generated content to the public, 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. Publishers might survive by using generative AI to produce stories or assign writers to edit AI-generated material at a much lower fee, and indeed even though many trade publishers today abhor that idea, investors might insist on it.[41] The creative middle-class professions, however, will be drowned out and decimated. Since all human creativity starts with the human creators, our literary works and arts will suffer tremendously as a result.

AI-generated content can never serve in the place of human art, and so it is not in any of our interests to foster it and put it on equal footing with human creativity. Humans necessarily put some of themselves, their thoughts, emotions, experience, and personalities into the works they create; and an original work of authorship must contain that spark of human intellect. That human authorship is what the Constitution protects. Generative AI outputs will never contain the human spark; nor are AI-generated outputs ever original in the sense of the "original authorship" required under the Copyright Act. AI outputs are always merely derivative of the works the AI was trained on, lacking any new meaning or expression. AI can mimic human creativity, but only by regurgitating what it has been trained on, and as such is always stuck in the past. Since generative AI outputs do nothing to promote the "Progress of Science and useful arts…"—the very basis of copyright law under the U.S. Constitution—they should not be given protection.

Noam Chomsky, in a New York Times guest essay written with Ian Roberts and Jeffrey Watumull, explained how different generative AI is form human intelligence and creativity:

> The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant

[41] We have not seen much of this in trade book publishing yet, but imagine the pressure will come soon enough. The New York Times is already looking to hire a Newsroom Generative AI Lead. Newsroom Generative AI Lead, https://nytimes.wd5.myworkdayjobs.com/en-US/NYT/job/Newsroom-Generative-AI-Lead_REQ-014906-1

> system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations.[42]

The creative professions are too precarious as it is, and without a class of creative professionals, we will have very little new real art or literature. Our culture will be frozen in the past and will not reflect our current human experience.

Anyone literate can write, and anyone with access to an LLM will be able to generate a book, but a person who has natural storytelling or writing talent who then trained their lives to be a writer makes for very different writing. Studies using MRI's that see what parts of the brain light up when writing show that professional writers actually use a different part of the brain than hobbyists. It is the same part of the brain that professional athletes, musicians and other "experts' use—the part of the brain that kicks in after you put in the 10,000-hours that Malcolm Gladwell talked about in his book *Outliers*.[43] The kind of writing that professional writers do is qualitatively different from amateurs. But if authors cannot afford to write as a profession, then they won't, and we will lose diversity as well as quality in new works.

<u>Infringement</u>

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

Yes. Just as is true with human-created material, an AI-generated output that is substantially similar to a preexisting copyrighted work will, absent an exception, infringe the copyright owner's reproduction and/or derivative work right. As discussed below, liability for the infringement will depend on whether and to what extent a party plays a role in the generation of the output—for example, by creating or training AI systems or prompting the AI system to produce outputs that mimic the work of a particular author.

Moreover, even where a party is not liable for the *generation* of infringing output by an AI system, they may nevertheless be liable for infringements based on subsequent uses of that material. For example, a party that makes infringing AI-generated output available to the public

[42] Noam Chomsky: The False Promise of ChatGPT, New York Times, March 8, 2023, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html?searchResultPosition=1
[43] Carl Zimmer, This is Your Brain on Writing, New York Times, June 20, 2014, https://www.nytimes.com/2014/06/19/science/researching-the-brain-of-writers.html

online would be infringing the copyright owner's reproduction, distribution, public performance, and/or public display rights.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

The substantial similarity test is generally adequate to address claims of copyright infringement based on AI-generated outputs. But it may not be adequate to address the separate harm that arises when AI is used to generate materials that mimic an author or artist's recognizable style. As discussed in response to question 5, such outputs are a new form of exploitation of copyright owners' works, and they undermine rightsholders' markets when they are sold in the marketplace in competition with the original works upon which they are based. State rights of publicity or unfair competition laws may address this type of appropriation in some cases, but will not always be sufficient, particularly given the interstate nature of online commerce. Congress therefore should adopt a new economic right, whether under copyright law, a federal right of publicity law, or as a *sui generis* right, to ensure that rightsholders retain control and can be compensated for AI outputs that copy recognizable style or are identifiably similar to or taken from a copyrighted work.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

AI companies should be required by law to keep records of what data, including what copyrighted works, they used to "train" their models and to publicly disclose those datasets and copyrighted works. Existing civil discovery rules, under which records are merely subject to discovery or subpoena, are not sufficient to allow the public to evaluate and researchers to ensure safety of the models. Requiring public disclosure will prevent use of sensitive, harmful, or illegally harvested data in the training. It 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 done in many cases.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

As with all infringement claims, the direct liability of each party will depend on whether they have taken an action that infringes one or more exclusive rights; and whether a party is secondarily liable will depend on whether their actions meet the required elements of contributory, inducement, or vicarious liability. In other words, it will require a separate evaluation in each case.

25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs?

Open-source models do not raise unique considerations with respect to infringement. Like other AI systems, such models can displace human creators by generating outputs that compete with the works on which they were based. Therefore, any laws or regulations in this area should not treat open-source models differently from other generative AI models.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

Our understanding is that copyright management information is typically stripped out when works are ingested for AI training. This means that when an author or artist's work is incorporated in a generative AI output, they are not credited for the work, any more than they are paid for it, even when it is in the style of the artists or closely resembles their work.

Section 1202(b) currently prohibits the removal or alteration of copyright management information, but only if is done with the party "knowing, or… having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement." Because the AI developers who have used copyrighted works to train AI without authorization have generally claimed fair use, proving knowledge of infringement could be difficult. Even if the copyright owner ultimately establishes in court that the use was not a fair use, the developers will argue that they did not know at the time that the removal of metadata would induce infringement because they believed that the use was non-infringing.

To close this loophole, section 1202(b) should be amended to delete the knowledge requirement. Alternatively, a new section could be added that states: "No person shall, without the authority of the copyright owner or the law, intentionally remove or alter any copyright management information in the course of using a work to train artificial intelligence."

27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI-generated output.

No response.

Labeling or Identification

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

Yes, Congress should adopt legislation to require labeling of materials substantially generated by AI. Such legislation would help protect consumers from being misled into purchasing or consuming AI-generated content that they assumed was human-created, and to otherwise identify when content has been generated by AI. It would also help reduce incentives to dump large quantities of low-quality AI-generated content into online and other marketplaces. In addition, it would protect consumers against fake text, imagery, and videos that are passed off as authentic news.

At least two labeling bills have been introduced in this Congress that could provide workable approaches. Under the AI Disclosure Act of 2023, H.R. 3831, any output generated by generative AI must include the following notice: "Disclaimer: this output has been generated by artificial intelligence." Under the AI Labeling Act of 2023, S. 2691, each AI system that, using any means or facility of interstate or foreign commerce, produces AI-generated content shall include a "clear and conspicuous notice" identifying the content as such. The labeling requirements differ depending on whether the material is text-based or image, video, audio, or multimedia content. Additional requirements are discussed in response to the questions below.

28.1. Who should be responsible for identifying a work as AI-generated?

We agree with the approach taken by the AI Labeling Act, which requires the AI system used to generate the content to provide a persistent identifier in the output that it is AI-generated. In addition, it requires both developers of generative AI systems and third-party licensees of such systems to prevent downstream uses of the systems without the required disclosures, including by (1) providing contractual prohibitions against removal, (2) requiring certification that users and licensees will not remove such disclosures, and (3) terminating access to the system where they have reason to believe a violation has occurred.

In addition, social media and marketplace platforms should be required to use best efforts to require users to disclose if content available to their platform is AI-generated and they must make that information publicly available to prevent consumer deception.

We believe that people will prefer human content over AI-generated (aside from the novelty of it) and have the right to know in advance of paying for or consuming such content that it is AI, not human-generated.

28.2. Are there technical or practical barriers to labeling or identification requirements?

We understand that the persistent watermarking and other available labeling technologies are applied to works as files, and that they do not work well for text since text is infinitely divisible. Text can be cut and pasted from any file, and so even if you watermark or add another persistent identifier to a text file, it can easily be circumvented. We nevertheless believe that such labels should be required and believe that those requirements will incentivize the development of better technologies.

28.3. If a notification or labeling requirement is adopted, what should be the consequences of the failure to label a particular work or the removal of a label?

The Federal Trade Commission or other agency should be given enforcement authority. Both the recently introduced AI Disclosure Act, H.R. 3831 and the AI Labeling Act, S. 2691, follow this approach, and provide that a violation shall be treated as an unfair or deceptive act or practice under section 18(a)(1)(B) of the Federal Trade Commission Act (15 U.S.C. 57a(a)(1)(B)). Congress also should consider adopting a private right of action for any person harmed as a result of obtaining substantially AI-generated content that was not appropriately labeled.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools? What are their limitations?

Some tools of this type currently exist; a well-known example is GPTZero. To date, however, these tools are faulty in that they all are over-inclusive in finding content to be AI-generated. This is especially true when the text is either very common (as in the Bible) or is written by non-native English speakers. We are aware of some universities that have tested these tools and decided against using them because they are biased against foreign students and are not accurate.

Such tools will likely improve over time, and we have spoken to one developer that says that their results are getting much better. At the same time, with each improved model of LLMs (e.g., GPT-5) we anticipate that it will become increasingly difficult to distinguish AI content from human.

Additional Questions About Issues Related to Copyright

30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

In the context of books, federal trademark law, including section 43(a) of the Lanham Act, may be applicable to the extent AI-generated material is passed off as the work of a particular author. State unfair competition and rights of publicity laws may also apply in these circumstances.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

Yes. See the above responses to questions 8.5 and 23. We take no position on whether such a law should preempt state laws in this area.

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works "in the style of" a specific artist)? Who should be eligible for such protection? What form should it take?

Yes. See the above responses to questions 8.5 and 23.

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? Does this issue require legislative attention in the context of generative AI?

No response.

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

We urge the Copyright Office to prepare a written report providing legal analysis and policy recommendations on these issues.

Respectfully submitted,

Mary Rasenberger
CEO, The Authors Guild

Umair Kazi
Director of Policy & Advocacy, The Authors Guild

Kevin Amer
Chief Legal Officer, The Authors Guild

AG The Authors Guild

APA AMERICAN PHOTOGRAPHIC ARTISTS

ARTISTS RIGHTS SOCIETY

ASCRL™ American Society for Collective Rights Licensing

DRAMATISTS GUILD

GRAPHIC ARTISTS GUILD

RWA Romance Writers of America® The Voice of Romance Writers

SFWA SCIENCE FICTION & FANTASY WRITERS OF AMERICA

SGA SONGWRITERS GUILD OF AMERICA

Sisters in Crime®

WOMEN'S NATIONAL BOOK ASSOCIATION EST. 1917

Legislative Proposals to Protect the Creative Professions and Mitigate Risk of Harm from Generative AI, August 10, 2023

We are organizations that represent creative professionals from diverse backgrounds. Collectively, our members include book authors, freelance writers, journalists, playwrights and dramatists, visual artists, songwriters, composers, lyricists, photographers, graphic designers, and other creative professionals.

Generative AI technologies pose a serious threat to our members' professional creative futures, and we believe that guardrails around their development and use are urgently needed to mitigate the profound financial and cultural harm that the unregulated use of generative AI will almost certainly bring to the creative professions. There is a serious risk of market dilution from machine-generated works that can be cheaply mass-produced, and which will inevitably lower the economic and artistic value of human created works.

We believe that it is inherently unfair to use and incorporate the works of creators in the fabric of AI technologies and their outputs without the creators' consent, compensation, or credit, including creating derivative works that will actually compete with those original human creators. We are asking for interventions to safeguard the incentives that fuel the creation of a rich and diverse creative culture and markets, so vital to our democratic culture that they are inscribed in the Constitution.

We are lobbying for laws, regulations, and policies that recognize the following and require:

1. Consent and Compensation: Require all generative AI companies to seek permission for the use of creators' works in generative AI systems, and to fairly compensate creators who allow their works to be used in "training"[1] of generative AI;

2. Credit and Transparency: Create obligations for all AI companies to disclose what datasets and works they use to "train" their AI systems in the past, present, and future;

3. Permission and Payment for use in outputs: Require all AI companies to seek permission and pay compensation when creative works are used in outputs, or when names or identities or titles of works are used in prompts—whether through adding a new economic right under copyright law or as a *sui generis* right, and through a broad, well-articulated federal right of publicity law;

4. Labelling AI-generated content: Require the conspicuous labelling of AI-generated works as such, with enforcement provisions;

5. Permission for Generative AI's Use of a Person's Identity, Persona, or Style in a Federal Right of Publicity: Create a federal right of publicity that would simplify bringing a claim for use of voice, name, image, or other indicia of a creator's identity (whether such creator is living or deceased). Unlike current state right of publicity laws, we believe that it should also encompass a creator's style where readily recognized by the relevant consumers; and

6. Prohibit Removal of Copyright Information in the Ingest/Training Process: Amend section 1202 of the Copyright Act so that it is a violation to intentionally remove "copyright management information" from a copyrighted work in order to train AI or create an AI training dataset without permission of the copyright owner, whether or not it can be proven that it was knowingly done to induce or enable infringement; and

7. No copyright for AI-generated outputs: Retain the current copyright law requirement that copyrighted works be human created; we oppose efforts to deem AI-generated content protectible under copyright law, or through the creation of even a limited *sui generis* right. Providing copyright or similar incentives to use AI to generate content will exacerbate the threat of AI-generated content flooding and overwhelming the market for human works that the Constitution seeks to promote and protect.

[1] We use the term "train" to refer to AI developers' use of pre-existing works in developing their AI only because it has become the standard shorthand. That said, we have reservations about the semantics of the word because it makes the incorporation of copyrighted works into generative AI sound like a one-time use and serves to anthropomorphizes machines—as if they are simply "reading" or "observing" texts and other works. The reality is that the works are used to build the AI program and remain part of its fabric. There is no generative AI without the material that AI is purportedly "trained" on.

Legislative Proposals for Protecting the Creative Professions

1. Consent and Compensation

Free-Market Collective Licensing

The current generative AI technologies were mostly "trained" on unlicensed copyrighted works. Neither consent, credit, nor compensation was provided. We ask Congress to clarify and remind AI companies that the law requires authorization to embody creative copyrighted works in generative AI technologies—outputs that will inevitably compete with and usurp the market for the ingested original human works on which they are based. Individual creators and small creator businesses also need congressional assistance so that they can establish a private, efficient, and cost-effective collective licensing system to provide AI companies with the appropriate rights in exchange for fair compensation.

Because it is not efficient for AI companies to attempt to seek licenses from each of the hundreds of thousands or millions of individual creators or small businesses who own the rights to their copyrighted works, they are unlikely to do so. The major AI companies have instead pushed the bounds of fair use and have simply risked getting sued by creators; and they will continue to use works without permission unless an efficient way to obtain licenses is developed.

Where the numerosity of creators or other copyright owners is a hindrance to licensing, collective licenses have proven an effective means of providing licenses on behalf of a very large number of individual creators for specific uses. Here, collective licensing could solve the problem of how to license a mass number of works to AI developers for AI training on behalf of individual creators and small business on an industry-by-industry basis. The licenses would be offered in the free market on a non-exclusive basis to licensees—meaning that copyright owners can always directly license and that there can be more than one organization offering a collective licensing solution. Some corporate copyright owners have indicated a preference for direct licensing but would also be welcome to join as well. Each collective management organization (CMO) would negotiate fees with the various AI companies and then distribute those payments to the creators and other copyright owners who have registered with the CMOs, setting aside funds for those who have not yet registered but might in the future. We envision offering licenses for past uses of copyrighted works in AI systems, as well as future uses.

Collective licensing is an established concept and an effective means of paying creators and publishers where licensing creates market inefficiencies. Such licenses have been used in the music industry for decades, for television programming, by the Copyright Clearance Center for text licenses, and by other entities. For over two decades, the Authors Registry and the Authors Coalition of America have distributed royalties received from foreign collective licenses to U.S. authors and visual artists. More recently the American Society for Collective Rights Licensing (ASCRL) has distributed royalties received from foreign collective licenses to U.S. authors, artists, photographers, and other rights holders, and the Artists Rights Society (ARS), a membership organization affiliated with the Paris-based CISAC, collects licensing income and manages licensing requests on behalf of more than 70,000 artist members, the majority of whom create works of fine (rather than commercial) art.

What stands in the way of collective licensing is the fact that antitrust laws impose risks to forming CMOs that set rates on behalf of their members. As such, the CMOs and their members might be exposed to public and private antitrust action.

<u>Legislative Request</u>: Accordingly, we seek clarification that these specific AI-use collective licenses will not violate U.S. antitrust laws.

Extended Collective Licensing

For the mass, indiscriminate training of AI that has already taken place, where the AI companies may not be able to identity all works that the AI was trained on, they will need blanket licenses that would cover all of the potential works. An extended collective license (ECL) would assist with this. It is a type of collective rights licensing where qualifying organizations that represent a large number of a particular class of creators can negotiate licenses in the marketplace for a specific type of use on behalf of the entire class of copyright owners, whether or not they are current members of the organization. There must, however, be an effective mechanism for non-members to opt out of the licenses.

Legislation is necessary to authorize these types of licenses (which otherwise must be provided on an opt-in basis). They are an efficient and rational way to license rights in cases of mass use, such as where rightsholders are numerous individual creators, and the users cannot negotiate directly with all of them due to their sheer numbers. The ECL legislation would specifically authorize qualifying organizations to negotiate blanket licenses on behalf of the entire class of creators on an opt-out basis.[2]

The Copyright Office could be given the authority to authorize organizations who meet certain criteria to enter into agreements with the users—the generative AI companies—for the extended collective license. The requirements might include, for instance, demonstrating that the organization 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 an entire class of creator and copyright owner members without government rate-setting or oversight responsibilities.

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

[2] For examples of ECLs, 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 (Consolidate 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.

Creators who authorize a CMO to license their works for such uses would receive a distribution based on algorithms that would take various factors into account, including for instance, the number of works published, the type or length of those works, and possibly any available sales data (with criteria varying by industry). The CMO could also provide compensation to creators for uses of creative works that were created as works made for hire.

A certain amount would be set aside for those who have not yet registered but may 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.

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.[3] The Office concluded that it was premature to create ECLs in 2017, but the proposal nevertheless presents a carefully, well thought-out model for an ECL in the U.S. Today, the new generative AI technologies have created renewed interest in extended collective licensing, as the works owned by many individual creators are already being used with impunity to train generative AI to produce material that competes with human creation.

Legislative Request: Accordingly, we seek legislation to authorize creator organizations to provide ECLs on behalf of a particular class of creators for generative AI uses under an opt-out regime. Such an ECL could apply to past use only provided that the law is clear that permission need be obtained going forward.

2. Credit and Transparency

Legislative Request: We seek legislation that will require AI companies to keep records of what data, including what copyrighted works, they used to "train" their models and to publicly disclose those datasets and copyrighted works.

Record keeping alone, where the records are merely subject to discovery or subpoena, is not sufficient to allow the public to evaluate and researchers to ensure safety of the models. Requiring public disclosure will prevent use of sensitive, harmful, or illegally harvested data in the training. It 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 done in some cases.

3. Permission and Payment for Use in Outputs

In addition to compensating creators for the use of their works in "training" AI, we seek to prevent AI companies from using creators' works in outputs without the creators' express permission. Where creators decide to permit this use, they should be compensated.

The CMOs (discussed above) could also license rights to allow AI to create derivative works or look-alikes/sound-alikes, and collect and distribute the fees on behalf of the individual creators

[3] https://www.copyright.gov/policy/massdigitization/

who wish to permit the use and be compensated, providing them with a means for earning additional income. This kind of licensing could be provided on a one-off basis (as many licensing organizations do today) or pursuant to blanket licenses as described above.

4. Labelling AI-generated Content

Legislative Request: We support legislation that requires labelling AI-generated works as such, with enforcement provisions that give it teeth, such as The AI Disclosure Act of 2023.

Such legislation would help protect consumers from being misled into purchasing or consuming AI-generated content that they assumed was human-created, and to otherwise identify when content has been generated by AI. It would also help reduce incentives to dump large quantities of low-quality AI-generated content into online and other marketplaces. In addition, it would protect consumers against fake text, imagery, and videos that are passed off as authentic news.

5. Permission for Generative AI's Use of a Person's Identity, Persona, or Style in a Federal Right of Publicity Law

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, impairing the copyright incentives as they are an uncompensated taking of the author's or artist's expression. Moreover they raise 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.

Legislative Request: Accordingly, we are seeking a new economic right, whether under copyright law, a federal right of publicity law, or as a *sui generis* right, to ensure that rightsholders retain control and can be compensated for AI outputs that copy recognizable style or are identifiably similar to or taken from a copyrighted work.

6. Prohibit Removal of Copyright Information in the Ingest/Training Process

When ingesting copyrighted works to train AI or to create datasets for training AI, information about work, author and owner is stripped out. This means that when an author or artist's work is incorporated in a generative AI output, they are not credited for the work, any more than they are paid for it, even when it is in the style of the artists or closely resembles their work. This is also unfair.

Proposal: We recommend amending section 1202 of the Copyright Act so that it is a violation to intentionally remove "copyright management information" from a copyrighted work without permission of the copyright owner, whether or not it can be proven that it was knowingly done to induce or enable infringement.

Section 1202 currently prohibits the removal or alteration of copyright management information (defined as information such as the title, author, owner, or terms of use for a work with which it is associated), but only if is done "knowing, or… having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement…" Because the AI developers who have used copyrighted works to train AI without authorization have generally claimed fair use, proving knowledge of infringement could be difficult. Even if the copyright owner ultimately establishes in court that the use was not a fair use, the developers will argue that they did not know at the time that the removal of metadata would induce infringement because they believed that the use was non-infringing.

Required Legislation: Section 1202(b) could be amended to delete the last phrase starting with "knowing, or… having reasonable grounds to know, that it will induce…"

Alternatively, a new section could be added that states:

"No person shall, without the authority of the copyright owner or the law, intentionally remove or alter any copyright management information in the course of using a work to train artificial intelligence."

7. No copyright protection for AI-generated outputs

We also continue to oppose efforts to expand copyright protection for AI-generated work, including through limited *sui generis* terms. 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. If AI-generated works were entitled to the same protection or similar 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.

Respectfully submitted,

The Authors Guild
American Photographic Artists
Artists Rights Society
American Society for Collective Rights Licensing
Dramatists Guild of America
Graphic Artists Guild
Novelists, Inc.
Romance Writers of America
Science Fiction & Fantasy Writers of America
Songwriters Guild of America
Sisters in Crime
Women's National Book Association