SFWA, Inc.
PO Box 215
San Lorenzo, CA 94580

To: Suzanne V. Wilson, General Counsel and Associate Register of Copyrights.
Maria Strong, Associate Register of Copyrights and Director of Policy and International Affairs.

The Science Fiction and Fantasy Writers Association (SFWA), formerly Science Fiction and Fantasy Writers of America, is a 501(c)(3) nonprofit organization whose mission is, in part, to support, defend, and advocate for writers of science fiction, fantasy and related genres. Formed in 1965, SFWA currently has over 2,500 commercially published writers in those genres across various types of media. Its membership includes writers of both stand-alone works and short fiction published in anthologies, magazines, and in other media. SFWA is not a subsidiary of any other entity. SFWA has no subsidiaries or other ownership interest in any other organization that may be affected by the Copyright Office's policies on AI.

We thank the Copyright Office for the opportunity to make a response to the comments that have been received; the sheer volume speaks plainly to the importance of this topic, and while we cannot hope to address everything that has been said—either to applaud the thoughtful or to demur from the mistaken—we greatly appreciate the interest this subject has raised. We call the CO's attention in particular to the many short but impassioned comments made by creators who are concerned for their livelihoods and their ability to continue to reach their audiences. Society relies on such creators to enrich our lives, and we ignore their plight at our peril.

As a thorough response is not possible, we seek to focus our reply on the many points raised concerning the question of fair use. We acknowledge that due to the sheer volume of comments, we may have missed valuable insights and egregious mistakes that we might otherwise have preferred to highlight. Nevertheless, in the comments we have seen, arguments on the subject of fair use seem to take similar forms, and we wish to address three broad categories:

- The source of human-created content for training;
- The effect of the use of generative AI on the marketplace;
- The very large scales involved.

SOURCE

On the subject of the source of human-created content, we find ourselves in the position of largely responding to silence on behalf of the creators of AI systems. In their responses to the Copyright Office, the creators of AI systems do not adequately address the provenance of the copyrighted works they defend using. They may mention the names of collections, such as CommonCrawl or Books3, but generally not how and why collections of copyrighted works are made available for this use.

For most published books, there are no legitimate ways to acquire the plain text of the work for free. To go from a published book in paper or ebook, to a text stripped of all formatting (and often copyright information) suitable for machine training, requires someone to make unauthorized copies. And these are copies in the commonly-understood sense: a human being can open the file and sit and read. Nor are these copies ephemeral; they are stored and passed around, and their use is not guaranteed to be restricted to only training AIs.

In other words, the creation of Large Language Models as currently practiced has relied on large-scale piracy of copyrighted works. The way in which they have chosen to acquire material has the effect of encouraging and legitimizing piracy. Authors who have made their work available in forms free of restrictive technology such as DRM for the benefit of their readers may have especially been taken advantage of. In addition, much of the market for professionally published science fiction and fantasy short stories relies on providing free access to those copyrighted stories on the Internet. It would be unfortunate if the freeloading of large corporations resulted in these artistic works being less available to human readers and scholars.

If the creators of these systems have been silent on the ultimate provenance of the materials they use, those who see the effects of piracy clearly have not been:

> “[A]s long as the content is available elsewhere, the opt-outs or blocks are not fully effective. AI developers and dataset curators often still access protected content through pirate websites, undermining the value of such prohibitions and exacerbating the harm to copyright owners.” (News Media Alliance, COLC-2023-0006-8956)

> “The unlawful presence of pirated material on the internet is a battle that educational publishers continue to fight, but the unchecked sourcing and scraping of educational content and material from around the internet to train generative AI models further propagates the harms of online piracy by dramatically expanding the reach of these pirated sources.” (Pearson, COLC-2023-0006-8703)

> “Moreover, the internet is replete with pirated copies of sound recordings, unauthorized compilations of song lyrics, infringing copies of photographs and images, and countless other unauthorized reproductions of copyrighted works. Training on pirated content is no more legal than training on copyrighted content ripped without authorization from an official source. Any further reproduction or distribution of these infringements only multiplies and perpetuates unlawful conduct through an infringing chain of distribution.” (Universal Music Group, COLC-2023-0006-9014)

> “[E]ven the most slavish pirate will merely claim that they are providing access to unprotected ideas for the purpose of analysis.” (Software & Information Industry Association, COLC-2023-0006-9041)

SFWA acknowledges the problem of generative AI scraping pirated material published as copy-protected ebooks by professional publishers, but SFWA additionally has the unique position of representing many authors who have fought to make their work available for free for human readers. Over the last twenty years, many science fiction and fantasy authors of short fiction have embraced the open Internet, believing that it is good for society and for a flourishing culture that art be available to their fellow human beings regardless of ability to pay. That availability is not without cost; it is quite difficult to bring an online magazine to market, and being freely available has never meant abandoning the moral and legal rights of the authors, nor the obligation to enter into legal contracts to compensate authors for their work and spell out how it may and may not be used. But on balance, many writers and fans believe that freely sharing stories is a good thing that enriches us all.

The current content-scraping regime preys on that good-faith sharing of art as a connection between human minds and the hard work of building a common culture. The decision to publish creative work online to read and share for free is not guaranteed; it is a trade-off of many factors including piracy, audience, and the simple (albeit elusive) ability to make a living. In too many comments to enumerate here, individual authors have made clear that they regard the use of their work for training AI to be another important factor in that mix, and the ultimate effect on the short fiction marketplace and its role in our culture is far from certain. Bluntly, many authors do not want their work taken for this purpose, and that cannot be ignored.

> "If my work is just going to get stolen, and if some company's shareholders are going to get the benefit of my labor and skill without compensating me, I see no reason to continue sharing my work with the public -- and a lot of other artists will make the same choice." (N. K. Jemisin, COLC-2023-0006-0521)

The developers of AI systems seem to believe that a green light to use scraped copyrighted work will result in a clear field for them to continue freeloading forever; we fear rather that it will result in large swathes of artistic work removed from the commons, locked behind paywalls and passwords to the detriment of all.

EFFECT OF USE

One of the difficulties of discussing fair use in this context is that the proposed uses for AI vary widely. Conversational chatting, answering questions about documents, producing snippets of computer code, suggesting words in an email, and writing novels are all technically text generation based on large language models, but have wildly different effects on the marketplace. Although the developers of AI systems might like to focus very narrowly on the instant and nature of training as determining fair use for all possible end results, creators have been consistent in addressing how AI systems, particularly generative AI systems, are ultimately used. The discussion is ill-served by attempts to conflate them all together and to bind all uses to a single yes-or-no determination that using copyrighted works as training for any machine learning for any purpose is or is not inherently fair use. 17 U.S. Code § 107 describes fair use considerations, and we would draw attention to the fourth element, "the effect of the use upon the potential market for or value of the copyrighted work".

The commercial nature of much of the use of generative AI is broadly the focus of the discussion, but it is worth first mentioning potential non-commercial uses, as there were a number of submitted comments regarding research and scholarship.

For example, researchers such as Project LEND at the University of California seek to preserve their ability to do research based on scanned books in the HathiTrust Digital Library. Analyses of fair use have always taken that sort of use into consideration and protected education and research. The examples they give of using generative AIs to generate abstracts and summaries, create annotated bibliographies, draw connections among award-winning books: these are not uses that have been understood to ultimately impinge on copyright. These would fall under scholarship and research as listed in 17 U.S. Code § 107, and they are not the sort of uses that creators have been protesting.

Indeed, creators' concerns about ethical and transparent sourcing align with those of scholars. Consider a use case they propose:

> "[A] digital humanities scholar wants to study the Pulitzer Prize for Fiction and use a non-generative AI model to help her discern the underlying themes, moods, and attitudes of the winning books." (Project LEND, COLC-2023-0006-8603)

It would be unfortunate for a researcher doing such work to rely on an AI for her conclusions about these books only to discover that her AI model was trained on pirated works that were incomplete or altered, or that the model included fan-written work scraped from the broader Internet and did not make distinctions. Clear licenses and provenance are vital even when a use is unquestionably fair; the needs of researchers and educators are not met by indiscriminate scraping.

Having established that we are primarily concerned in our comments with commercial uses, there are two harms that we wish to address here. First, the harm from unfair competition, and second the harm to licensing markets.

In reading through the provided comments, we find that the harm from competition is consistently mischaracterized. This comment was typical:

> "[M]any rightsholders more broadly focus on the possibility that the output of Generative AI models might in some sense "compete" with either the original works they were trained on or against the potential future output of authors, even when that output does not embody any expressive content of any particular work in the training set. But a use that enables creation of

> new, non-derivative works that might compete with the original does not result in cognizable market harm." (Meta, COLC-2023-0006-9027)

Or else, they narrowly define "market" to the point of absurdity, as in:

> "[I]n the context of textual works, it is highly unlikely that a reader interested in reading a specific book included in GPT-3's training dataset would turn to ChatGPT to obtain information about that book in lieu of purchasing or checking out that book from the library." (Authors Alliance, COLC-2023-0006-8976)

These comments betray misunderstanding of the nature of the harm to the fiction marketplace that has been inflicted by generative AI systems. AI developers seem to believe that creators are worried about being outcompeted by brilliant computer creations, and be deserted by our human audiences. Nothing could be further from the truth. The harm creators and audiences are already experiencing is a flood of trash, directly enabled by generative AI with no restrictions on output. A fuller accounting of a portion of these harms can be found in Neil Clarke's remarks as recorded in the Federal Trade Commission's comment attachment (COLC-2023-0006-8630), but the problem essentially is that genre writers rely on access to markets in which to sell their works; AI-generated material clogs them up and literally crowds human writers out. The danger is not that readers cannot find specific already-published work that they already knew they wanted to seek out, but that human writers' voices, especially new and marginalized writers, will be silenced in a sea of noise. Readers will lose books they never knew they needed.

When Andreessen Horowitz said, "Moreover, since this technology enables production of creative work at an unprecedented rate, the problem will compound over time." (Andreessen Horowitz, COLC-2023-0006-9057) they were referring to the question of tracking royalties, but is just as apt a way of describing the future of the harms these technologies will cause our markets.

Although fiction writers have had reason to be less fearful of being replaced outright by machines, supplanting human creators with machine-generated art is not academic; we need only look to the market for art to see our own future. One respondent enthuses,

> "Generative AI also enables authors to express themselves in new ways: image-generating systems like DALL-E and Midjourney enable authors to create illustrations to accompany their textual works where they otherwise might not be able to." (Authors Alliance, COLC-2023-0006-8976)

Let us be clear that "might not be able to" has traditionally been handled by hiring a human artist. The harm to the market for art is already clearly spelled out by those eager to inflict that harm for their own benefit; equivalent harm to the market for writing will not be far behind.

Aside from the harm being done to the publishing market by generative AI, the "take what we want" approach also harms the ability of authors to license their work for use in training AI, should they wish to.

We disagree with Meta's comments (COLC-2023-0006-9027) regarding the potential for a licensing market; they acknowledge that "it is possible that AI developers will strike deals with individual rightsholders, to develop broader partnerships or simply to buy peace from the threat of litigation" but their conclusion that "those kinds of deals would provide AI developers with the rights to only a miniscule fraction of the data they need to train their models" ignores the fact that for fans of individual authors, that "miniscule fraction" is extremely important.

As an example, fans of the works of author JRR Tolkien have driven licensing of his work in not only television, movies, and video games; but also LEGO sets, lunchboxes, replica helmets, etc. Meta ignores the reality that for popular fandoms, there is a licensing market for everything. Indeed, it does not take much searching online to see that fans are already investigating the use of Generative AI to produce new works in that vein.[1]

Meta's discussion in their lengthy section entitled, 'How Large Language Models "Learn"' uses the sentence, "Susan's aunt planted the flower in the garden" to illustrate how these tools would draw from a wide variety of sources to learn individual words. This enables their tools to fluently use words like 'flower' and 'Susan'; however, their explanation holds less true in how these tools would learn the words in a sentence like, "Frodo and Gandalf faced the Balrog in Moria." Learning the words in that sentence both requires a much more targeted input data set, and—with due respect to Susan's aunt—seems likely to be more commercially interesting.

[1] For example, https://www.reddit.com/r/tolkienfans/comments/11qju02/i_had_chatgpt_write_a_summary_of_a_sequel_for/

While Prof. Tolkien’s works may represent a “miniscule fraction” of the overall training data, it is clear that it is for some potential users a great attraction. It is easy to conclude that a savvy developer of AI technology might do well to license these works for an AI that caters to these fans, and would distinguish themselves in the AI marketplace thereby; Meta’s dismissal of such a licensing market seems highly premature.

SCALE AND COST

> "The second characteristic of training data to keep in mind is its scale. We intend that our AI tools will benefit from and reflect the full breadth of human reasoning and understanding." (OpenAI, COLC-2023-0006-8906)

As a rule, we have found that those who create or fund the creation of Generative AI systems argue in their responses that their use of copyrighted work constitutes fair use. This is unsurprising, given the significant amount of money at stake for them; a rule in favor of creators’ rights would increase their costs and it is natural (though unethical) for them to attempt to obtain what they need for free. Broadly speaking, they appear worried that they have appropriated copyrighted works in such volume that to pay for them would be very expensive. That may be the case, but ability to pay for what one takes is seldom a factor in whether one is entitled to take it.

Still, there is no reason that this should result in, “tens or hundreds of billions of dollars a year in royalty payments.” (Andreessen Horowitz, COLC-2023-0006-9057) or any other particular number, large or small. The value of these rights is at present unknown. The mention of such lofty numbers seems to us less a realistic valuation of the likely cost of royalties, and more a sign of the size of the budgets they need to pretend royalties would exceed. We suggest that creators’ groups such as SFWA be involved in determining the specifics and negotiations to determine how payment amounts are calculated. Needless to say, the payments should not make the use of copyrighted material prohibitively expensive, but at the same time, the potential value of these AI systems is potentially limitless and the major corporations that are currently involved are not paupers.

Comments on the topic of royalties essentially boil down to the complaint, “Our business needs so much data that we cannot pay for it all, and therefore we should be entitled to have it for free.” Were one to substitute any other commercial product for “data” – “electricity”, “water”, “bandwidth”, “flour” – the argument would be absurd on its face. To make such an argument even as they make little effort to undertake the negotiations needed to determine what those costs may actually be, underlines how unserious they are. Indeed, it seems just as likely that they use such volumes of data merely because they can, and that were they obliged to pay a fair price for it, they would develop algorithms that do not need nearly so much. Human beings learn to speak and write fluently without first consuming the entire Internet, for which we may be thankful in many ways; it seems reasonable to suppose that future generations of LLMs will too.

In any case it should not be ignored that creators may choose to make their work available for free for this purpose; the widespread use of Creative Commons licenses suggests that in fact many creators would not mind such a use and would permit it so long as their rights to opt in and attribution are respected. Those who produce or defend AI systems argue in their responses to the CO that the benefit to society outweighs the interests of individual creators, and perhaps they are right; let them make such an argument to those creators whose rights they seek. For corporations possessing war chests in the billions of dollars, it is perhaps easier to persuade the government to erase creators' rights than to persuade those creators to voluntarily participate, but persuasion and remuneration remain the better path.

We do not believe scale to be a significant hurdle to finding a way to respect creators' rights. The actual problems associated with any collective license are identifying, finding, notifying, and paying the creators of the original, as well as determining a fair way to determine the appropriate payments. These are not easy, but they are not intractable. Some have proposed an opt-out compulsory extended collective license to simplify the effort and covering many, probably most, authors who have never heard of the license and will never see a penny in compensation. SFWA prefers a voluntary, opt-in system that requires the author's consent; the test of whether it is truly unworkable should come from the genuine attempt.

CONCLUSIONS

Fundamental to creators' rights is the right to say "no, my work may not be used in this way." The principle that authorship conveys control is not only a commercial right but a moral one, and the range of comments in response to the Copyright Office's call make clear that it is a cherished right. Questions of "how" and "when" and "how much money" all come later; first and foremost the author must have the right to say how their work is used. And authors will always have that right: at the extreme, they can choose to never share their work with the public. It was with the intent to persuade authors to share their work and thereby build a culture for society that copyright was created in the first place.

Once an author publishes and gets the benefits of copyright, society has worked out rules to support research, scholarship, education, and fellow authors–to ensure, in effect, that society retains its own benefits from the bargain. The advent of generative AI has changed the nature of that bargain, however, and the well-funded developers of these systems have sought to take for themselves all of the benefits of the thriving culture they want to exploit, without accepting any responsibility or acknowledging any debt.

All that is needed is transparency, and the ability for individual authors to opt in. That is not to say that all problems can be easily solved; the federal government may have to provide a legal framework for this. We believe that the opt-in provision allows whoever creates the license to provide a standard offer and, if it isn't reasonable, authors will simply not opt in. If they don't, their work will not be included. The question of achieving the scale needed for AI training can then become a negotiation: they can make their arguments and offers, and we can make ours, and a balance will be reached where AI companies have persuaded or paid enough creators to get what they want.

So long as authors retain the right to say "no" we believe that equitable solutions to the thorny problems of licensing, scale, and market harm can be found. But that right remains the cornerstone, and we insist upon it.

Respectfully submitted on behalf of the Science Fiction and Fantasy Writers Association,

SFWA Board of Directors

SFWA Legal Affairs Committee

Contact: John Murphy, SFWA Vice President (vp@sfwa.org)