TRANSCRIPT OF PROCEEDINGS

In the Matter of,)
)
COPYRIGHT AND ARTIFICIAL)
INTELLIGENCE LITERARY WORKS,)
INCLUDING SOFTWARE, LISTENING)
SESSION)

Pages: 1 through 139

Place: Washington, D.C.

Date: April 19, 2023

HERITAGE REPORTING CORPORATION
Official Reporters
1220 L Street, N.W., Suite 206
Washington, D.C. 20005-4018
(202) 628-4888
contracts@hrccourtreporters.com

LIBRARY OF CONGRESS
UNITED STATES COPYRIGHT OFFICE

In the Matter of,)
)
COPYRIGHT AND ARTIFICIAL)
INTELLIGENCE LITERARY WORKS,)
INCLUDING SOFTWARE, LISTENING)
SESSION)

Suite 206
Heritage Reporting Corporation
1220 L Street, NW
Washington, D.C.

Wednesday,
April 19, 2023

The parties met remotely, pursuant to the notice, at 1:00 p.m.

ATTENDEES:

ANDREW FOGLIA, U.S. Copyright Office
MARK GRAY, U.S. Copyright Office
SHIRA PERLMUTTER, U.S. Copyright Office
JENÉE IYER, U.S. Copyright Office
RACHEL BROOKE, Authors Alliance
ALI STERNBURG, Computer and Communications Industry Association
KEITH KUPFERSCHMID, Copyright Alliance
MATTHEW SAG, Emory University
LEIGH HENNIG, Humanity in Fiction
JOHNATHAN BAND, Library Copyright Alliance
JULE SIGALL, Microsoft
EDWARD HASBROUCK, National Writers Union
MARY RASENBERGER, The Authors Guild
CHRIS CALLISON-BURCH, University of Pennsylvania
HEATHER WALTERS, U.S. Copyright Office
BRANDY KARL, U.S. Copyright Office
KEYANA A. PUSEY, U.S. Copyright Office
SY DAMLE, Andreessen Horowitz
TERRY HART, Association of American Publishers
TRACY CHABALA, Author
CATHERINE ZALLER ROWLAND, Copyright Clearance Center
DEREK SLATER, Creative Commons
PETER ROUTHIER, Internet Archive
ATTENDEES: (Cont'd.)

CYNTHIA ARATO, News Media Alliance
BETSY ROSENBLATT, Organization for Transformative Works
CHRIS MOHR, Software & Information Industry Association
MEHTAB KHAN, Yale Law School

P R O C E E D I N G S

(1:00 p.m.)

MR. FOGLIA: Hello, everyone, and thank you for joining us for the U.S. Copyright Office's Listening Session on AI and Literary Works. I'm Andrew Foglia, Deputy Director of Policy and International Affairs. My co-host today is Assistant General Counsel Mark Gray, and we will be joined by other moderators. But, to kick off today's listening session, it is my pleasure to introduce Shira Perlmutter, Register of Copyrights and Director of the U.S. Copyright Office.

MS. PERLMUTTER: Welcome to the Copyright Office's Listening Session on AI and Literary Works, Including Software and Print Journalism. This is the first of four listening sessions we're holding this spring all focused on the copyright law and policy issues arising from the training and use of artificial intelligence.

As you know, AI developments are rapid and are now reported in the mainstream media virtually every day. Everyone is talking about the astonishing capabilities and potential ramifications of the newest generative AI and what it will mean for society, and copyright is an important part. How does current law

apply? Should it be changed? And how will the copyright community, from creators to users, be impacted?

The Copyright Office has a role to play both in addressing practical concerns and in advising on policy. Our listening sessions are part of a larger AI initiative that will continue well beyond this spring. We're analyzing the issues, helping claimants in registering works that incorporate AI-generated material, and establishing a process for gathering information to guide future policy decisions.

Today's session focuses on literary works, and we've seen the remarkable texts that large language models or coding assistants can produce and heard concerns from journalists, authors, and publishers about what the training and deployment of AI will mean for their industries.

There was tremendous public interest in participating in today's session, and we have more than a thousand registered online. Unfortunately, we weren't able to accommodate all of the requests to speak, but this isn't the last chance to share your views on AI with the Copyright Office. There are the three additional listening sessions, and we will be soliciting written input in the coming months. We

encourage everyone who's interested to submit comments.

Let me thank our panelists in advance for contributing to today's conversation. This is a complex topic and a deeply important and personal one for our participants, whether they're users or developers of AI technology, artists whose work helps train that technology, or creators contemplating how AI will affect their careers. Each of your perspectives is critical in informing sound public policy, and we look forward to an enlightening discussion.

So let me now turn the proceedings back over to Andrew.

MR. FOGLIA: Thanks, Shira.

As Shira mentioned, today's listening session is the first in a series of AI listening sessions that the Office has scheduled over the next six weeks. Future sessions will have different topics, different panels, different formats. On May 2, we'll be hosting a listening session on AI and visual art. On May 17, AI and audiovisual works, including movies and video game. And our final session on May 31 will be about AI and music and sound recordings. You can sign up to attend these sessions

at copyright.gov/AI, and speaker registration remains open for the last two sessions.

These listening sessions are going to inform further steps in the Office's AI initiative. Questions our panelists raise may be ones on which we seek written comments later this year. So please take note that in addition to the folks you see on camera today, the whole Copyright Office is listening.

With that, I'll turn it over to Mark Gray for some housekeeping.

MR. GRAY: Thank you very much, Andrew.

Hi, everyone. Before we get started, I just have a few housekeeping notes for everyone. First, for those of us who are joining as panelists but not in this first session, please keep your camera turned off and your microphone off until your session begins. Likewise, for those in this first session, once we change sessions after the break, please do the same as well so everyone on screen is someone actively speaking.

Second of all, for those of you in the audience, we are recording this session today. This will be available on our website in a few weeks, so please stay tuned. There will be a recording that lives at this event later in the past. And for those

of you who are interested in captions, we have activated Zoom's transcription function for today.

So today's panels are going to start with a brief introduction and short statement by each participant if that participant wants. We'd like to ask everyone to limit your statements to three minutes. We will have moderators watching the time.

Once the introductions are completed, we will have a moderated listening session. There are some questions from the moderators that the panelists have received in advance. Those are intended only to guide and to spark discussion, but we welcome participants to share any relevant perspectives or experience that they think is important for the Office to hear.

And so, with that, I will hand it over to our moderators for the first session, which are my colleagues, Jenée Iyer and Heather Walters. Jenee is a counsel in our Office of Policy and International Affairs, and Heather is our Barbara A. Ringer Fellow.

Jenée, the mic is yours.

MS. IYER: Thank you so much, Mark.

And welcome again, everyone. So we will begin our introductions in the order that is stated on the agenda, so with that, Authors Alliance, would you

like to go first and introduce yourself?

MS. BROOKE: Thank you. Good afternoon, everyone. My name is Rachel Brooke, and I'm a senior staff attorney with Authors Alliance, a nonprofit membership-based organization that exists to advance the interests of authors who want to serve the public good by sharing their creations broadly.

Today, I'd like to emphasize the potential for generative AI programs to support authorship by first increasing efficiency in some of the practical aspects of being a working author and, second and more importantly, by aiding in the creation of new works of authorship.

In the first category, generative AI programs can support authors by, for example, helping them create text for pitch letters, produce copy for their professional websites, and develop marketing strategies. Making these activities more efficient frees up time for authors to focus on their writing.

In the second category, generative AI has tremendous potential to help authors come up with new ideas, develop characters, summarize their writings, and perform early-stage edits to manuscripts. Moreover, and particularly for academic authors, generative AI can be an effective research tool for

authors seeking to learn from a large corpus of texts.

These programs undoubtedly have the potential to serve as powerful creative tools that support authorship in these ways and more, but it's important to remember just how new these technologies are, and because generative AI remains in its infancy and the costs and benefits for different segments of the creative industry have yet to be seen, in our view, it is sensible to preserve the development of these tools before crafting legal solutions to problems they might pose.

In fact, Copyright already has the tools to deal with many of the issues we'll speak about today. When generative AI outputs look too much like the copyrighted inputs that they're trained on, a substantial similarity test can be used to assess claims of copyright infringement to vindicate authors' exclusive rights in their works. In any case, in order for generative AI programs to be effective creative tools, it's necessary that they are trained on large corpora. The holdings in Google Books and HathiTrust indicate that it's consistent with their use to build a large corpus of works, including works protected by copyright, for research purposes.

And the question of the copyright status of

text created by generative AI programs is an important one. Authors Alliance agrees with the Copyright Office's recent guidance regarding registration of AI-generated works, and we believe that under ordinary copyright principles, the lack of human authorship means these texts are not protected by copyright.

This being said, we do recognize that there may be challenges in reconciling existing copyright principles with these new types of works, but, again, while this technology is still in its early stages, it serves the core purpose of copyright, incentivizing creativity and furthering the progress of science and useful arts to allow these systems to develop and confront new legal challenges as they emerge. Thank you.

MS. IYER: Thank you very much.

And CCIA?

MS. STERNBURG: Hi. My name is Ali Sternburg, and I'm Vice President of Information Policy at the Computer and Communications Industry Association, CCIA, a more than 50-year-old nonprofit trade association of internet and technology companies.

CCIA, our members, and their users have a significant interest in ensuring that new and emerging

types of AI-related creativity and technology are fostered rather than hindered by the U.S. copyright system. Courts have regularly and successfully applied longstanding technology-neutral copyright precedent to new technology, ensuring that progress is promoted and not stifled, consistent with the purpose of copyright law.

Settled copyright doctrines like fair use, substantial similarity, and authorship are well-balanced and flexible enough to keep up with new innovation and technology. Attempts to regulate these technologies in their very early stages could have significant negative impacts on beneficial tools and models that are important for a variety of use cases, use cases like public health or other societal challenges.

I look forward to a productive discussion today to discuss the benefits of AI tools and applicability of existing copyright law and policy. Thanks so much.

MS. IYER: Thank you.

And Copyright Alliance.

MR. KUPFERSCHMID: Thank you. My name is Keith Kupferschmid. I am CEO of the Copyright Alliance. The Copyright Alliance represents the

copyright interests of individual creators and organizations across the spectrum of copyright disciplines. What unites our members is their reliance on copyright law to protect their rights in the creation and dissemination of copyrighted work for the public to enjoy, and the copyright law is critical not only to their success and prosperity but also to the short- and long-term success of the U.S. economy.

For my members, there is no bigger copyright issue these days than the impact of artificial intelligence. Every single one of my members is interested and concerned about AI's impact on copyright.

Now let me begin by making clear that the Copyright Alliance supports the responsible and ethical advancement of AI technology. Many in the creative industry are already using or planning to use AI for the creation of a wide range of works that benefits society.

As with many advances in technology, new opportunities also come with new challenges. Advancements in AI have led to difficult legal questions surrounding the ingestion of copyrighted works into AI systems, legal liability for infringing outputs, and the copyrightability of the output. As

AI technology continues to evolve and questions arise about how copyright laws apply to the creation of AI-generated works, it's critical that when the Copyright Office makes determinations about AI policies, that the underlying goals and purposes of our copyright system are upheld and that the rights of creators and copyright owners are respected.

The interests of those using copyrighted materials for AI ingestion purposes must not be prioritized over the rights and interests of creators and copyright owners. Small- and large-scale creators produce high-quality works. These works are often ideal for ingestion by AI machines to generate output because of that high quality, and that is why AI developers want to use and copy them. So it should come as no surprise that there's already a high demand for large amounts of copyrighted work for AI ingestion purposes. Importantly, copyright owners are meeting those demands by entering into voluntary license agreements for TDM use.

However, both small and large creators face significant risk of being harmed when their works are copied without their authority for ingestion purposes. In particular, individual creators, who have little to no negotiating power with AI system developers, are

most at risk of such harms. All these issues are playing out in real time in other fora. There are numerous court cases pending that will shape how copyright law applies to AI, and many federal agencies, and, of course, the U.S. Copyright Office are also reviewing these issues.

These cases and these reviews are in their very early stages, so at this stage, we should all proceed cautiously and thoughtfully and let our guiding principle be one of respect for creators' rights and longstanding principles of copyright law. Thank you very much.

MS. IYER: Thank you.

And from Emory University, Matthew Sag.

MR. SAG: Hi, everyone. Thanks for the Copyright Office for organizing this. So I really appreciate this opportunity. My name's Matthew Sag. I'm a professor of artificial -- sorry, a professor of law in artificial intelligence, data science, and machine learning. I just want to make a few quick points.

This technology is new and exciting, but many of the legal issues are not new. The test for infringement is copying in fact and substantial similarity, and that remains the same no matter how a

work is created.

The copying required to collect the training data for these large language models is a classic form of non-expressive use that was upheld as fair use in iParadigms, Google Books, and, of course, HathiTrust. What's different is that it's possible that because of the size of large language models being used for generative AI that they can actually memorize the training data sufficiently to produce infringing works. That is a really interesting and important development.

What I propose is a constructive role for the Copyright Office not in rewriting existing copyright law but in promulgating best standards or best practices very much modeled on the fair use best practices which have been so successful. These best practices would give guidance to people training large language models, and they'd be focused on ways to avoid infringing output.

I'm going to be developing these things in a forthcoming paper, but I would just say quickly, you know, there are basic steps about de-duplication, about abstracting key words and tags, about controlling the ratio of the training data to the model size that will have a very important effect on

the likelihood that these models actually in the wild generate infringing works, and, for me, I think that's where our attention should be focused, and I think there's a good opportunity for the Copyright Office to play an information leadership role here, and I look forward to your comments. Thank you very much.

MS. IYER: Thank you.

And Humanity in Fiction.

MR. HENNIG: Hi. My name is Leigh Hennig. I'm speaking on behalf of Humanity in Fiction as the founder. We are an open advocacy group of authors, editors, publishers, academics, and others concerned with the ethical development of AI in creative spaces. I'm a published author, and I've spent much of my career working with automation at scale from the technology side. I'm also deeply involved and embedded with the speculative fiction-writing communities and have a keen understanding and awareness of the concerns within those communities as they relate to AI-generative text, such as chatGPT and GPT-4.

We recognize a number of ways in which AI can benefit our creative processes, such as with world building, character creation, with prompts to help move past creative log jams that everyone knows as

writer's block, and we're certain that there will be other additional creative uses for these tools.

We want to be clear that we don't see AI generative text as an enemy or a threat provided that its development is seen in a responsible and ethical manner. Like the invention of the camera and the impact it had on painters and illustrators, we think that AI can and should have a place in our creative processes.

President Biden recently called on tech companies to ensure their products are safe before making them public, using social media as an example of "the harm that powerful technologies can do without the right safeguards in place." The President, we believe, was as right as he was wrong. Without the right safeguards, this kind of technology has the potential for great harm. But it's not companies that we should be calling on for implementing safeguards. It's our regulatory bodies.

Corporations will not prioritize making their products safe. They will prioritize what will make money, as they have always done. If that happens to align with a public benefit, then fantastic. But what we need is thoughtful regulation considering the voices of artists and publishers, who are often seen

as very small and overridden in many of these discussions, and we need to consider everybody that lacks the powerful of -- excuse me, lacks the powerful lobbying arms behind them.

We don't fear that -- we don't feel that this is anti-corporate fear-mongering, and according to a report published on March 14 by the tech journal, *Ars Technica*, an entire team responsible for making sure that Microsoft's AI products are shipped with safeguards to mitigate social harms was eliminated. This has raised alarms within communities that follow such ethical developments, such as Emily Bender, a University of Washington expert on computational linguistics and ethical issues in natural language processing, which the report goes on to quote.

Twelve months ago, chatGPT was not part of our public discourse in the way that it is today. The velocity of innovation is as astounding and exciting as it is concerning, and we applaud the early attention given to this technology by the U.S. Copyright Office and the other organizations represented here today. We further encourage --

MS. IYER: Thank you, and at that, I'm going to have to pause as we hit the three-minute mark.

MR. HENNIG: Apologies.

MS. IYER: Thank you, and we look forward to hearing more from you during the discussion sessions.

Library Copyright Alliance, please.

MR. BAND: Thank you very much. I'm Johnathan Band. I represent the Library Copyright Alliance, which includes the American Library Association and the Association of Research Libraries.

So I just have four brief points. First, generative AI promises to be an amazing research tool of great benefit to librarians, students, academics, and all kinds of other creators, including, of great relevance to this group, lawyers.

Two, generative AI poses interesting copyright issues, but the U.S. copyright framework is flexible and robust enough to address these issues and, as others have mentioned, it is already in the process of doing so.

Three, because of the enormous benefits of generative AI to creators and users alike, the courts and the Copyright Office should apply existing doctrines in a generous manner so as to foster the growth of AI, in other words, to foster it and not erect roadblocks.

And, finally, a discussion of copyright legislation relating to AI is premature. We are in

the early days of AI, and the problems that may arise in the future are completely speculative. Thank you very much.

MS. IYER: Thank you.

And Microsoft.

MR. SIGALL: Thank you. My name is Jule Sigall, and I am Associate General Counsel at Microsoft. Thank you for affording us the opportunity to participate in these listening sessions. The Copyright Office should be commended for convening these sessions and exploring these timely and important topics. I'd like to make some observations about AI and mention some principles that we at Microsoft are using to consider the issues raised by AI on copyright.

The first observation, AI may well represent the most consequential technology advance of our lifetime. Today's cutting-edge AI is a powerful tool for advancing critical thinking and stimulating creative expression. It makes it possible not only to search for information but to seek answers to questions. It makes it easier for us to express what we learn more quickly. In the words of the copyright clause, it has the potential to "promote the progress of science, the spread of knowledge, to more people

and in more useful ways than ever before."

Second, the new generative AI tools are being adopted very quickly by hundreds of millions of people worldwide. Practically every corner of work and play is figuring out how AI can help improve the way we get things done, and all of this change is occurring at a rapid pace.

Third, authors and creators are also adopting these new technologies in their expressive work in all fields of creativity. In particular, the new generative AI tools offer individual creators the ability to express themselves in ways they could not before and are opening up creative expression to people who might have never thought of themselves as creators without this technology.

Finally, at Microsoft, we use the metaphor of co-pilot for these new AI technologies. They will sit alongside a human to help them create, analyze, learn, and understand, and a human will remain at the center of that activity.

Of course, such rapid change raises many questions about the impact of AI, especially in the copyright communities. Microsoft is committed to building and using AI in a responsible and ethical way. Here are some principles that we use at

Microsoft when thinking about AI and copyright.

First, AI tools and users must respect copyright. Authors and creators have the rights afforded to them by copyright and other IP laws, and these laws must be respected when developing AI and AI applications.

Second, the public has the right to extract knowledge from copyrighted works, to read, to learn, to understand, and to develop ways to create new works. The public also has the right to use new technologies like AI to develop and advance their knowledge.

Third, AI tools must benefit society broadly, not narrowly. The economic benefits of AI should be broad and inclusive, and authors and creators should meaningfully participate in them. For example, creators should be able to use AI to help them create new works and should receive copyright protections for works created using new technologies.

These principles can be in tension with each other, especially where the new technologies change the status quo, but it's important that copyright law and policy be interpreted and developed to promote each of them as much as possible. Microsoft is committed to forming new and deeper partnerships with

the creative communities, civil society, academia, governments, and industry to make progress in that development. Thank you.

MS. IYER: Thank you very much.

And the National Writers Union.

MR. HASBROUCK: Thank you. My name is Edward Hasbrouck. I'm a freelance and self-published independent journalist, book author, and web content creator. I'm speaking as a member and volunteer for the National Writers Union, which includes writers in all genres and media. The NWU's digital media division, the Freelance Solidarity Project, who you may hear from at later sessions, also includes creators of digital graphics, audio, and video.

Our members have created works which have been scraped from the internet, copied, and used for training generative AI without permission or payment and without respect for our moral rights. The NWU sees (1) moral rights, (2) the right to organizing and collective bargaining for freelance and self-published creators, and (3) feasible and affordable registration of web content as prerequisites for protection of our rights as creators of works used to train generative AI.

Training of AI language models begins with

copying, which we believe has infringed our copyrights and has already deprived us of hundreds of millions of dollars in rightful revenues. The additional violation of our moral right of attribution makes it impossible to tell which of our works have been copied to train AI and thus frustrates redress for either the economic infringement or the violation of our moral right to object to use of our work to train AI to generate prejudicial content.

Even if copying of our work to train AI is fair use, we have the moral rights to attribution and to object to prejudicial use of our work. Congress need not wait for courts to resolve any doubt as to whether copying for AI training is fair use to create a means of redress for the massive ongoing violations of our moral rights. Generative AI reinforces the urgent need and treaty obligation for Congress to enact effective protection for our moral rights.

As for our economic rights, payments to authors are likely to require collective licensing, but the ability of millions of freelance and self-published creators whose work is used to train AI to bargain collectively with billion-dollar AI companies depends on our ability to organize and act collectively, which is significantly deterred by fear

that organizing by freelancers and self-publishers might be held to be an antitrust violation.

Congress could best facilitate organizing, collective bargaining, and collective licensing for AI training by explicitly clarifying the right of freelancers and self-publishers to organize and act collectively as workers, including but not limited to collective bargaining over the terms of collective licenses. We should not have to fear that we will be accused of violating antitrust laws if we seek to organize and act together to exercise our rights as writers and digital media workers. Thank you, and I look forward to addressing your questions.

MS. IYER: Thank you.

And The Authors Guild, please.

MS. RASENBERGER: Thank you. Hi. I am Mary Rasenberger. I'm CEO of The Authors Guild, the largest and oldest organization for professional writers in the country, with over 13,000 members and growing. Our members include all kinds of journalists, of book authors. They write virtually every kind of book you can think of both traditionally and self-published. Our mission, together with our foundation, is to protect the precarious profession of writing in order to protect our literary culture.

We recently conducted a survey of our members and of some other organizations on how authors are using generative AI today and also the ramifications that they think AI will have for their work and the writing profession. We received over 1700 responses. I will talk about that a little bit today. About a quarter already use AI in some capacity to help them in their writing and they find these tools very helpful. Very few, very, very few, however, only seven out of 1700, use AI-generated text in any meaningful capacity in their final published work.

I do want to say that we believe that AI technology can help more people write and all writers write better, including writers who suffer from disabilities, and this is fantastic. But we do need some very important guardrails put into place. We agree with the Copyright Office that it is important to not provide copyright to AI-generated work or elements of a work because doing so will incentivize the use of generative AI to replacing writers, and as a result, AI-generated works will flood the market, devaluing human-created works. Publishers and writers will feel compelled to turn to generative AI to produce texts to remain competitive because of the

speed and low cost of producing them, but AI-generated writing will never be able to replace human thought in writing, so let's not let that happen.

Also, the Authors Guild is in the middle of our periodic income survey, which was sent to over 200,000 writers in the U.S. We have almost 6,000 complete responses so far. Close to 90 percent of those authors said they believe they should be compensated for the use of their works to train generative AI, and of the other 10 percent, most of them did not know. I will discuss today the need for collective licensing to enable that. Writers need to be compensated for the use of their work. And I just want to also mention that I agree with everything that Edward just said and won't repeat any of it.

We also learned that --

MS. IYER: Thank you.

MS. RASENBERGER: -- most writers fear that AI --

MS. IYER: Thank you very much. Ms. Rasenberger, I'm going to pause you right there --

MS. RASENBERGER: Yeah.

MS. IYER: -- because we're at the three-minute mark.

MS. RASENBERGER: Okay.

MS. IYER: And I'm going to welcome our last panelists to introduce themselves from the University of Pennsylvania.

MR. CALLISON-BURCH: Hi, everyone. My name is Chris Callison-Burch. I'm an Associate Professor of Computer and Information Science at the University of Pennsylvania. I've been working in the field of natural language processing, which is the sub-field of AI that's most relevant for these generative technologies, for more than 20 years. I have more than a hundred publications in the area that have been cited 20,000 times. At the University of Pennsylvania, I teach courses in artificial intelligence, natural language processing, and a seminar course on interactive fiction and text generation.

I've been using large language models and GPT in particular since June of 2021, when I had early access to the open AI API. My most recent Ph.D. student, Daphne Ippolito, did her Ph.D. thesis focused on creative writing applications of large language models. During her Ph.D., she worked at Google Brain in the Magenta Group, which is focused on creative applications of AI technologies for things like creating writing and music generation.

So, with respect to copyright and artificial intelligence, I think there are two distinct cases that the Copyright Office should consider giving updated guidance about. First is whether the use of AI to create creative works should exclude those works from being copyrightable, and second is whether using copyrighted works to train AI systems without the affirmative consent of the copyright holder should be considered fair use.

So I'd like to argue that creative works produced using a generative AI system should be copyrightable by a human because human users of the system will tend to perform selection and arrangement of the generated output in a non-trivial way, and I believe that the guidance from the Copyright Office dated March 16 of this year underestimates the level of interaction that human authors tend to have with the AI systems, and so I believe that there's more creative input from people than is currently given credit.

For the second point regarding guidance on whether training AI systems on copyright materials without affirmative consent from the right holders should be considered fair use, I'd like to argue that it should be considered fair use because, first, fair

use or learning, if the Copyright Office decided that it were not fair use, then that would make training of these AI systems effectively impossible and would shut down this interesting development. Second, the learning process is transformative. And, third, although there may be memorization of some of the materials, I believe that there are technological solutions to minimize that. Thank you.

MS. IYER: Thank you very much, and thank you very much again to everyone joining us today, and as we get ready to move into the discussion portion of our listening session, I kindly ask that those of you who are panelists in this first session to please turn your cameras on, and when we present some of our first questions, if you have a response, please do use the Raise Hand function. That helps us make sure that everybody who would like to speak has a chance to.

And so, with that, I'd like to present our -- we'll open the discussion with a question of, what artificial intelligence technologies are you or others in your industry using in the creation of new works? And if anybody would like to kick us off? Ms. Rasenberger?

MS. RASENBERGER: Thank you. I can speak to the tools that writers are using. GPT is the most

commonly used to help writers in the writing process, with 50 percent of those who use generative AI saying they use chatGPT. GPT4 comes in second, as well as Bing. About 8 percent are using Google's Bard. Grammarly and other grammar tools are also used. Sudowrite, which is a platform based on 3.5 built to help in writing in novels, is also being used. It can be used for other writing besides novels. And Jasper, which focuses on business writing but can be used for anything, is also a popular tool.

There are, I will say, dozens of other writing tools on the market already. They are all based on GPT 3.5 or 4 because back in '20 and then '21 OpenAI opened to the public a portal to allow others to connect to GPT by an API and develop tools based on it. So, when you hear about writing tools, other than Bard, they're all based pretty much on GPT.

When we surveyed our writers, about 30 percent said that they use AI to help with brainstorming, and this is a really interesting way to use it. It helps with plot, ideas, character-setting, to develop ideas. It also helps some writers to structure and organize drafts. About a quarter say they use generative AI for marketing ideas and communications, and 50 percent use AI to help with

grammar and writing to perfect their sentences.

Only 7 percent said that they use generative AI to generate actual text other than for marketing, and approximately 90 percent of those who use generative AI tools to generate text reported that less than 10 percent of the final work comprised generative AI. Only seven writers out of the 1700 responded that their work comprised 50 percent or more AI-generated text.

Almost all of the authors we surveyed, and we followed up with a couple hundred who do use generative AI, they all said that it is important to them to have writing in their own voice and their own expression because that is what they do. And as the Office thinks through copyrightability issues, I think that's important to bear in mind.

Writers also reported that businesses, their clients have turned to using AI, such as Jasper, to replace humans for writing branding, marketing, and web copy and that their work in that area has dried up. One writer said that she has lost 75 percent of her income already, so we are already seeing the impact.

I should note that our surveys and interviews were with professional writers. There are

also others using AI now to write stories and books, and you will see lots of videos, social media about how to get rich quick writing fast with AI, which we all laugh at because we know few people get rich writing, but these people do rely heavily on AI-generated text and those books are now mostly self-published on Amazon's Kindle platform. Thank you.

MS. IYER: Thank you.

Ali?

MS. STERNBURG: Sure. I just wanted to mention a few other kinds of tools that may be used in this space, including -- just also that show the breadth of what can be considered an AI technology, things like translation tools, translating into other languages, speech recognition, computational photography. There are types of direct AI tools or toolkits that can be used to create other works, so, yeah, I just wanted to mention a few other types of technologies. Obviously, this session is literary works. It's also literary works including software, so there's, yeah, a wide range of type of tools that I'm sure other colleagues will mention as well. Thank you.

MS. IYER: Thank you.

Johnathan?

MR. BAND: Yeah, very briefly. It seems that as with other new types of technology and their interaction with creators, there's going to be a generational issue, so we're going to have generative AI with a generational impact. In other words, I think that, as we've seen with other technologies, younger people, people newer in the field, are more likely to be able to use the new technologies and adapt to them and figure out ways in which they'll help. Those of us who are older and have more established ways of doing things will find it harder to use these new technologies.

And, you know, certainly, you know, the Copyright Office and other entities can help with education and training, but I think, at the end of the day, it's just, as we've seen within photography and other areas, it's just sort of a fact of life that younger writers are going to be able to assimilate these new technologies more quickly than those of us who are older. Thank you.

MS. IYER: Thank you. And we'll go to Jule and Keith, and then I'll pass it over to Heather for a second question. Jule?

MR. SIGALL: Thank you. I think I just wanted to highlight probably the most common AI tool

used in the software industry today is developer aids like GitHub's Copilot, which sort of sits alongside inside the development environment for software developers and suggests code for them and reacts to the code they've written, the comments they put into their code and other prompts to help them develop code much more quickly and much more efficiently, especially around the more common sub-routines and function calls that they use to build code.

What's an interesting note about how it's being used is it seems to have a primary benefit for developers who are proficient in one particular software language being able to bridge that expertise into a new one where they may not be as familiar, but using a tool like Copilot, they can start generating code, running code much more quickly in that new language that they may not be as familiar with thanks to the interface that it provides to translating and developing code in a new language.

And I guess the last point to make is, as we are sort of reviewing the Copyright Office's guidance about what copyright you can register under that might have been developed involve using AI tools, it may be very problematic in the coding space for a couple reasons. I don't think the line between what is the

developer's and what is suggested by the AI is very clear, yet I think almost every developer at the end of the process, when they have some code that they are trying to run and deploy, none of them would think that they are not the creator and author for copyright purposes of that code, and, in fact, they often rely on that to invoke open-source licenses and other things in the industry. So I think there may be some challenges with the way the Copyright Office at least has articulated what you can register when it comes to code under the guidance that's been published.

MS. IYER: Thank you.

And Keith?

MR. KUPFERSCHMID: Yeah, thank you. Just one general comment in response to something Johnathan said. Really, I want to just go on the record as saying I don't agree with what he said. It sort of smacked of ageism. I don't think age is a dividing line between who can use a new technology and who can't, and so just I don't really agree with what Johnathan said. I do wholeheartedly agree with a lot of what Mary said in terms of how our experiences in terms of how creators are using new AI technologies.

I do want to also be careful to make sure we kind of limit or focus what we're going to talk about

today and that we make sure we talk about generative AI which is used for generative purposes. Like things like just correcting grammar, spell check, or even something that just potentially does a translation that may or may not be on the line there, whether that's being used for generative purposes, you know, remains to be seen.

The question also, I'm assuming what you were asking in terms of the scope of the question that you are interested in knowing about how people are using AI to generate, you know, fully generative output as opposed to using it in the workflow. I'm assuming that to be the case, but like I said, as Mary said, at this point, many creators are not using it, but they are beginning to see that change, and to the extent they're using it, they're using it for ideation and to help with writer's block and things like that, and Mary went through some of the other different technologies that people are using.

I will put out an offer. We have a lot of creators who are members at the Copyright Alliance, and we're happy to bring them in or just do demos somehow to show exactly how they're using it if you think that would be helpful.

MS. IYER: Thank you very much, Keith, and

some of the questions and points you made actually flow right into our next question, which touches on generative works, so I'm going to pass the mic over to Heather, and then I see, Ali and Chris, you have your hands raised, so if you'd like to also touch back on your thoughts there and also responding to the question that Heather's going to pose, then we can move forward. So, Heather, the mic is all yours.

MS. WALTERS: Thank you, Jenee.

For our next question, what do you think the Copyright Office should know about how AI systems generate literary material, whether that's fiction, non-fiction, or code?

MR. CALLISON-BURCH: I'd be happy to kick us off. So I think one of the interesting elements of this that touches on the current copyright regulation is the idea that it's mechanically produced, so generative AI systems are pre-trained on huge amounts of textual data if we're talking about large language models or images, and then they're sampling from their probability distribution in a way that could be construed as a random process that's explicitly excluded by copyright law.

However, I think that in order to generate a particular work, you need a prompt and you need a

random feed and you need a model, and so I think that it should be relatively straightforward for the Copyright Office to ask human authors to demonstrate that they've done substantial selection or modification by providing the original piece that the system generated or simply the prompt, random feed, and model to check to see whether the work that they produce is substantially different than this automatically generated and reproducible process.

And then a brief comment on Keith's point about generative AI excluding translation systems and grammar correction software. So, just to be clear, those are also generative AI that are pre-trained in exactly the same way that all these other technologies are, and the stymieing effect of excluding copyrighted works from training such systems would also have an impact on translation systems and on grammar correctors.

MS. WALTERS: Thank you.

Next up, Ali, and then Edward.

MS. RASENBERGER: Oh, sure. I know there are others ahead of me now. Yeah, I was also just going to -- Chris did a great job explaining that. I was just going to ask about whether the question was about generative AI or AI generally because it's not

always clear, but that's a really great point from a technical export.

On to the current question, I would just add a few different points here. Different AI systems operate in fundamentally different ways, so as I'm sure we hear a lot in this context, it's not like a one-size-fits-all thing. For example, a large language model will have different mechanisms and constraints than a diffusion model and both will differ from a convolution system. This is from our AI experts. I don't have any idea what any of that means.

The Office should endeavor to understand these differences as they may affect Office guidance here, but, generally, however, an AI synthesizes information from its input materials similar to how a human might learn from existing creative works, and it's also creating new material. It's not, like, making a mash-up of existing text. It is doing something entirely new. Thank you.

MS. WALTERS: Edward?

MR. HASBROUCK: Yeah, I think the most important fact for you to keep in mind about generative AI is that it depends entirely on copying valuable training material as part of the input from

which to generate any valuable output. Without that training material and the prompts provided by users, generative AI would generate only garbage. Never has the axiom of computer science "garbage in/garbage out" been more apt, and I think it's an open question how much of the value of the output is attributable to the training material, how much to the prompt, and how much to the AI software.

That division of revenues should be a matter for negotiation between creators and those who want to use our work to train AI, but, clearly, the value contributed by the training material is more than zero, which is what we've been paid to date. OpenAI, for example, has received a billion dollars in venture capital, none of which has been passed on to the authors of the training corpus even though, without that training corpus, chatGPT would be worthless.

Even if the software and the prompts has contributed as much to the value of the output as the training material, creators of the training works should already have received half of that billion dollars. Creators of works infringed by copying our work for AI training have already been deprived of hundreds of millions of dollars to which we are rightfully entitled from OpenAI alone.

It's also important to recognize that the works most likely to have been copied to train AI are those on publicly accessible websites and to focus on that work and on the creators of that web content. You know, the low-hanging fruit for unauthorized copying to train AI language models is publicly accessible web content, not works available only in print. We remind the Copyright Office that decades after the creation of the Worldwide Web you still haven't created any procedure for registering most web content, especially large and/or dynamic websites that isn't prohibitively burdensome.

The violations of copyright by copying web content to train AI heighten the urgency of making it possible to register copyright in these works, and we again implore the Copyright Office, as long as prohibited registration formalities are retained, to implement a realistically feasible and affordable procedure for bulk registration of web content.

MS. WALTERS: Matthew and then Mary.

MR. SAG: Yeah. So the question is what should the Copyright Office understand about generative AI in this context. What I would want to make clear is that things like GPT, et cetera, are fundamentally predictive tools, so a model like GPT

has no internal mental state, but it's a very long equation that does effectively model latent relationships in the training data.

Why is this important? It's important that you understand that when a model like this produces some text, it's not just making a collage of a few different existing texts. It's recombining the things that it has learned from the training data at a much more abstract level and then combining those, the things it's learned about the structure of language, narrative form, et cetera, and so the link between the training data and the outputs, which, you know, definitely exists, is significantly attenuated. It's attenuated by this process of abstraction. It's attenuated by the remixing of latent concepts at this very abstract level, and it's also attenuated in weird ways by the way random noise is used in the training process. So I think that's one thing at least the Copyright Office should understand. Thank you.

MS. WALTERS: Thank you.

Mary?

MS. RASENBERGER: Thank you. As you know, LLMs are trained on works of the same nature that they are programmed to generate, so GPT and Bard, the two main engines, were developed by copying and ingesting

large amounts of text, including potentially millions of books and articles found online without permission, and in generating the outputs, these programs merely re-scramble inputs. Nothing new is added. Generative AI cannot think or feel itself. It cannot express emotion. It can only mimic what it has been fed.

And so, by its nature, it is always derivative of what it's been trained on. There would be no GPT without the pre-existing works. OpenAI on its website says that the GPT 3 training data sets included text posted on the internet or uploaded to the internet and also two internet-based book corpora referred to by OpenAI as Books1 and Books2 without any further explanation. Researchers have attempted to recreate the data, and they have reported that Books1 is Books corpus which maintains that it's free books scraped from smashwords.com.

Books2, no one knows exactly what this is, but it must be massive as it appears to include pretty much every published book that you look for, that you try to get it to mimic. It is highly unlikely as a result to be legitimate, and some suspect that it's Libgen, which is one of the major piracy sites, book piracy sites. In any event, most books and most articles online and other works on the internet have

been copied and used to train GPT to generate text. We believe that this use is not fair and that it should be compensated. We do not, however, want to impede the development of AI, so we would like to see collective licensing that makes it possible for AI developers to license the data they need and compensate authors.

We believe that writers should be compensated also for past training since it appears that the massive training that has already occurred for GPT and Bard to teach the engines to think and to write has already occurred, and as Edward mentioned earlier, we may need an antitrust exemption to do it as effectively as possible, and we do hope that the AI companies will come to the table.

I also should mention that there are some writers who simply do not want their work used as training data, including when they use AI. They do not want anything that they upload or generate to be used to train AI, and we believe that those writers should have the right to opt out.

MS. WALTERS: Thank you.

Leigh?

MR. HENNIG: Thank you. I'd like to say that I agree with Edward. We conducted our own survey

of over a hundred editors and publishers, and nearly all were concerned with the use of their works without at least attribution. One issue that Chris raises, and I have to say I disagree there as well, I understand that his point is that people are interacting with these AI tools in a much more in-depth way, and his point is that they're interacting heavily with them for their output and not just running with whatever comes out of a single prompt or two, and while I agree that this is often true, we're also finding that this really isn't necessarily the case at all times or even most of the times at least as far as what editors and publishers working through their slush piles are seeing.

Editors are getting crushed by the massive increase of submissions which are cheaply generated and just sent in oftentimes by people looking to take advantage of get-rich-quick schemes. *Clarkesworld*, as an example, recently had to close, and that was so remarkable and shocking to the speculative fiction world that it actually made the news. So this is especially impactful as well to marginalize peoples who already face an uphill battle in getting their work seen and recognized.

So, you know, I think that the issue is

Chris's point is well taken, but I want to make sure that we're not completely overlooking the impact that these tools are having not only on writers but on editors and publishers.

And I also would like to say that I disagree as a final point with Matthew. While we've recently seen that AI-generative output is not wholly unique, there are a number of lawsuits currently underway because output that is generated by these things is too closely resembled to works that authors are coming out with. And as another example, CNET was recently in the news for the heat that they received for their articles which are generated by AIGT, essentially copying a lot of the stuff that they have.

So some tools are better than others, and to somebody else's point, large language models have different ways that they produce text, but I think it would behoove us to pay attention to the collateral damage that these tools are having even as we do recognize the benefits that they have to society and to creatives. Thank you.

MS. WALTERS: Thank you.

Jule and then Johnathan.

MR. SIGALL: Thank you. I'll be quick because both Professors Callison-Burch and Sag made

the points I wanted to make about how language models work only in the sense that they are not collecting pieces of the work and reassembling them in collage style. Instead, what they, in our view, are doing is capturing the unprotected elements of works, sort of the ideas, the concepts, the facts, and the relationships between those facts, concepts, and ideas in order to generate new works, not to actually reproduce existing works.

And, in fact, the memorization phenomenon that Professor Sag mentioned is an interesting one because, in most of the field of AI development research, that memorization is seen as a bug and not a feature in the sense that the whole point is not to provide access to the underlying material and the training data for various purposes, including privacy and other sensitivities, so the goal is to create a model that can take the elements of those works and create new works for them.

And, of course, as others have mentioned here, there are questions and important copyright questions to answer about the outputs of these systems and whether they infringe existing works, and those are an appropriate focus of where the copyright law and policy should apply. But the model itself is not

really designed to be a reproduction machine or to be a collage machine. It's designed to understand the core components of knowledge in existing works and make them accessible to people so that people can develop new things based on that knowledge.

MS. WALTERS: Thank you.

Johnathan?

MR. BAND: Yes. Just briefly in terms of what the Copyright Office should know about this technology is that, you know, following on from what both Edward and Mary said, a lot of the material that makes up the training corpus is material that has been uploaded to publicly accessible websites, and whereas, in some cases, it may have been done involuntarily, in the vast majority of cases, it's been done voluntarily by the author or whoever is the rights holder.

And Mary said that, you know, the rights holder should have the ability to opt. Well, of course, they do have the ability to opt out. They can use bot exclusion headers. The way this technology works, like the way search engine works, is you have bots that are crawling the internet and gathering information, downloading material from websites, and the websites can use bot exclusion headers, and that would prevent the prowling of their sites. So, again,

the technology exists to address this problem. Thank you.

MS. WALTERS: Thank you. I am going to pass it back to Jenee for the next question.

MS. IYER: Thank you. As some of the conversations have started to also talk a little bit about outputs, we'll pose our next question here, and those of you who have your hands raised, feel free to go ahead and continue your thoughts as they relate to the previous question and incorporate any response to this next question that you have.

So the next question is, how is the training or the output of artificial intelligence affecting your field or industry? And again, for those of you who already had your hand raised, feel free to continue your thoughts on the previous question, and we will start with Keith.

MR. KUPFERSCHMID: All right. I'm going to answer the last question because I had my hand up and for some reason wasn't selected, so I'm going to mention three things. Before I do that, let me just point out that that last question asked actually two different questions. It talked about whether it's okay -- it talked about the ingestion and the input and the output, if you will, and there was some talk

about output, which you mentioned. I'm going to focus on the input. I'm going to talk about three different things.

One, I'm going to raise to your attention the article that was in *The Washington Post* today. It refutes what Johnathan just said, okay? *The Washington Post* article does an analysis of chatbots that are using Google C_4 data sets, and what it does is it reveals that proprietary, personal, and often offensive websites go into the AI's ingestion of data, okay? Now I'm going to quote from the article here. It says, "High on the list, b-ok.org," which is No. 190 on the list, "is a notorious market for pirated e-books that has since been seized by the U.S. Department of Justice. At least 20 other sites identified by the U.S. Government as markets for piracy and counterfeits were present in the data sets." It also revealed that works with over 200 million copyright symbols were part of the data sets. So I suggest you take a look at that article.

Something else, a second thing you should know is about data laundering. Some AI developers have, without authority, used training data sets as pre-trained AI created by non-commercial third parties in their commercial products. That's known as data

laundering. Neither this kind of unauthorized use nor the work of the non-commercial entity necessarily qualify as fair use.

And that brings me to the last point I was going to mention, which is on fair use, okay? There's been some talk, without getting into the details, of whether ingestion -- like I said, we don't have to address output because, when you copy all these works, you are copying, you are infringing the copyright in those works. Nothing could be clearer. You can say, oh, we're copying relationships, we're copying data points. You're copying the works. What you use them for, that may be something else, okay?

Now, on fair use, obviously, it depends on a case-by-case basis. You have to look at each case, okay? But, certainly, I think you have to look at whether a TDM license is available, and they are available in many instances, whether the use is going to be commercial. We're seeing mostly commercial uses these days, and most importantly whether the resulting AI-generated work harms the actual or potential market for the ingested work. That gets into the output a little bit.

But let me address the Google Books case and some of these other cases that we're talking about

because I think, if there's anything the Copyright Office takes away from this listening session, it should be this: The Google Books case could not be more different than what we have going on here. In the Google Books case, Google did not copy books to make new books. That's what AI does, the copying works of expression. They were copying copyrighted works to make new copyrighted works that compete with the works that they are copying. In Google Books' case, Google used the works for informational purposes. They used it for the information in the works, not the expressive content of the works. That is exactly what AI is doing. They're using the expressive content to produce new works.

Google implemented numerous safeguards to ensure that the use did not harm authors or publishers. That is not what AI technology is doing. It's allowing you to put in the style of and it's allowing me to put in -- if I put in SpongeBob, then I get back SpongeBob images and things like that.

And lastly and very, very significantly, in the Google Books case, there was no licensing market for Google's use, but in AI case, there is a vibrant license for AI training materials that would be completely destroyed if AI use is considered to be

fair use.

So, in sum, the Google Books case I used as an example, but there are other similar cases, very, very different than what we have going on now, and no one should assume that just simply because, in those cases, the court held to be fair use that we would have fair use taking place here. And I'll stop there.

MS. IYER: Thank you. I just note I do see a couple of audience hands raised, and for the moment we are just calling on the panelists, so thank you.

And Mary?

MS. RASENBERGER: Thanks. I also just wanted to respond to Johnathan's point from the last question. It is not a matter of using robot text. First of all, you can't be crawled if you use robot text. You can't be searched, crawled for search purposes, and you'll hear from the news organizations about that probably in the next session. So it's not that simple. Also, we believe that much of the content that was crawled and used to train GPT is pirated. There are very few books available online without DRM that are not pirated.

To this question, our biggest concern with the use of generative AI is that if there are incentives to use generative AI to produce books and

other written material, publishers will feel compelled to use them, and that's because they are faster, cheaper, and this, in turn, will dumb down our literature and our journalism.

As an example, one of our members, she writes for corporations. She was asked by one to write 30 pieces a month instead of 10. She was told use AI and then just edit it, and as she responded, that's impossible because, if I start with something AI-generated, it will take twice as long because I will need to add my own voice and add thought and what AI produces is not just usable. But there will be pressure because that's the way capitalism works. There will be a need to stay competitive in markets.

We're very concerned about AI-generated works surrounding the market for certain kinds of books and journalism, making it harder for indie authors and traditional publishers to earn money in these sectors and decreasing the whole pie. We are already seeing a flood of AI-generated books in certain markets, namely, certain types of non-fiction books, such as self-help and cookbooks, children's books, and genre fiction, such as sci-fi. These are the very books that bring in profits that allow publishers to invest in other books that do not have

as broad a market but that are critical to literary and civic culture.

I also want to note that writers make a living from multiple writing sources, not just books. Most need to take on multiple jobs to support their book writing, freelance journalism, corporate writing and copy writing, and the most immediate effect of generative AI has been to replace human writers for web content writing, copy writing, marketing, newsletter writing, and other communications.

We have already heard from many writers who have lost a lot of their income to AI. Also, entry-level journalism jobs face significant risk as simple news articles are being generated by AI, and the problem is this is the first early career stepping stone for most writers. Most book authors do not get their first book published until well into their 30s or 40s, and they work in journalism and odd writing jobs until then, so we need to find a way to protect those jobs as well.

I do believe that we will see a loss of jobs because of AI. What we want to prevent, we have to prevent is a breaking down of the writing profession to the point where so few writers are able to enter the profession that we really see a decline in the

books that are published and the quality of the books, and I can talk more about that if we get to the next question. Thank you.

MS. IYER: Thank you, and I'm going to exert a little bit of moderator's privilege here. So, Rachel, I know you haven't had a chance to speak yet, so I'm going to toss the mic to you, and then to the remaining hands that are up, I'm going to request that you please hold your responses at about one minute so that we can get to the last question here and also be respectful of everybody's time here today.

So, Rachel, I'm going to toss it to you.

MS. BROOKE: Well, thanks very much. So I want to return a bit to the Office's question as to how the training or output of artificial intelligence is affecting our field or industry. So, as I mentioned in my introductory statement, generative AI is making it easier for authors, and while I mentioned academic authors, this, in fact, applies to all different types of authors, to do the early-stage research that's foundational to their writing, such as getting short and simple summaries of complex issues, surveying the landscape of various fields in which those authors don't have a strong background, or even getting guidance on what types of human-authored works

to turn to in their research, like a sort of bibliography-like output.

Making research more efficient means that these authors can spend more time on their writing and making valuable contributions to their fields. I'll say that Authors Alliance is really committed to protecting authors' rights to conduct research, and we see generative AI as a new, innovative, and, as I said multiple times, efficient tool of conducting this kind of research. Making research easier helps authors save time and has a particular benefit for authors with disabilities or other reasons that make it difficult to go to multiple libraries or otherwise rely on analog forms of research. So, in conclusion, I would say it has a strong benefit on what I see is our field or industry.

MS. IYER: Thank you.

And we're going to do next Chris, Edward, and then Ali with about -- if you could keep it to one minute, please, that would be much appreciated. Chris?

MR. CALLISON-BURCH: Thanks. I appreciate the impassioned pleas from my co-panelists. I wanted to just address Keith's -- he made two points that I think misrepresent -- or several points that I think

misrepresent what's happening, including discussing ingestation and data laundering, which are not technical terms. The copying is, of course, present in the same way it is for your web browser where, in order to display it on your computer screen, a work must be copied. This is non-expressive copying that does not violate copyright.

Secondly, he suggested that fair use should be determined on a case-by-case basis, but, in fact, I would argue exactly the opposite. Fair use should be determined as a general class of things, and if it is left to a case-by-case basis, then that opens -- that stymies this technology and makes it so that no company could develop anything without the permission from all authors.

Finally, this is just a bit of speculation, but I think it might be worth raising, especially for the speculative authors panel, where I suspect we may be hitting an escape velocity point where AI systems could be trained on the output of AI systems and, thus, if it were determined that they were not copyrightable, then we would be done and we would have a high-quality AI system not derived from human-generated text. Thank you.

MS. IYER: Thank you.

And I see we have a few more hands raised. After Edward and Ali, we will move to Heather and closing statements, so I'll invite you to make your comments then. Edward?

MR. HASBROUCK: Yeah. All members are already being adversely impacted by both the training and the output. As far as the training, as I earlier noted, that has deprived us of our rightful share of the income that AI developers have already obtained from our work.

And contrary to what Johnathan said, I want to push back further. Robots.txt or other such measures are not a solution. They are not enforceable. They're not actually practiced. The largest web scraper in the world, the Internet Archive, announced retroactively that years previously they had stopped across the board honoring Robots.txt files even when they included specification, not generically, but specifically intended to exclude the Internet Archive's crawler in the fashion the Internet Archive itself had specified. So it's not enforceable, it's not practiced, and it can't help you opt out if your work has already been scraped and ingested, so that's not a solution.

As far as the output, the output is being

used to substitute for our work, and just because the displacement of human creators is a one-for-many rather than a one-for-one substitution doesn't mean it isn't substituted for and an infringement of the works in the training corpus. The output is being used to generate formulaic news stories by using box scores as prompts to generate sports reports, using trading statistics as prompts to generate routine financial news, displacing freelance and staff journalists.

It's being used to generate web content and other marketing and business communications, displacing freelance business and advertising copy writers. So I think the effects are growing, but contrary to some earlier comments about future effects being speculative, the effects to date are already real and adverse on our incomes. Thank you.

MS. IYER: Thank you.

And Ali?

MS. STERNBURG: Thanks. Yeah, I'll just keep it brief. Just wanted to respond to two quick points. One was I don't think John specifically mentioned Robots.txt. There may other types of inclusion protocols in the works. I think one is part of the EU's most recent, I think, text and data mining assumption in the copyright directive maybe, so these

are things that may be in development anyways, especially for other regions that are similarly looking at AI and these issues.

And then also there was a point about styleries, and I just want to make sure. It sounds like something that likely wouldn't be protected by copyright. There's already a lot of limits on what can be protected and what cannot, so just did want to make sure we made that point. And then just in terms of the original question, I think just these types of tools can provide tools for authors but also can help them, for example, create some kind of illustration alongside. I know that's a thing that will be addressed more later, but it can help people that don't have all these different skill sets to supplement their work and don't have the technical skills. So, yeah, that's it. Thank you.

MS. IYER: Thank you, Ali.

And so, Leigh and Keith, we're going to circle right back to you after we throw it to Heather for a final question and the opportunity for closing statements. So, Heather?

MS. WALTERS: Thank you, Jenee.

For our closing question, are you aware of the Office's registration guidance with respect to

works containing AI-generated material, and, if so, what questions or concerns do you have about that guidance? And, again, if you have any closing statements, please include that. Leigh and then Keith.

MR. HENNIG: Thank you. We were aware of the guidance that the Copyright Office put out. One potential question that we would like to see considered is where exactly the delineation rests when considering works that should be eligible for copyright. For example, if there's an anthology or a collection of short stories and that contains a mix of human and AI-written stories, where does that stand?

I think, for my closing statement, there is a clear distinction in the utility and application of AI it relates to things like fiction writing versus research, coding, or academic work. As our own survey of fiction authors and editors have shown, many writers see AI tools as a benefit. I think that I would like to close by encouraging the Office and others here to consider how we can be flexible in our consideration and application of potential regulation. We don't want to completely close ourselves off to the ability to use AI, especially since the utility of it in, you know, again, those research and academic and

coding situations is very clear, but at the same time, we want to be respectful of how that impacts more creative functions, such as, you know, fiction writing and other such endeavors. Thank you.

MS. WALTERS: Thank you.

Keith?

MR. KUPFERSCHMID: Sure. Quickly, before I get to my other comments, in response to Ali's comment, to be clear, I was not talking about in the style of being protected by copyright. I was talking about that being a safeguard in the context of a fair use analysis. There is a difference. Read the case.

For Chris, I don't even know how to respond other than to channel my inner Chandler Bing and go, could you be more wrong?

And in terms of the Copyright Office's guidance, I've got about six pages of notes. Obviously, I cannot go through all those. There are a lot of concerns and confusion with the guidance. I'm going to go through a couple highlights, but I would love the opportunity to somehow be able to make sure that the Copyright Office is aware of these concerns that we've heard from all of our members.

Significantly, the Copyright Office should not and does not have the capacity, frankly, to be

engaged in investigations of what is within and outside the boundaries of what is disclaimed as AI-generated and whether there is sufficient human involvement in each case, as it did in the cash de nova (phonetic) case, and, hopefully, that is not the plan.

We get the idea that you do need to identify what is and what is not claimable, but there is confusion with the guidance in terms of, you know, where to draw that line. Further guidance would certainly be helpful. People are very -- I'll finish up. People are very concerned about the invalidation or cancelation of their registrations and people challenging these in courts.

And lastly, there are a lot, a lot of inconsistencies between the guidance and the compendium and how like the de minimis standard, which is inconsistent with the compendium, which talks about an appreciable amount. The standard application presents problems. Anyway, as I said, I can go on for days talking about some issues. I would love the opportunity to be able to do that. I thought this listening session would be such an opportunity, but, obviously, we've run out of time.

MS. WALTERS: Thank you.

Mary and then Johnathan.

MS. RASENBERGER: Thank you. I applaud the Office for drawing a clear line in the sand with the guidance even though some of the reactions have not been popular, but I think drawing this line is important for protecting the incentives for human authorship. In fact, I think it is absolutely crucial. Human authorship requires that a human conceived the work and executed or closely monitored the execution, and if you accept that AI-generated material is not copyrightable -- if you accept that it's not copyrightable, then you can't say, well, in this instance, because there was a lot of human work, some of it highly creative, that went into the prompting, the results should be copyrightable. I think you have to look at whether a human created the actual expression that's in question.

The Copyright Office does need to do a better job of explaining the guidance, though. There's a lot of confusion. First, it is not well understood that only AI-generated expression in the deposit copy needs to be disclaimed and that that can be done in a word or two, so if a writer is using AI as a tool to brainstorm and is not adopting AI-generated text or if they're using it for spell check

or grammar, the writer does not need to disclaim AI-generated material.

Second, if there is only some AI-generated material in a work, it can be disclaimed in a word or two as some text, some images, or all images or all text, so if a writer uses some sentences that AI generated in their final work, they can simply exclude some text. They do not need to identify exactly what it is that AI generated in the application because that's going to be impossible to keep track of as a practical matter, and that is also not understood.

Third, in conclusion --

MS. WALTERS: Excuse me just one moment.

MS. RASENBERGER: Yeah.

MS. WALTERS: We do have some hands raised, and as the session is going to end in about six minutes, I want to make sure everybody has a moment to speak before we break.

MS. RASENBERGER: Okay. Can I just -- I hope we have an opportunity then because I have some other things that I wanted to mention.

MS. WALTERS: Thank you.

MS. RASENBERGER: If we have an ability to follow up in writing. Thank you.

MS. WALTERS: Thank you.

MS. IYER: Yes, this is the first -- this is the beginning of the Office's initiative on AI, so there will be many forthcoming opportunities to continue to provide the Office with feedback, and we do appreciate all the feedback we've heard here today.

MS. WALTERS: Johnathan?

MR. BAND: Yes, thank you. So just with respect to the Copyright Office guidance, I think I agree with a lot of what Mary and Keith said, that in general it's on the right direction, but on the edges, there are issues that need to be examined, especially, as some of the others have mentioned, you know, that the Office may not have appreciated how much back and forth there is going to be between the author and the AI and that the delineation and the disclaiming could be pretty difficult.

With respect to closing, I'd just like to make two points. One is I'm glad that Keith is so supportive of the Google Books decision. I've never head him be so positive about it, certainly not while the case was pending. But then, more seriously, you know, the whole discussion that we had back and forth about bot exclusion headers shows two things. One is that there's obviously a lot to be examined. You know, I read the *Washington Post* article in a very

different way to suggest that, you know, if you look at the top sites, you know, like, you have Wikipedia.

I mean, you have a lot of -- that's where a lot of the information that's being included is coming from and other public domain sites, government sites and so forth, so that's something obviously for the Copyright Office to look at. To some extent, though, maybe it's really beside the point because, as some of the other people have indicated, that what is being extracted ultimately is non-protectable material.

Now, to be sure, you might have to make a temporary copy in order to extract unprotected material, but, you know, we have 30 years of case law about that, about sort of intermediate copying or, you know, copying that doesn't get its way into the final product, and so, you know, maybe all these issues with respect to the ingestion are even more settled than we thought, but, in any event, there's certainly a lot of things for the Copyright Office to look for, to look at going forward. Thank you.

MS. WALTERS: Thank you. And as we finish up the session, if everybody could keep their closing comments to under a minute? There will be more opportunities to communicate with the Office about this topic in the future. Chris and then Ali and then

Jule and Rachel.

MR. CALLISON-BURCH: Thanks, everyone. In terms of the guidance that the Copyright Office issued last month, I think I would like to advocate for a more expansive view of what constitutes human authorship. I think that Microsoft's metaphor of a copilot where it's really actually being driven by humans is absolutely a good way of conceptualizing this, and the prompt completion that was outlined in that document is a limited view on what's possible. Here's a figure that shows interactions that allow where a human has asked the system to rewrite an original input sentence using many different transformations that clearly show that the human is driving the types of transformations that result from a final output. So thank you to the co-panelists for the spirited discussion.

MS. WALTERS: Thank you.

Ali?

MS. STERNBURG: Yes, a few brief on the question, on the final question. From our perspective, the Office helpfully applied a lot of existing precedent going back to the 1800s on authorship. I have my little monkey selfie here. There may be instances where we might want more

clarifications -- where it might be helpful to have more clarification on the sufficiency of human input, but, obviously, these things are really still developing in its really early stages.

And then just as a closing statement, again, I guess to continue to put back on the point most styles and competing in the market and fair use. I understand we're looking at what an AI could do. It's also important to think about whether you'd want the same restriction on, like, a human artist. That's how people learn. That's how people create, looking at prior work to develop an understanding of artistic styles and attempting to recreate them, and so, if a human can do it, an AI should be able at least by default. So thank you so much for this today.

MS. WALTERS: Thank you.

Jule and then Rachel.

MR. SIGALL: Thanks very much and thanks again for the chance. This is a really useful and interesting discussion, and I was happy to be a part of it. I touched on the challenges that the Office's registration guidance might have in light of computer code and computer software especially that's using AI tools, so I won't repeat those.

I do think, though, I think the Office -- I

would suggest to the Office to consider that that discussion about how new creators using AI tools fit into the registration system is a good place to start for the broader discussion around how AI tools should fit into the copyright system because I think the reality, as you've heard from myself and others today, is everyone will be using these technologies, so the question is, what kind of copyright system should we have, registration system, liability system, policy? What kind of system should we have for those using these technologies?

And I think, when you put the author at the center of that discussion and understand how they're using these technologies, and, in fact, you maybe broaden our conceptual notions of who is an author, including newer authors using these technologies in ways that incumbent authors are not, I think you may get a better sense of how to draw the lines on some of the other broader issues as well, and I think, if we start there, I think we'll end up with a copyright system that may be built more for the future than for the past.

MS. WALTERS: Thank you.

Rachel, and then we will end with Edward.

MS. BROOKE: Thanks, and thanks for

convening this session today. It was so great. So, as I mentioned, Authors Alliance really approves of the registration guidance, but I agree with others that there still may be open questions around the edges, like, for instance, about the copyright status of things like co-authored works with both human and AI authors where the contributions can't be easily disentangled. I think this underscores the point that these technologies and the development of the uses that they facilitate are still in their nascent stages.

So, to close out, I'd like to return to the purposes of copyright, which we've touched on but not focused on today. It's important to remember that copyright is not only about protecting the rights of copyright holders but incentivizing creativity for the benefit of the public, so registration issues aside, the new forms of creation made possible through generative AI can incentivize people who wouldn't otherwise create expressive works to do so, bringing more people into these creative industries and adding new creative expression to the world, which I think we can all agree is strongly in the public benefit.

MS. WALTERS: Thank you.

Edward?

MR. HASBROUCK: Thank you. Your questions have focused on economic rights, but I want to conclude by reminding you of the imperative to include moral rights throughout this inquiry and particularly to call to your attention the intolerable position that the lack of enforceable moral rights to attribution and integrity place on authors even if all of this use is fair use. Today, anything I publish on the web can be and probably already has been ingested. My work is being used, has already been used to produce fake news, fascist propaganda, spam, and defamation, and I have no means of objecting to its use by AI companies to generate that prejudicial material. This is wrong. It needs to be acted upon, and I hope you will include that in your legislative recommendations. Thank you.

MS. WALTERS: Thank you. I'm going to pass it back to Jenee for the closing of this session.

MS. IYER: I am, in turn, going to pass it over to Andrew.

MR. FOGLIA: Thanks, Jenee, and thank you again to all of our panelists for your participation today. As Jenee and Heather both mentioned and as we said at the outset, there will be other opportunities for you to comment on these various questions.

We are now going to take a 10-minute break. We will resume at 2:43, and I look forward to seeing some of you back then. Thanks, everyone.

(Whereupon, a brief recess was taken.)

MR. FOGLIA: Welcome back, everyone. We will begin the second panel now. For those of you who are just joining us, a few Zoom housekeeping points before we begin. If you are joining this session but are not a panelist for this particular session, please keep your camera turned off and your mic on mute. We are recording this session today. The recording will be available on our website, and the transcription function is activated as well.

This panel will work exactly the same as the previous one. We're going to start with a brief introduction and short statement from each of the panelists if they desire. We request that these statements be limited to three minutes, and the moderators will be watching the time. After these introductions, we will have a moderated listening session. The moderator questions, which panelists received in advance, are intended only as prompts for discussion, and we welcome participants to share relevant perspectives and experiences they feel are important for the Office to hear.

With that, I will hand it over to the moderators for the second session. Brandy Karl is an Assistant General Counsel in our Office of General Counsel. Keyana Pusey is a Barbara A. Ringer Fellow.

The mic is yours, Brandy.

MS. KARL: Thank you, Andrew, and thank you to all of our panelists joining today. We're going to begin in the order stated on the agenda, and that is with Andreessen Horowitz.

MR. DAMLE: Hi, everybody. I'm Sy Damle from the law firm of Latham & Watkins. As Brandy mentioned, I'm here representing Andreessen Horowitz or also called a16z. a16z is a venture capital firm based in Silicon Valley that invests in companies that both build and rely on artificial intelligence technologies. a16z's interest in these proceedings is in ensuring that responsibly designed AI technologies remain both lawful to create and open to use, so I want to start the panel today by making two important factual observations.

First, we've found in many discussions around these issues that there's some confusion about what the output of these models tends to look like in relation to the input, so I just want to be crystal-clear that as an empirical matter the overwhelming

majority of the time the output of a content generation AI service standing alone is not substantially similar in the copyright sense to any particular copyrighted work that was used to train the model. To just give one data point, one research team tried to get one of the popular image generation tools to output 350,000 images from the original training set used to train that model. It succeeded only 0.03 percent of the time.

In other words, 99.97 percent of the time the output did not replicate the images used to train the AI, and that was while researchers were trying to infringe, not in your normal use case, so important for us as we're collectively considering an appropriate public policy approach here to keep in mind that what we're talking about is an innovation whose output taken alone would constitute prima facie copyright infringement only in the genuinely rarest of cases.

The second point I want to make, it was notable to me to hear today from the first panel how many authors and creators themselves are relying on these new generative AI tools to help them with their work. In other words, these tools are not substitutes for human creativity but are themselves engines of

human creativity. So the point that I want to be sure to emphasize is that, really, the only practical way for these tools to exist is if they can be trained on massive amounts of data without having to license that data.

In fact, the data needed is so massive that even collective licensing really can't work. What we're talking about in the context of these large language models is training on a corpus that is essentially the entire volume of the written word. That volume creates complications that are way more complicated than what the Office was faced with when it was attempting to set up an extended collective licensing scheme around mass digitization, which, as we know, ultimately failed.

So, with those two points in mind, I just want to say that I'm grateful to the Copyright Office for having me here and for opening these conversations and look forward to today.

MS. KARL: Thank you, Sy.

Next, we have the Association of American Publishers.

MR. HART: Hi. Good afternoon, everybody. I'm Terry Hart, the General Counsel of the Association of American Publishers, or AAP. AAP represents the

nation's leading book, education, and journal publishers. I want to thank the Copyright Office for convening these listening sessions and for its thoughtful and careful consideration of these very important issues. I just wanted to start with a couple of overall points before jumping into the moderated discussion.

So, first and foremost, my members, publishers of all types, are using and investing in AI technologies as we speak and plan to do so going forward. At the same time, the high-quality works that they publish are very valuable for training AI models. So, certainly, there's a lot of promise with AI across all sectors, a lot of opportunities, but I would urge policymakers not to be tempted to sacrifice copyright in a race to advance AI. A strong publishing industry is just as vital to prosperity and the progress of science and the useful arts in an AI era as it has been for centuries. Publishers and, indeed, all creators need to be a part of these conversations.

I think, two, that licensing solutions remain the best tool for facilitating AI development while protecting the rights of copyright owners and licensees. Licensing preserves the incentives for

authors and publishers to create and it encourages investment in high-quality data sets. The U.S. copyright system has been successful in adapting to new technologies for decades, centuries even, and can accommodate to continue development of AI, and I think overall changes to the copyright law framework at this time would be premature. Thank you.

MS. KARL: Thank you, Terry.

Next, we have Ms. Chabala.

MS. CHABALA: Hi. Thank you so much for allowing me to participate today. It's been great to get everybody's input on this topic. So I'm a writer, formerly an editor at Shondaland.com, and I have two degrees in writing from USC, so I'm very connected to a lot of creative writers and authors of books, fiction, non-fiction, poetry, et cetera, and I also have lots of writer friends and freelancers, and so I'm really excited to kind of present the perspective of all of us. Additionally, I'm also a budding music producer and I'm surrounded by musicians, music producers, artists, so, you know, I'm coming with the perspective from those guys.

So the first thing real fast, just having listened to everybody, it's not that myself or the people that I associate with are against AI, but the

idea that we're all going to be using these programs, that we all want to use these programs, that we want to make everything more and more efficient, that's something that, you know, I don't think should be assumed. From where I sit, from where my writer friends sit, and, you know, we're people who studied the humanities, we're kind of asking ourselves why things need to go faster.

Our society is just racing ahead faster and faster and for what? I'm not sure. Technology is a great thing, but, yeah, we've extended life span, but then, at the same time, we've also extended disability and incapacitation. So I think these are things that, like, it's a really general idea, but I think it's important for everybody to know that that's where, you know, a lot of creators are coming from.

So I'm not sure how much more time I have. I think that everybody's really covered the training data and all of that very well. One of the things that I just wanted to talk about is how we can use copyright to ensure that the public knows what is and is not written by AI or even partially written. I think this is -- and we might all align on this, right? But I think we can use copyright for that even if it includes flexible copyright.

So I'm probably running out of time here, but I think that doing so, ensuring that we know what is created by AI and what isn't respects and reinforces the sanctity and specialness of human-generated works, championing human narratives, ideas, and reason, which by virtue of being human are filtered and shaped through invaluable lived experience, inductive learning, and on-the-ground research.

I think this is really important that we understand how important that is and that it protects writers' works from the devaluation that might ensue as a result of a literary and journalistic marketplace saturated with low-quality, low-effort manuscripts, articles, queries, and submissions, and it can almost most importantly protect us from disorienting disinformation on a mass scale, which obviously can have very real, very grave, real-world repercussions, and this kind of system of whether it's, you know, flexible copyright can allow AI to take up its own space in the literary marketplace, one where AI is competing transparently with other works, be they human or from machines, which can be good for, you know, AI and the creativity that can flow from that. So I'll end with that.

MS. KARL: Thank you. Thank you, Tracy.

Next, we have Copyright Clearance Center.

MS. ZALLER ROWLAND: Thanks, Brandy. It's really nice to be here today. Thank you for the opportunity to speak. We're looking forward to participating in this initiative and support the Copyright Office in its efforts to develop policies that will, as the U.S. Constitution mandates, promote the progress of science. So CCC strongly supports a well-functioning copyright system, one that respects copyright ownership and licensing and enables lawful uses and robust markets. We opened our doors on the day that the 1976 Copyright Act went into effect, and ever since then, we've been providing solutions for a variety of copyright matters.

We offer voluntary licenses for millions of literary works, including transactional and collective options. We also offer software solutions that help you search, discover, access, elaborate on and analyze copyrighted works while being mindful of copyright compliance. So one example, and it's pretty apt for today's discussion, is that we've offered a right line XML dilution which powers AI discoveries for over a decade. That's all to say that CCC is fundamentally committed to supporting the copyright system and to

provide solutions that help copyright work.

We appreciate that the issues surrounding copyright and its intersection with AI technologies are many, ranging from routine to incredibly complex. Copyright comes into play at several points in this AI journey, including when AI technologies use copyrighted works as part of a corpus, include copyrightable software, which is one thing I don't think anyone has really talked too much about yet today, and also result in various outputs.

While we are discussing literary works today, these touch points are applicable to every type of copyrightable work, and it's important to pay due attention to each touch point and it's importation when we are discussing AI and copyright. So one way to pay due attention is to ensure that licensing continues to be a key part of this discussion. There are already licenses in place for various AI-related uses, and licensing is an obvious solution to many of the issues that are raised by the use of AI technology vis-a-vis copyright.

Licensing, including collective licensing, offers an effective and efficient way for AI technologies to use copyrighted materials while respecting creators and copyright owners. Enabling a

robust licensing market and continuing to respect balanced copyright systems will benefit creators, owners, users, and technology overall.

So, again, thank you for letting me participate today. It's been really interesting so far, and I'm looking forward to discussing these important issues further, so thank you.

MS. KARL: Thank you, Catherine.

Next up, we have Creative Commons.

MR. SLATER: Thank you. Good afternoon. My name is Derek Slater. I'm a founding partner at Proteus Strategies, which is a tech policy consulting firm. I'm here representing Creative Commons today. For those who don't know CC, different than CCC, Creative Commons is a leading global nonprofit organization that helps overcome legal barriers to the sharing of knowledge and creativity, and CC is the steward of the widely used Creative Commons license suite for open content, and, you know, CC was really built, the founding insight is all creativity built on the past, all creativity built in the commons in one way or another.

And that's true whether it was some of the authors groups that we heard from in the last session, people today or people who are going to be using AI

and certainly AI itself, and so, for many years, Creative Commons has been looking at the interplay between copyright and AI not just because we're interested in the technology but because we're interested in fostering people building on the commons and contributing back to it to ensure better sharing.

And so, as generative AI has become really prominent the last few months, we've been engaging widely with different stakeholders, artists, technologists, policymakers, in settings quite like this to help think about the benefits and the challenges, and I just want to summarize two of the key takeaways.

You know, echoing some of what we heard this morning, you know, AI isn't a homogenous technology. There's also a huge diversity of uses and many creators are benefitting. Professionals are benefitting, creating fiction, non-fiction. Organizations and companies are using it in various settings as a productivity tool, and also there are just sort of amateur uses or communities who use it. You know, I'm, in fact, not in a jail cell today but in a co-working spot we're right out there. There are people working on new sets of tools to help people translate, for instance, fiction writing into visual

storytelling, so bridging some of the sessions, and this is early days.

For our part, you know, when it comes to copyright law, it's important to keep that variety of uses in mind. We think training on copyright works is generally going to be lawful under fair use and other exceptions. Similar to what the Copyright Office has said, we think there should be significant human creativity for something to be copyrightable. Otherwise, AI outputs should not be copyrightable. Nuances here matter. The facts are going to matter, but we should be serving copyright's purposes, building on the commons, growing it with further material and tailoring solutions accordingly.

Finally, we also think norms and tools outside of copyright are helping address key concerns, so just echoing I think what Matt Sag said earlier this morning, you know, there's still room to think about what are best practices, and we've seen companies who are adopting opt-out regimes of various sorts. You know, Robots.txt, actually contra to what we heard in the first session, is more flexible than what people may think.

We've also had people come to us and say, look, we know there's fair use, but we want to train

on Creative Commons licensed works. Help us do that because that's how we want to build our tool to build a sort of training commons. So we think collaboration in settings like this can be really important and look forward to the discussion. Thanks.

MS. KARL: Thank you, Derek.

Next, can we please have Internet Archives?

MR. ROUTHIER: Thanks. Hi. Good afternoon. My name is Peter Routhier. I'm policy counsel at the Internet Archive. Thank you to the Copyright Office for organizing this session. Thanks to all the fellow speakers sharing their views today.

I want to follow on something Derek just said and note that I think it's important that we keep in mind the purpose of copyright when we're looking at new technology like this. As the Supreme Court said in Harper & Row, copyright is intended to increase and not to impede the harvest of knowledge. Put another way, copyright ought to further the public's interest in obtaining knowledge and learning and, of course, in the progress of science and the useful arts, and already today artificial intelligence is helping to do this. At the Internet Archive, for example, we digitize many texts that have only ever been made available in physical form. Our Democracy's Library

Project digitizes many government works for preservation, access, and a host of other uses.

There are significant constraints on our ability to do so because there are generally no commercial incentives to digitize these works even though it serves the goals of copyright by increasing the harvest of knowledge. So it is a great benefit that we, as we digitize these works, can use artificial intelligence and machine learning technologies to help at many steps along the way. That starts with something as simple as OCR tools, which have been greatly improved through the use of machine learning technology in recent years, to metadata extraction and summarization using the latest large language models.

Better and more efficient ways of digitizing these works serves the purposes of copyright and ought to be encouraged, and this is not, of course, the only example. We've already heard many other ways today of how this technology's helping ordinary people learn from and create new works. We should let the robots read. Artificial intelligence has the power to learn things that no human could and has the potential not only to serve copyright interests but to be of tremendous value to society as a whole. Thank you.

MS. KARL: Thank you, Peter.

Next, we have News Media Alliance.

MS. ARATO: Hello. I'm Cynthia Arato of Shapiro Arato Bach. I speak for the News Media Alliance, which represents the most trusted publishers in print and digital media in the United States, from small local outlets to national and international publications read around the world. Every day, the Alliance's members invest in producing high-quality, creative content that is engaging, informative, accurate, and trustworthy.

The Alliance's members make significant contributions to the U.S. economy. They play a crucial role in informing our communities and sustaining our democracy, and their ability to serve these pivotal roles is increasingly imperiled when they do not get fair credit or compensation for their contributions. The critical task for this Office and for legislators and stakeholders too is to facilitate the growth of generative AI while ensuring fair credit and compensation for the creators whose works make the field possible.

Generative AI systems, while holding promise, are commercial products that have been built and are run on the backs of creative contributors.

These systems have been developed by ingesting massive amounts of the creative output of publishers, often without authorization or compensation, and they disseminate that same content in response to user queries, again, without authorization or payment and often with little or no attribution or link to the original news source.

Such disassociated output necessarily results in zero clicks for the publishers, severing the publishers' relationships with their readers, reducing traffic to publisher sites, and damaging publisher brands that have been built for decades. Copyright laws should protect and not harm publishers in this setting. Developers and deployers of generative AI should not use expressive works without authorization and should respect publishers' rights to negotiate fair compensation for the use of their valuable works.

The system should also be transparent to publishers and users. They should identify the content used to fuel their products and connect and not disintermediate users with publishers. Protecting publishers' legitimate intellectual property interests will strengthen, not impede, generative AI innovation because authorized use of publisher content can

improve the reliability and accuracy of AI products, which will enhance system output and bolster consumer confidence.

This is not unchartered territory. There are existing functioning markets for licensing content where compensation frameworks are already in place to permit use in return for a payment, and copyright laws have previously navigated issues of comparable scale and complexity, resulting in a wide range of mechanisms for consent and payment. We believe copyright can do so again here. Thank you.

MS. KARL: Thank you, Cynthia.

Next, we have the Organization for Transformative Works.

MS. ROSENBLATT: Hi. I'm Betsy Rosenblatt. I'm a Professor of Law at University of Tulsa College of Law, and I'm Legal Chair of the Organization for Transformative Works, which is a volunteer-operated nonprofit that advocates for fans and fan works, including fan fiction, and we're in the unique position, I think, of being creators, users, and re-creators of works, as well as an online service provider of a volunteer-coded website.

I want to start with the understanding that not all language learning models or other AIs operate

the same way. They operate differently both in the way they create models and train models and the ways they generate output, and so we may not be able to make generalized rules about AI, and that's fine. We shouldn't require ourselves to.

On to what the Organization for Transformative Works thinks, if only one thing comes out of this process, I think it should be this: That it is crucial to divide copyright's relationship with generative AI into three wholly separate questions. Each of these questions is independent of the others, and mixing them together muddles the relevant legal questions and will lead to incoherent results.

The first is, when is and isn't crawling and scraping for training purposes infringement? The second is, when an AI generates or is used to generate something substantially similar to a copyrighted work, who is responsible for the infringement? And the third is, when an AI is involved in generating a work, who, if anyone, owns copyright in the work? These are wholly separate questions, and I want to address the first one, which may well be the thorniest and most polarizing, perhaps especially among Organization for Transformative Work members and volunteers. This is deeply, viscerally tied to people's senses of

morality, fairness, and even their senses of self.

Regarding this question, precedent indicates that many types of scraping for purposes of machine training either does not implicate copyright at all or constitutes fair use under, for example, the Google Books precedent, but different systems are different. We must consider what a training model actually fixes in a tangible stable medium of expression. Many do not fix works in a tangible medium of expression at all. They process works into math without saving them in any specifically recoverable way and thus don't implicate copyright at all, and if they do reproduce works, we need to consider fair use, how the use transforms the works, how the use affects the market for the works.

At the same time, we must consider the deeply, intensely-held reactions of those whose works are incorporated into training models. Here's what bothers them: not being asked, not being allowed to opt out, not getting attribution for their contributions, and this is true especially if and when those contributions have the ability to generate works similar to theirs.

MS. KARL: Thank you, Betsy. I'm sorry. We're going to have to move on. If you can save for

the rest, I appreciate it.

Then, next, we have the Software and Information Industry Association.

MR. MOHR: Hello. Hi. Thank you for the opportunity to participate. My name's Chris Mohr. I'm the President of the Software and Information Industry Association, or SIIA. We are the principal U.S. trade association for those in the business of information. Our 500-plus member companies include platforms, financial information providers, scientific, technical, and medical publishers, database publishers, and educational technology firms. We are deeply involved in many of the questions surrounding AI development, ranging from its implications for privacy to automated decision-making to broader implications for U.S. global competitiveness.

We believe that in order for the technology to reach its full promise it must be transparent and ethical, and we're supportive of efforts by NIST, among other agencies, to develop guidelines for responsible AI use. Our mission is to protect the three components of the information life cycle, namely, creation, dissemination, and productive use. A healthy copyright system is essential to the health

of that life cycle, and we thank the Office for convening this group to discuss the implications of AI on that system.

One of the strengths of the copyright law is its technological neutrality, and if you don't believe that, I've got a digital audiotape machine I'd love to sell you. Copyright’s evolution is driven by new technology, and that pattern continues in its application to AI.

I'd like to make three quick points in advance of today's conversation. First, a robust licensing environment is essential to the health of the business of information, and we as a group have dedicated much of our institutional existence to enforcing the certainty of licensing arrangements.

Many of our members already license protected works for text and data mining. Copyright has always functioned as a property right against which that licensing occurs, and that has to continue. It does not follow that all AI uses must be licensed. Whether a particular use is or is not fair will depend on one of several of its well-known factors. Our overall view is that the doctrines of equity contained in fair use are going to be more than adequate to sort the proper use from the improper use.

And, finally, we don't believe that existing copyright law requires change to handle AI output or authorship. The Copyright Office both in its registration denials for AI-generated works and subsequent guidance have reached the right conclusions, though we have some concerns around the edges similar to what you heard in the last panel. But, overall, the good news is, I think, that these registration decisions line up with judicial and agency interpretations, for example, in other areas, like patents, which find that human beings are required to meet both statutes' inventorship or authorship requirements. Thank you again for inviting me to be part of the conversation.

MS. KARL: Thank you, Chris.

Our final panelist today is from Yale Law School.

MS. KHAN: Hi. Good afternoon. My name is Mehtab Khan. I am an Associate Research Scholar at Yale Law School's Information Society Project. I work on the intersections between intellectual property, specifically, copyright law and data governance, and at present, I'm studying the connections between the ex-ante processes of AI development and how they impact downstream intellectual property rights. I'm

interested in the multifaceted nature of the development of AI tools, including but not limited to generative tools, and the various stakeholders that are implicated at the input stages.

I have a few comments based on my current work that might be helpful as we move forward in this conversation. The first comment is that we need to take into account the steps involved in the creation of these tools and the copyright issues that arise at each stage that may be beyond just determining whether there has been unauthorized copying or whether that copying is fair use or not.

There are important policy reasons for taking into account these various stages. It helps firstly lend clarity to who the stakeholders are. There seems to be a disconnect between the copyright holders, who are concerned about the output, and who might they direct their complaints towards. Is it the platform that is hosting the tool? Is it the developers who are this abstract category of people and entities who are responsible for creating the tools? And who are the collectors of these copyrighted works that go into training of these data sets? So it's important to take into account the various stages involved and who these stakeholders

might be.

Another reason to take into account these different stages is that we don't want to risk placing too much responsibility to interpret and react to the output created on the users and on the downstream providers and users of these tools because they lack the capacity to access and understand the core components of the technology. There's more power and more information available at the input stages, and so it makes it more critical to identify who these entities and stakeholders are so that we might be able to discern who might be responsible for what action purpose.

The second comment I have is that we need to take into account sector-specific issues and not move towards a one-size-fits-all solution, so authors might have very different concerns from artists, especially small-scale artists, who might have very different practices and expectations when it comes to their work being used, and that also means taking into account norms of an industry. Software sharing might be very different from how artists share their work or how fans share their work in fan fiction or how open-knowledge projects like Wikipedia have distinct guidelines on how knowledge may be shared.

The last point I would make is that we should remember that this issue of how copyrighted works are used to create AI tools is and should not be limited to just generative AI but also to how AI development takes place in general to create various applications and purposes and how copyright is implicated over there as well both at the input and the output stages. Thank you.

MS. KARL: Thank you, Mehtab. If you can continue that thought when we move to the questions, that would be wonderful.

We're going to start off with our first question. Thanks again to all of our panelists for introducing themselves. Welcome again, and we're going to start with the question, what artificial intelligence technologies are you or others in your industry using in the creation of new works? Please use the Raise Hand function and we'd ask everyone, all of our panelists, keep their cameras on. And, first, we have Sy.

MR. DAMLE: Hi. So I want to use this question as an opportunity just to sort of get a little more granular about how these technologies actually work, how they're trained. Some of that was discussed earlier. So I think the important point to

understand is that in the context of these large language models like chatGPT that the algorithm fundamentally is learning facts about language, unprotectable facts about language, not actually retaining the content of the works themselves.

So I think other people in the prior panel talked about the fact that this is not a collage machine. What it's doing is it's sort of taking the works apart, all the works in a training set apart and trying to learn about things like statistical correlations between words, so, like, just to get a little more granular, what typically happens in one of these tools is that you take all of the works and you don't ingest them sort of one at a time. What you're doing is you're breaking them into pieces and you're feeding those pieces into the AI tool sort of randomly, and so the model isn't learning the entire work. It's, like, actually not, like, possible to given the way that the works are fed into them, but they're learning about statistical correlations between the pieces of the works.

So just to give, like, a very specific example, the model might learn that across the entirety of the English language the words "Today is a" are much more likely to be followed by the words "a

beautiful day" or "Tuesday" rather than the word "green" because "Today is a green" wouldn't make sense. Those words don't appear together in the English language at all, if maybe very infrequently. And it might also learn that if the word "rain" is in close proximity to the words "Today is a," that the word "Tuesday" is more likely to be the next word that follows rather than the words "beautiful day." So just a very tiny example of the kinds of statistical correlations that it's learning, right? These are unprotectable facts about language that it's learning, and that's what it's storing. It doesn't store the works themselves.

And then, on the output side, the model is taking these factual statistical correlations and then using them to decide with a fair amount of randomness what word should come next over and over, sort of a very, like, advanced auto-complete, and so, as we've all seen for those of us who have used chatGPT and other tools like it, what is output are works that are sort of radically different than anything that the model was trained on because all the model "knows" are these statistical correlations about the relationship between concepts and language. Those are all unprotectable, and that's what's stored in the model,

so I just wanted to make that point about the technology.

MS. KARL: Thank you, Sy.

Next, can we have Terry?

MR. HART: Yeah, thanks. So I wanted to answer this question but quickly respond to Sy and something one of the previous panelists also said just to make this point, you know, I think one of the benefits of the Copyright Office doing these listening sessions and learning as much as it can about AI is to understand, you know, how it works and how that intersects with how the Copyright Act works, and so, specifically -- and I think, certainly, there's a lot correct about what Sy has said about how certain large language models work.

But I think also it's correct that at some point any large language model or any AI that's being trained on a corpus of textual works is at some point going to be making a reproduction or some type of use of a copyrighted work that at least on its face would be considered protected by one of the exclusive rights under 106, whether it's actually a reproduction, whether it's some kind of derivative work where there may be tokenizing the book, sends it into a version that the machine can read and understand, which, you

know, going back to my original point about the Copyright Act being technologically neutral, you know, it does cover reproductions that are made and versions that are machine-readable rather than human-readable.

So I think, in many cases, there is at least a prime facie instance of reproduction or some type of other protective copyright use. Now, whether, you know, I think it's a much more challenging question whether in the end that's protected by fair use or excused by fair use or not, you know, certainly, a much more complex question. I think, in certain instances, it could, in certain instances, it could not, but did want to just at least, you know, make that point that I think, in many instances, there will be a prima facie instance of copying that if not excused by fair use would constitute infringement.

So that said, I'm going to the question about how AI technologies are being used by my members in the industry, and, you know, they are widely divergent between the trade book side, between the scientific journals, scientific and scholarly publishers, educational publishers, higher ed, K through 12, so they're all using AI technologies in different ways, including using them for translation, for research integrity, for marketing, assessing

scholarly impacts, and plenty of other applications.

And what I would offer to the Copyright Office is that most of my members would be eager to demonstrate, you know, in one-on-one sessions with the Copyright Office how they're using certain AI technologies, so I wanted to offer that. You know, AAP, I would be happy to help facilitate those sessions, you know, to the extent that you think you would find those types of demonstrations helpful. Thanks.

MS. KARL: Thanks so much, Terry.

Chris?

MR. MOHR: Thanks. So a few things. I mean, so our members are using this technology in a wide variety of circumstances. One is a large language model. Another is, ironically enough, in compilations in ways that have been going on for quite some time in the sense that some of what our members do is to provide, you know, your customer data. When you do that, there are probabilities that are used -- this is AI -- to comb through a huge data set to figure out which John Smith are we talking about here and which information is relevant to him.

Answering those questions requires selection, coordination, and arrangement of what's

likely to be relevant and what isn't. That has been around for a while. It's going to continue. It is not, I think -- and in order to do that, it requires copying of a variety of different works.

We have folks who are in what we call the or they call the alt-data business, and what that is is they use AI to -- the easiest example is to track market sentiment based on, around a given security, using publicly available data that may be on the internet, and we also see it in the use of plagiarism detectors, as well as even we see our educational technology providers experimenting with it to do things like have it write draft questions at the end of a learning module that are then reviewed by humans to be sure they're okay, and they are attributed to AI generation. So we use it -- our members, rather, use it in a wide variety of circumstances.

MS. KARL: Thank you, Chris.

Catherine, can we have you next, please?

MS. ZALLER ROWLAND: Sure, and I wanted to start by saying that everybody uses AI in some ways even if you don't realize it, right? So, if you're in your email and you're typing, you know, something and it just pops up through the auto-fill, then you've got some sort of technology involved there. There are so

many different ways of doing it.

The way that CCC is involved is their licensing, as I've mentioned before, but the licensing is for all sorts of AI-related uses, machine learning, training, that kind of thing. What we're looking at is how you can use technology and licenses to go ahead and help the entire system. So, for example, we have our license that people use that is mostly based on scientific material, so there are people using it in those fields, and then you also have it, as Chris was mentioning, we have a license that involves some sort of abilities for curriculum if you have kind of -- or you answered one question one way and then you have to ask another question, et cetera. There are some sort of AI capabilities that are licensable through that as well.

So there are licenses that do cover some of these things, many of these things, in fact, so there is a market that's there that is operating, and one thing I wanted to mention about that is that, you know, regardless of what country you're in or what license you operate, this is a global economy that we live in and a global copyright world, right? So you might have a law here in the United States or not a specific law but an interpretation, and a completely

different country might have a different way of doing things, right?

So there are a variety of countries with different views, so licensing is one of those ways that we've been involved that helps people be able to understand, like, what can they do in an ethical, compliant way and be able to make sure that they know what they're getting. And one thing I just wanted to quickly mention about, you know, the facts that, you know, what if we're just trying to get the facts? What are we trying to do with this? There are so many different kinds of AI technologies and use cases, but, if you have something that's in a copyright protected work, you still have a copyright protected work at issue that you have to consider and how to deal with that, so, you know, you've got the expression of these facts. You've got the context around them. A lot of these things are really important for training your systems. With that, I will turn it back to you.

MS. KARL: Thank you, Catherine.

We are going to transition to the next question and on to my colleague, Keyana, but, Betsy and Cynthia, if you would please include your answers to this question along with the next one, thank you so much.

MS. PUSEY: Thanks, Brandy.

So the next question is, what do you think the Copyright Office should know regarding how AI systems generate literary material, whether that be fiction, non-fiction, or code?

And, Betsy, you can go ahead.

MS. ROSENBLATT: Sure. I wasn't ready for that one. I was ready for the one that we were just talking about, but the Organization for Transformative Works does not make much use of generative AI itself, although it may use it to assist in creating and refining its code, which is all open source.

That said, many fans, including disabled fans, especially disabled fans, depend on generative AI to create and consume works. Other fans are exploring what various generative AIs can make, but, by and large, fan works are expressive of a fan's own particular creative interest in self-expression, and so AI cannot replace fan-created works. It may well and I think be encouraged to contribute to the body of fan works overall, and so fans may not oppose the use of AI to create works or engage in brainstorming but do have serious concerns about AI being used deceptively, especially without disclosure, and I think that's a combination of strongly held anti-

plagiarism norms and pro-attribution norms in fandom.

And I'm genuinely not ready for the previous question, but I would love to continue what I started to say at the beginning, which is that what fans are concerned about mostly is they're not concerned with payment because they're making a different kind of work. They are concerned with their deeply held beliefs about plagiarism and attribution, and so, when we think about how works are used to be interpolated into training models, I think fans would say we need to consider a number of things.

We need to consider how the learning models work, what kinds of works they scrape, the retrievability and perceivability of scraped material, how the model uses what is scraped to create new works, the serious social and communicational drawbacks of limiting scraping to only public domain works, which are archaic, and we know that perpetuates bias and outdated ideas, and also when and to what extent opt-out is feasible. So I may come back later, but those are my thoughts at the moment.

MS. PUSEY: Thank you, Betsy. Sorry to have caught you off guard there, but thank you for your insight.

Cynthia, did you want to answer this

question as well, or was your hand up for the last question?

MS. ARATO: No, I'd like to answer this question, and one feeds into the other. So I'd say I want to talk about three things, the quantity and the quality of the copying that is being done, as well as the lack of transparency. So I think, when the copying is being done to fuel these AI products, the entirety of works are being copied, and that includes the entirety of the expression in what are valuable creative expressive works. It seems artificial to talk about how the copying is being done just so a system can figure out one isolated word to put after the next.

Ultimately, those words form sentences which form paragraphs which form entire creative works, so I think that the comment Sy made seems to disaggregate the process into nothing and ignores the reality that the expressive work's being copied, and I think some AI systems are able to generate whole paragraphs in the style of particular authors, so I think that it goes way beyond isolated facts.

And then there's just a lack of transparency. There's a lack of transparency to users and to content owners. So many of these systems may

fail to provide source attribution, so when responses are given to users, they're done without knowing the original source of the material and the same for publishers. The systems ingest massive amounts of content without identifying the works that have been copied or from where those works have been obtained.

MS. PUSEY: Thank you, Cynthia.

We're going to go to Tracy.

MS. CHABALA: Yeah, I think that there's very broad use of these models, obviously. There's also many different types of writing, very, very broad there, and we'll have to distinguish between using AI to follow the style rules of Strunk & White from using AI to write whole books, fiction, non-fiction. Obviously, we need to look at research-based works differently than literary works.

There's a big difference between using an AI to aggregate static information for a research paper or using it to generate ideas for a philosophical work, and I think once again we're kind of getting caught up on writers using AI as tools rather than really looking at generative AI, which is about creating whole works, so I think it's really important to not be disingenuous about the fact that that's happening, right? It's not just writers using it, you

know, as little ways to streamline their work.

And then, again, I'd like to emphasize that not all writers are using generative AI at all, right? I use an AI transcription program. I don't believe that's generative AI. It might be and I'm unaware of it, but, you know, yeah, I think it's important to emphasize that.

MS. PUSEY: Thank you.

Derek?

MR. SLATER: Sure. I think, in helping to think about how these tools are trained, again, it's not a homogenous group. There's a whole variety of systems. It's really useful to go to the framework that I think Betsy started to tease out of separating training and then the output itself and even go further, so you could start with a sort of collection of the data, and as we heard, some of that can be scraped from websites. You know, I think, Cynthia was just saying, well, it shouldn't make a difference if it's the whole work. It feeds, you know, a word feeds into a page, feeds into an article.

But, in fact, I mean, that really does and has made a difference in copyright. That is to say, the copying of a whole work as an intermediate step in non-infringing uses can be legitimate. The whole work

getting copied is not dispositive. Talked a little bit earlier about Google Books, but there are other cases like Sega v. Accolade and Sony v. Connectix where, yes, they copied the whole software in order to take the unprotectable elements and create a whole interoperable video game or video game system, so that has the same sort of logic that I think Cynthia was speaking to and again was a fair use. So there's that part of collecting the data, potentially scraping the LLM.

Looking at that in its own right is also important because you can think about interventions, and these don't have to be legal interventions, but the sort of voluntary ones that, you know, we've talked about with opt-out and so on. Robots.txt we've heard about so far is one, and just to be really clear, it's not true that it's sort of all or nothing. You can choose to opt out of certain user agents to not have them scrape your site or parts of your site and allow other agents to do so, and, yes, libraries and research institutions might decide, you know, we're going to scrape it anyway because that's our archival mission, and it's good that the law doesn't prohibit it, instead we have voluntary best practices to help mediate that exchange. That's the scraping

portion.

Then, again, there's sort of the training. We're subtracting that sort of uncopyrightable elements looking at the works as data to create a model. Again, that's an intermediate step to then the eventual output, and this is where I think the rubber really meets the road. I think Tracy was just speaking as a bit of, well, in some cases, you're using it maybe to assist you to do -- I love it for citation, for doing my citations, so Strunk & White style is a great example or maybe just like, look, I need to break my writer's block. Like, help me think of something for this character. But there are other examples that are more difficult where the output may incorporate or be used in a way that incorporates something that is substantially similar to the original.

And, again, as I think The Authors Alliance said earlier, we have legal tools to think about that, the substantial similarity test, and then the question is, well, you know, who is responsible? The user is the one doing the prompt. We have tools like secondary liability to think about whether the tool creator themselves is contributing or not. I think, in most cases, they aren't. It's the user who's doing

it. But I think this helps explain why it's really important to separate out those stages, to think about the types of interventions that the law does and should make and then other sorts of interventions we might make around transparency, attribution, and so on.

MS. PUSEY: Thanks, Derek.

Sy?

MR. DAMLE: So I want to just emphasize a point that Derek just made and sort of fold it back into a point that I was making, that the way that the law works in this area is that you don't look at the sort of, you know, intermediate copy in isolation. Every case is looking at the purpose of that intermediate copy. Sony, Sega, Google Books even, all look at what is the ultimate purpose, what is the output.

And so, as I started with, I think there are sort of two points here that I haven't really heard anyone dispute. One is the copies are being made not to store those copies. The copies are being made here to learn in the service of extracting unprotectable facts from them, right? So that's point one. I don't think anybody is really disputing that.

Second is that the output, except in the

really rarest of circumstances, is not going to be substantially similar in a copyright sense from anything that the AI model has been trained on, and so what are we left with? We're left with -- well, I think Cynthia made the point that, well, it may replicate the style of an author, you know, that the tool was trained on.

Well, I mean, like, it would really extend copyright, you know, copyright law beyond its recognizable bounds to say that creating something in the style of an author is copyright infringement. If I were to, you know, compile the collected works of Stephen King simply so I can emulate his style of writing, I don't think anybody would say that I have infringed his copyrights by doing that, and so I think the fact that the same thing is achievable by a computer doesn't really alter that fundamental copyright point. And so I think all of that, again, points very strongly towards the conclusion that I think the entire industry's been operating under, which is that what is happening to train these AI models is quintessentially fair use.

MS. PUSEY: Thanks, Sy.

I'm going to pass it to Mehtab and then to Brandy for the next question.

MS. KHAN: I just wanted to make a quick point about the technology and how I would caution against equating copyrighted works with less or more bias and examining that relationship more critically. The creative works embeds certain world views and points about communities or views about certain people regardless of the copyright status of that work, and so using copyrighted works does not necessarily mean that the output that we produce is going to be biased or not biased or not have certain representations. What the output will do is simply entrench or reproduce some of the existing features and representations that are in the input stages that already exist there.

MS. PUSEY: Thank you.

Brandy?

MS. KARL: Thank you.

So our next question, is are you aware of the Office's registration guidance with respect to works containing AI-generated material? What questions or concerns do you have about that guidance? And it looks like we have Cynthia. Oh, wait. No, you were still just hands up. Terry. I'm sorry.

MR. HART: Thanks. So, yes, we're aware of the Office's registration guidance. I think, in

principle, like many of the panelists on the first panel recognized, that the Office has taken the correct approach and, really, the devil is in the details, so I wanted to just point out one suggestion I had and then highlight what I thought were some of the biggest concerns there.

So my suggestion is to encourage the Office to commit to transparency and stakeholder consultation going forward, as it has with this guidance, as it's done with its compendium and its other resources which are extraordinarily helpful.

But I would extend that here also to its development of its own internal registration policies and procedures and training materials because, in this field, where things are kind of advancing very rapidly, where there's a lot of unknown unknowns about the registration and copyright questions that'll come up, I think it's really important to have that level of transparency into how the Copyright Office is approaching things and that ability to consult with copyright registrants so that they have a level of certainty about how they're registering their works and are able to offer, you know, really useful feedback.

So, you know, I'm not suggesting, you know,

throw open the doors and throw all your internal policies out there. But, you know, I think there's certain things the Office can do in terms of, you know, just informal consultations with stakeholder groups to get feedback on discrete policies or maybe even a working group of stakeholders to see, you know, as things progress and as the Office confronts more and more registrations where there may be AI-generated work incorporated in some fashion that it's able to refine and develop its practices in a way that makes sense to the overall copyright system.

So that said, you know, I just wanted to highlight some quick concerns. Some of these have been mentioned already. One, of course, you know, there's, I think, a lot of lack of clarity as to what degree of human interaction or editing is going to create copyright authorship over a generative AI work. Relatedly, I think there's a concern that in some instances that distinction may be indeterminate, you know, as far as, you know, kind of working iteratively with generative AI tools. Like, what part of that is attributed to human authorship? What is attributed to just the tool operating in a way that doesn't give rise to human authorship?

I've heard concerns that the disclosure rule

for disclosing AI-generated content may differ from the other types of disclaimers that registrants are already supposed to make, so, for example, disclaiming public domain materials where maybe a generalized statement may suffice.

MS. KARL: Terry?

MR. HART: And finally -- could I just -- one quick last point. I just finally wanted to point out some of the burdens that the rule may put on registrants going forward and then, with my members, in particular, with publishers, they may actually lack the knowledge of what their authors that they are putting out there have used in terms of AI-generated tools, but the publishers themselves are making the application, so there may be a burden there and some uncertainty. Thanks.

MS. KARL: Thanks, Terry.

Derek, and also can I have the remaining hands two minutes because we need to move on to the rest. We're tight on time. Thank you.

MR. SLATER: Yes, just to echo the, yes, aware of it, aware of the guidance. We were happy to see affirmatively put forward there to have some sort of requirement for significant human creativity, the Copyright Act be incentivizing human ingenuity and

creativity. That should remain the case, so we were happy to see that. We also think that helps fuel the commons of more stuff that people can build on in useful ways, and that's consistent with the Copyright Act and with Creative Commons' mission.

I think, as been discussed, it's going to get more complicated as people mix more deliberately their creativity with the automatically generated works. I think, you know, that's true in literary works, and in the subset, as Jule said earlier, of software, I think it's even more dicey, so I'll just echo, I think, I'm sure we'll get to have more discussion about this, but I appreciate you driving the conversation forward.

MS. KARL: Thank you, Derek.

Chris?

MR. MOHR: Thanks. Briefly, so as I've already talked about we are supportive of the conclusion in the guidance. There's a couple areas where we think it could be fleshed out a bit more. There was some concern, I think, that in places the guidance could be read as a bit draconian in tone in terms of its consequence, and that is a problem retroactively for folks who may already have registered their works under a different set of

assumptions, and so that would be a problem that I would encourage the Office to reflect on as it fleshes this issue out.

The second piece, and I think Terry alluded to this, is that there is a mention of de minimis contribution in the guidance. I think we know what that means. Spell or grammar checking is a de minimis contribution. But, if the Office has particular examples in mind, it might be useful to flesh those out in either a revision or in the Compendium. And, finally, I think, you know, part of this conversation, I think, is a little bit confusing, and it may just be because it's a conversation, not a series of legal briefs going back and forth.

But I got to tell you I had a hard time finding much of what Sy said as inaccurate in terms of describing how these models work, whether the reproduction right is implicated. I think, in most cases, we're assuming that it is, unless there's an excuse, whether through an implied license or through a fair use analysis, but in terms of finding statistical relationships between different pieces of work, language, what's likely to come next, that, I think, is standard for how all of this stuff works, and so, I mean, if we are proceeding off of mistaken

assumptions, it would be helpful, I think, to have some record of why specifically those assumptions are wrong, and at least for myself, I didn't necessarily hear that at least with respect to the technical operation of these bots.

In any event, I know we're getting close to wrap up, so thank you again for inviting us to this conversation, and we look forward to further engagement with the Office as this develops.

MS. KARL: Thank you, Chris.

Just to preview what's happening next, we're going to take Tracy, Betsy, and Peter, and then we will transition to our last question, along with combo closing statements and it is going to be tight, so please let's try to get through our statements. Thank you so much. Tracy?

MS. CHABALA: Yeah. So this isn't so much of a concern as much it is just a thought that as we move into the future, you know, this prompt engineering, despite everything I've sort of said, can be a really creative endeavor that's quite innovative, and so I can see eventually somebody kind of arguing that that work in and of itself is substantial and, therefore, you know, something deserves a copyright on those grounds, especially because I have kind of just,

like, watched what goes on on MidJourney, and I can see there's all these revisions. There's, you know, all sorts of effort that does go in there. It's just a thought.

MS. KARL: Thank you, Tracy.

Betsy?

MS. ROSENBLATT: Yeah, I agree with Derek's comments about the importance of valuing expression that originates with humans. I also want to identify a couple of places where I think the guidance is going to require additional difficult line-drawing. Many, many works are and long have been generated with the assistance of AI but not by AI. We need to consider that line. The line between original authorship and detailed prompting, as Tracy just brought up, the line between selection and arrangement of otherwise uncopyrightable AI-created works and the creation of AI works themselves. And, also, I would encourage the Office to consider to what extent the rules encourage lying and/or self-delusion among authors.

MS. KARL: Thank you, Betsy.

Peter?

MR. ROUTHIER: Thanks. Yes, we've seen the guidance and like I think almost everybody has said, we generally support it, in particular because it's

based on the longstanding principle that copyright law is for human authorship and that that's part of what furthers -- that's what copyright's supposed to be about, and I just wanted to note that I'm sort of actually struck by the fact that I've been watching, I think, both panels. I've heard almost everybody comment on it, and it seems like there's broad agreement that the registration guidance is pretty good in that regard, and I think that's evidence that our existing copyright rules and the existing copyright structure is actually working just fine in this area right now.

MS. KARL: Thank you, Peter.

Keyana?

MS. PUSEY: Thank you.

So, again, just asking this last question and ask that you briefly provide any answers you may have with your closing statements. So the question is, how is the training or the output of artificial intelligence affecting your field or industry? And, Tracy, did you already have your hand up from the last time or should we go to Betsy? I'm sorry, I couldn't tell.

MS. CHABALA: No, you can go to Betsy. I said -- yeah.

MS. PUSEY: Betsy?

MS. ROSENBLATT: Thank you. Right now, I think what we're hearing from fans is that some fans want to be able to opt out of having their work scraped and they have expressed the idea that if their works are going to be scraped they might not be worth making. On the other hand, we've heard a lot of enthusiasm about the potential for what AI can do and bring to fan communities.

Some of the concern, I think, is tied to the very idea of scraping being emotionally charged, but more and more I think it seems tied to the idea that scraping their works will result in generating works very similar to their own without attribution, and I think we should consider the role of attribution in this area in a way that perhaps copyright law may not do a lot of yet. Thanks.

MS. PUSEY: Thank you.

Cynthia?

MS. ARATO: Thank you. Generative AI is impairing the traditional licensing markets that exist between content creators and the other companies. It's also harming the relationship between publishers and users by providing more proprietary content from original sources without attribution and

disassociating the output from their sources.

One point, I think we do dispute that the output from these systems can be -- that is not -- sorry. We do dispute that the output from these systems would not be substantially similar to creative content that we own. Sy gave a visual example of artwork but no example regarding text, and we think that it's very easy to have text that's generated by the AI be substantially similar.

And then the last point I just want to make is echoing what Edward said in the original session. There are tremendous roadblocks to registering dynamic websites. It cannot be done in an easy, efficient, and group manner, and I think, while everyone is free to disagree about what might be fair use or not in generative AI, I think everyone can agree that there shouldn't be roadblocks put in place to register web pages and therefore sort of artificially put the thumb on the scale against content owners who are not able to easily register the work so that they can pursue whatever claims they feel they should be able to pursue in court on these issues. Thank you.

MS. PUSEY: Thanks.

Sy?

MR. DAMLE: So just on the point of whether

there is any evidence at all that these large language models in the text space can generate or regularly generate except in the rarest of circumstances output that is substantially similar to the input, I think, I mean, I haven't seen any, and it's certainly something we've looked at to see whether researchers have been able to do it, and just based on sort of how I describe these large-language models are trained, it would be pretty remarkable, it would be almost impossible for the large language models to put out output that is really substantially similar to a individual particular work in the input corpus, so I just want to make that point.

Second, just to answer the specific question, you know, a16z as a venture capital firm has a really broad picture into both companies that are building these tools and also companies that are using these tools, and I can tell you from the companies that are using these tools that it really is being used in a way to increase productivity, increase creativity. I know this is a panel about literary works, but just to give one example, you know, there are game developers that are using generative AI tools as part of the art production process not to replace artists but to help those artists generate new ideas

or realize their vision more easily.

And, beyond that, I just want to emphasize that the benefits to society for these tools are essentially limitless. If you think about the medical field, companies are using these kinds of tools to help doctors more quickly reach diagnoses when they're looking at, you know, the X-rays and CT scans and things like that. In the legal field, you know, which I'm sure we're all interested in, AI is being used to speed tasks like document collection and document review. All this makes medical and legal services cheaper, easier to access, and all of these tools depend on the ability to train on data.

And the sort of final point I'd make is, you know, if we're thinking about imposing new costs on the creators of AI models, I think one of two things is going to happen. I think either these tools just won't be able to be built, and I think that's probably the most likely outcome because, because of the way these tools are built, they require just way too much data for any licensing scheme to be able to work.

At best, what will happen is that the ability to build these tools will be preserved for those companies that have the deepest pockets and the greatest incentive to keep AI models closed, so the

result of that will be less competition, far less innovation, and closed AI models, which are hard to investigate. So I think we ought to be very, very cautious about imposing new costs on the creators of these tools, you know, without being mindful of those downstream consequences.

MS. PUSEY: Thanks, Sy.

Terry?

MR. HART: Thanks. So real briefly, just in the scientific and scholarly publishing world, I wanted to say that it's clear AI's going to increasingly be used to examine scientific and scholarly works in the public space, notably through open science databases in search of new breakthroughs and cures. The STEM community is only beginning to understand how AI will be trained and understand how the most authoritative and scientifically accurate works will be incorporated.

But, more broadly, I think I would again go back to the point I started with to caution about publishers and creators being sacrificed in a race for AI. I think, you know, as a panelist mentioned, there are certainly a host of issues regarding ethical and responsible deployment of AI outside of the copyright space, you know, and I think those each individually

were a thoughtful consideration as a society, and I don't think copyright should be any different from that. I think publishers, authors, creators of all types need to be part of these conversations, should not be written off at the outset because they, I think, share the hope of the opportunities that AI brings not only to their own industries but to society as a whole. So I will end with that. Thank you to the Copyright Office for this discussion.

MS. PUSEY: Thanks, Terry.

We're going to do Catherine, Peter, and Chris, and just ask that you please keep it brief as we do have closing remarks coming up next.

MS. ZALLER ROWLAND: Thanks. I will be as brief as possible here. I appreciate the opportunity, and as I said previously, the licensing landscape is incredibly important here and is a way that can help these markets function and to be able to do things in an ethical and a compliant way while also being able to advance technology. I just want to quickly say that, you know, the impacts of AI technologies on the constitutional purposes of copyright are really tied to the overall copyright system, so appropriate respect for copyright, including by using voluntary licensing, is very, very important, and it's going to

incentivize the creation industry to use the things that can feed into additional works that could be used by AI. You need to have copyright protection to promote the innovation and the creation of works that will be used to be training other AI machines in the future, so I think copyright plays an incredibly crucial role in that, and having respect for that system and potential options to use it in a compliant way are key.

MS. PUSEY: Thanks, Catherine.

Peter?

MR. ROUTHIER: Thanks. Yeah, thank you and I'll be brief and we can consider these certainly my closing remarks, so thank you very much for hosting this event and for having us all here today and thanks to everybody else for their thoughts.

So the two points that I have are just about making sure we're thinking about and urging that the Office make sure that it's thinking about the sort of full panoply of interests and parties that are involved with these issues.

So, in the first one, you know, as people have remarked, the training data, it's not always fully transparent and clear where it's coming from, but we know that they come from authors of many types,

for example, Wikipedia and other Creative Commons license-sourced, open-source software of various types, lots of general web content are all often disclosed as having been used to train machine learning models.

So, when representatives of some of the older industries suggest sort of opt-in or compensation-based schemes to replace the status quo, I think it's worth keeping in mind that those voices are not fully representative of the interests of the author communities that are included within, as far as we can understand, a lot of the training data that's used to train these models, and it also raises all kinds of issues about practicability and things like that. So this is not to say, of course, that their perspectives are not important, just to sort of put in context that this is a new area and the scope is quite extraordinary.

The second point is a slightly smaller point but also I think an important one. A lot of the questions and participants were sort of focused on industries today, and I know that's a useful shorthand and it can mean a lot of different things, but I just want to suggest that we make sure we're just as interested in non-commercial uses as we are in

commercial uses.

In the European Union, for example, when they did copyright reform a few years ago, they made a distinction between the exception for text and data mining, which was applicable to research organizations and cultural heritage institutions, and the exceptions available for commercial uses, and, of course, fair use when it's properly applied makes a similar distinction between commerciality and non-commerciality. So, as the Office continues its study of the issue, I just want to urge it to keep in mind that it make appropriate distinctions where necessary between commercial and non-commercial uses.

So that's it for me, and thanks again for hosting this session today.

MS. PUSEY: Thanks, Peter.

And lastly, Chris?

MR. MOHR: Thanks, and so a few points to conclude. I mean, I think, as we think about this, as our members think about this, there are two where the -- in the existing ecosystem right now, there's a distinction, to really summarize it, between gates up and gates down, and you see that distinction in cases like Field. You also see it in other doctrines, like cases like Van Buren interpreting the Computer Fraud

and Abuse Act, and so we're optimistic about -- it is important, I think, that when the gates come down that unauthorized use stop, and those gates can come down in a couple of different ways. One is a license and control. Another may be the use of signals, and I would refer you to according to what is it, I think it's the MIT *Technology Review*, stable diffusion is now going to be using a tag, a do not train tag similar -- and we are optimistic about the potential for this type of standard to alleviate some of these issues, and that's different from whether or not something can be found for purposes of being retrievable.

Again, when the gates are up, however, we don't see a need to re-examine -- wrong word -- revise the existing copyright regime, but, again -- and that's it, so thank you for the opportunity to contribute to this conversation, and, again, we look forward to more of them.

MS. PUSEY: Thanks, Chris. I'm going to pass it to Mark for the closing remarks.

MR. GRAY: Great. Thank you, Keyana.

So, first off, thank you so much for everyone today, all the panelists, as well as the several hundred folks in the audience. We really

appreciate you all joining us for this. We had a lot of insights and perspectives today. We will be keeping those in mind obviously as we continue our work on our AI initiative and we keep thinking about these copyright law and policy issues that are raised by artificial intelligence and different sorts of technologies within that field.

Looking to the future, I just wanted to let everyone know our next listening session is going to be on Tuesday, May 2. That is going to be focused on the visual arts. And then going forward, we will have a session on audiovisual works on May 17, and we will have a session on music and sound recordings on May 31. You can sign up for those on the Copyright Office website both to attend and to sign up to request to speak for the last two sessions.

The visual arts session signup is closed for participants. Keep in mind this is not the last chance to talk to us. This session was not the last chance. These listening sessions generally are not the last chance. We will keep providing opportunities throughout the next year to talk to us. There will be other chances to submit all sorts of comments and ideas to the Office, so please keep in mind, if you're in the audience and you didn't get a chance to speak,

there will be further opportunities. So thank you again, everyone. We really appreciate you joining us, and we hope you have a wonderful afternoon or evening depending on your time zone. Thank you.

(Whereupon, at 4:05 p.m., the listening session in the above-entitled matter adjourned.)

//

//

//

//

//

//

//

//

//

//

//

//

//

//

CERTIFICATE

CASE TITLE: Copyright and Artificial Intelligence Literary Works, Including Software, Listening Session

HEARING DATE: April 19, 2023

LOCATION: Washington, D.C.

I hereby certify that the proceedings and evidence are contained fully and accurately on the tapes and notes reported by me at the hearing in the above case before the Library of Congress, U.S. Copyright Office.

Date: April 19, 2023

Gigi Sanchez
Official Reporter
Heritage Reporting Corporation
Suite 206
1220 L Street, N.W.
Washington, D.C. 20005-4018