

**ARTIFICIAL INTELLIGENCE AND INTELLECTUAL
PROPERTY—PART II: COPYRIGHT**

HEARING

BEFORE THE

SUBCOMMITTEE ON INTELLECTUAL PROPERTY

OF THE

COMMITTEE ON THE JUDICIARY

UNITED STATES SENATE

ONE HUNDRED EIGHTEENTH CONGRESS

FIRST SESSION

—————
JULY 12, 2023
—————

Serial No. J-118-25

—————

Printed for the use of the Committee on the Judiciary



U.S. GOVERNMENT PUBLISHING OFFICE

COMMITTEE ON THE JUDICIARY

RICHARD J. DURBIN, Illinois, *Chair*

DIANNE FEINSTEIN, California	LINDSEY O. GRAHAM, South Carolina,
SHELDON WHITEHOUSE, Rhode Island	<i>Ranking Member</i>
AMY KLOBUCHAR, Minnesota	CHARLES E. GRASSLEY, Iowa
CHRISTOPHER A. COONS, Delaware	JOHN CORNYN, Texas
RICHARD BLUMENTHAL, Connecticut	MICHAEL S. LEE, Utah
MAZIE K. HIRONO, Hawaii	TED CRUZ, Texas
CORY A. BOOKER, New Jersey	JOSH HAWLEY, Missouri
ALEX PADILLA, California	TOM COTTON, Arkansas
JON OSSOFF, Georgia	JOHN KENNEDY, Louisiana
PETER WELCH, Vermont	THOM TILLIS, North Carolina
	MARSHA BLACKBURN, Tennessee

JOSEPH ZOGBY, *Chief Counsel and Staff Director*

KATHERINE NIKAS, *Republican Chief Counsel and Staff Director*

SUBCOMMITTEE ON INTELLECTUAL PROPERTY

CHRISTOPHER A. COONS, Delaware, *Chair*

MAZIE K. HIRONO, Hawaii	THOM TILLIS, North Carolina, <i>Ranking</i>
ALEX PADILLA, California	<i>Member</i>
JON OSSOFF, Georgia	JOHN CORNYN, Texas
PETER WELCH, Vermont	TOM COTTON, Arkansas
	MARSHA BLACKBURN, Tennessee

JAMES BARTON, *Democratic Chief Counsel*

SETH WILLIFORD, *Republican General Counsel*

CONTENTS

JULY 12, 2023, 3:03 P.M.

STATEMENTS OF COMMITTEE MEMBERS

	Page
Coons, Hon. Christopher A., a U.S. Senator from the State of Delaware	1
Tillis, Hon. Thom, a U.S. Senator from the State of North Carolina	3

WITNESSES

Witness List	33
Brooks, Ben, head of public policy, Stability AI, San Francisco, California	5
prepared statement	34
Harleston, Jeffrey, general counsel and executive vice president, business and legal affairs, Universal Music Group, Santa Monica, California	12
prepared statement	45
Ortiz, Karla, concept artist, illustrator, and fine artist, San Francisco, Cali- fornia	10
prepared statement	51
Rao, Dana, executive vice president, general counsel, and chief trust officer, Adobe, Inc., San José, California	7
prepared statement	63
Sag, Matthew, professor of law, artificial intelligence, machine learning, and data science, Emory University School of Law, Atlanta, Georgia	9
prepared statement	71

QUESTIONS

Questions submitted to Ben Brooks by Senator Tillis	98
Questions submitted to Jeffrey Harleston by Senator Tillis	104
Questions submitted to Karla Ortiz by Senator Tillis	110
Questions submitted to Dana Rao by Senator Tillis	115
Questions submitted to Matthew Sag by Senator Tillis	121

ANSWERS

Responses of Ben Brooks to questions submitted by Senator Tillis	128
Responses of Jeffrey Harleston to questions submitted by Senator Tillis	138
Responses of Karla Ortiz to questions submitted by Senator Tillis	151
Responses of Dana Rao to questions submitted by Senator Tillis	161
Responses of Matthew Sag to questions submitted by Senator Tillis	171

MISCELLANEOUS SUBMISSIONS FOR THE RECORD

Submitted by Chair Coons:

Computer and Communications Industry Association (CCIA), letter, July 12, 2023	192
Copyright Alliance, letter, July 17, 2023, and attachment	195
Digital Media Association (DiMA), letter, July 12, 2023	201
Motion Picture Association (MPA), statement	203
Screen Actors Guild–American Federation of Television and Radio Artists (SAG–AFTRA), statement	211
Songwriters Guild of America (SGA), et al., letter, July 18, 2023	214

**ARTIFICIAL INTELLIGENCE
AND INTELLECTUAL
PROPERTY—PART II: COPYRIGHT**

WEDNESDAY, JULY 12, 2023

UNITED STATES SENATE,
SUBCOMMITTEE ON INTELLECTUAL PROPERTY,
COMMITTEE ON THE JUDICIARY,
Washington, DC.

The Subcommittee met, pursuant to notice, at 3:03 p.m., in Room 226, Dirksen Senate Office Building, Hon. Christopher A. Coons, Chair of the Subcommittee, presiding.

Present: Senators Coons [presiding], Klobuchar, Hirono, Padilla, Tillis, and Blackburn.

Also present: Chair Durbin.

**OPENING STATEMENT OF HON. CHRISTOPHER A. COONS,
A U.S. SENATOR FROM THE STATE OF DELAWARE**

Chair COONS. This hearing will come to order. I'd like to thank all of our witnesses for participating today, and I'd like to especially thank my friend and colleague, Ranking Member Thom Tillis and his staff for working with us on such a collaborative basis to put this hearing together.

Welcome back from Vilnius. Senator Tillis was over at the NATO summit, and I am thrilled he is able to join us, and we are able to do this hearing today. This is our second hearing in as many months on the intersection of artificial intelligence and intellectual property law and policy.

You and your team have been great partners in pursuing this. If you will indulge me for a moment, Senator, and before I proceed with my remarks, I'd like to ask that we play just a little clip of something to frame the challenges of this topic—made with the permission of all the relevant rights holders.

[Laughter.]

[Video is shown.]

Chair COONS. Thank you for your forbearance. Yes, a round of applause is certainly welcomed.

[Applause.]

Chair COONS. My team actually produced a version of that where it is a duet between me and Frank Sinatra, but my voice came out so horribly flat, I didn't want to impose that on any of you.

Senator TILLIS. Thank you, Mr. Chair, for your judgment.

[Laughter.]

Chair COONS. Creating the song, “AI, AI” to the tune of “New York, New York” was great fun and I appreciate my team that worked so hard on pulling that off, but the very existence of it highlights a couple of the core questions around copyright raised by generative artificial intelligence.

ChatGPT wrote the lyrics following the style of “New York, New York,” although perhaps not quite as moving and inspiring as the original words to any but IP enthusiasts. Another generative AI tool was used to take Mr. Sinatra’s recorded songs and his voice and his phrasing and his style and set that to music.

So, a couple of those core questions: Did ChatGPT infringe the copyright on “New York, New York” when it drafted lyrics representing its lyrical style? What about the AI tool that set those lyrics to music? Did either that tool or did I run afoul of Mr. Sinatra’s rights by mimicking his voice?

In my case, no, because we got specific approval. Or did I just use AI tools to enhance my own creativity? And if so, should this newly created song be entitled to copyright protection? These are just a few of the questions I hope we will explore with our panel of talented and insightful witnesses as we consider the impact of artificial intelligence on copyright law and policy and the creative community.

As we all know, AI is rapidly developing. Generative AI tools have brought many new people into the creative fold and have opened new avenues for innovation. People who may not have ever considered themselves creatives are now using these tools to produce new works of art. Artists themselves have used AI tools to enhance their own creativity.

Paul McCartney recently made headlines announcing that AI helped create the very last Beatles song 50 years after the band broke up. As I have previewed, AI creates new copyright law issues, including whether using copyrighted content to train AI models is copyright infringement, whether AI-generated content should be given copyright protection, and many more.

These questions are working their way through the courts. How the courts decide them, and the decisions we make here in the Senate and in Congress about how to respect and reinforce existing copyright protections and works will have significant consequences, not just for the creative community, but our overall competitiveness as a country. While generative AI models are trained on copyrighted content, IP considerations haven’t been included or sufficiently considered in proposed IP regulatory frameworks here in the U.S.

In contrast, some of our competitors recognize IP policy as an important tool. The EU is currently planning to require AI companies to publish the copyrighted materials used in training models. The UK provides copyright protection for computer-generated works. These are just some initial concerns, and I think there are initial steps that we can take to ensure sustained U.S. leadership on artificial intelligence.

First, it is critical to include IP considerations in any regulatory framework for artificial intelligence and to give our Copyright Office, in this framework, a seat at the table. We should also consider whether changes to our copyright laws or whole new protections

like a Federal right of publicity may be necessary to strike the right balance between creators' rights and AI's ability to enhance innovation and creativity.

I am excited to explore these issues today. We have got a great panel, and great partner, and great Members of the Committee. With Senator Tillis' cooperation, we have assembled this wonderful panel. I will introduce them shortly, but first I will turn it over to my Ranking Member, Senator Thom Tillis.

**OPENING STATEMENT OF HON. THOM TILLIS,
A U.S. SENATOR FROM THE STATE OF NORTH CAROLINA**

Senator TILLIS. Thank you, Mr. Chairman, and thanks for everyone here. It is great to see the number of participants in the audience. It is even more amazing to see an equal number who would like to get in here, but who says AI and IP can't be sexy?

But, you know, in all seriousness, I appreciate that we are having another hearing on the opportunity to highlight the importance of intellectual property when it comes to emerging technologies, and today we are talking about AI. During our last hearing, we really discussed the impact of AI on patent—in a patent context, which explored ideas such as whether or not AI can be considered an inventor.

And it cannot, and hopefully it will not in the future. But while many of these issues we discussed in the last hearing were perspective, the creative community is experiencing immediate and acute challenges due to the impact generative AI in a copyright context.

Strong, reliable, predictable IP rights are paramount to incentivizing U.S. innovation and creativity. It is this innovation and creativity that fuels growth of our country's prosperity and drives enormous economic growth. In fact, core copyright industries added \$1.8 trillion of value to U.S. GDP, accounting for almost 8 percent of the U.S. economy.

These copyright industries also employ 9.6 million American workers. The sales of major U.S. copyright products overseas in markets also constitute \$230 billion and outpaced exports of other major U.S. industries. Advances in generative AI have raised new questions regarding how copyright principles such as authorship, infringement, and fair use will apply to content created or used by AI.

We must not only consider how our current IP laws apply to the field of generative AI, but also what changes, if any, may be necessary to incentivize future AI innovators and creators.

So, Chairman Coons, I am happy to have this Committee. I will submit the remainder of my statement for the record. But for those of you who have watched our Committee over the past several Congresses where either Senator Coons or I were in Ranking Member or Chairmanship, I think if anything, I hope people understand that we are very thorough and we are very persistent in our approach, and we are inclusive.

I have told everyone on this issue, whichever end of the spectrum you are, if you are at the table and the work groups, we are going to find a reasonable solution and compromises. If you are outside of the work group process and you are just taking shots at it, you may find yourself on the table, from my perspective.

So, we encourage you to get to the table and make what we are doing better. The reason why I think it is so important, and I am glad the IP Subcommittee is leading on this in terms of formal hearings with a focus on potentially drafting legislation, is I think we run the risk of some in Congress who think AI is bad—that it’s a threat to the future.

I am not in that camp. I think that AI is good. It is something that I first developed expertise in back in the late 80s, and have followed it every sense. It is a matter of getting the regulatory construct, the intellectual property construct, all the other underlying policies that you need when a new, I think positive—in a positive term, disruptive technology hits the field.

So, the reason that we need to move forward, address potential concerns is precisely because I want the United States to lead in innovation. And so much innovation is going to be premised on properly exploiting the capabilities responsibly, and that is what I hope we learn in this hearing and subsequent hearings and work group. So, thank you all for being here. And thank you, Mr. Chair, for having the hearing.

Chair COONS. Thank you, Senator Tillis. I am now going to turn to our witness panel today. We welcome five witnesses to testify about the intersection of artificial intelligence and copyright law.

Our first witness is Mr. Ben Brooks, head of public policy at Stability AI, a company that develops a range of AI models that help users generate images, text, audio, and video.

Next, we have Dana Rao, executive VP, general counsel, and chief trust officer at Adobe. I’d like to be chief trust officer in the United States.

Senator TILLIS. They don’t have titles like that.

[Laughter.]

Chair COONS. Mr. Rao leads Adobe’s legal security and policy organization, including Adobe’s Content Authenticity Initiative, which promotes transparency principles around the use of AI.

Next, we have Professor Matthew Sag, a professor of law in artificial intelligence, machine learning, and data science at Emory University School of Law. Professor Sag is a leading U.S. authority on the fair use doctrine in copyright law and its implications for researchers in text data mining, machine learning, and AI.

Next, we will hear from Karla Ortiz, an artist, a concept artist, illustrator, and fine artist who has worked on a variety of well-known and widely enjoyed projects, including “Jurassic World,” “Black Panther,” “Loki,” and she is most famous for designing, in my assessment at least, for designing Doctor Strange for Marvel’s first “Doctor Strange” film. Welcome.

Last but certainly not least, we have Jeffrey Harleston, general counsel and executive VP of business and legal affairs for Universal Music Group. Mr. Harleston is responsible for overseeing business transactions, contracts, litigation for all of Universal Music Group’s worldwide operations in more than 60 countries.

After I swear all of you in, each of you will have 5 minutes to make an opening statement. We will then proceed to questioning. Each Senator, depending on attendance, questioning, and time, will have a first round of 5 minutes. We may well have a second round

in 5 minutes, and we may be the only two left for a third round of 5 minutes, but we will see.

So, could all the witnesses please stand to be sworn in. Please raise your right hand. Do you swear or affirm that the testimony you are about to give before this Committee will be the truth, the whole truth, and nothing but the truth, so help you God?

[Witnesses are sworn in.]

Chair COONS. Thank you. Mr. Brooks, you may proceed with your opening statement.

**STATEMENT OF BEN BROOKS, HEAD OF PUBLIC POLICY,
STABILITY AI, SAN FRANCISCO, CALIFORNIA**

Mr. BROOKS. Thank you, Chair Coons and Ranking Member Tillis, for the opportunity to testify today. AI can help to unlock creativity, drive innovation, and open up new opportunities for creators and entrepreneurs across the United States. As with any groundbreaking technology, AI raises important questions, and we recognize the depth of concern among creators.

While we don't have all the answers, we are committed to an open and constructive dialog, and we are actively working to address emerging concerns through new technology standards and good practices.

At Stability AI, our goal is to unlock humanity's potential with AI technology. We developed a range of AI models. These models are essentially software programs that can help a user to create new content.

Our flagship model, Stable Diffusion, can take plain language instructions from a user and help to produce a new image. We are also working on research for safe language models that can help to produce new passages of text or software code. AI is a tool that can help to accelerate the creative process.

In our written testimony, we shared examples of how Broadway designers, architects, photographers, and researchers are using our models to boost their productivity, experiment with new concepts, or even study new approaches to diagnosing complex medical disorders. We are committed to releasing our models openly with appropriate safeguards.

That means we share the underlying software as a public resource. Creators, entrepreneurs, and researchers can customize these models to develop their own AI tools, build their own AI businesses, and find novel applications of AI that best support their work. Importantly, open models are transparent.

We can look under the hood to scrutinize the technology for safety, performance, and bias. These AI models study vast amounts of data to understand the subtle relationships between words, ideas, and visual or textual features, much like a person visiting an art gallery or library to learn how to draw or how to write.

They learn the irreducible facts and structures that make up our systems of communication. And through this process, they develop an adaptable body of knowledge that they can then apply to help produce new and unseen content.

In other words, compositions that did not appear in the training data and may not have appeared anywhere else. These models don't rely on a single work in their training data, nor did they store

their training data. But instead, they learn by observing recurring patterns over billions of images and trillions of words of text.

We believe that developing these models is an acceptable and socially beneficial use of existing content that is permitted by fair use and helps to promote the progress of science and useful arts. Fair use and a culture of open learning is essential to recent developments in AI. It is essential to help make AI useful, safe, unbiased, and it is doubtful that these groundbreaking technologies would be possible without it.

The U.S. has established global leadership in AI thanks in part to an adaptable and principles-based fair use doctrine that balances creative rights with open innovation. We acknowledge emerging concerns, and these are early days, and we don't have all the answers, but we are actively working to address these concerns through safe technology, standards, and good practices.

First, we have committed to voluntary opt-outs so that creators can choose if they don't want their online work to be used for AI training. We have received opt-out requests for over 160 million images to date, and we are incorporating these into upcoming training. We are hoping to develop digital opt-out labels as well that follow the content wherever it goes on the internet.

Second, we are implementing features to help users and tech platforms identify AI content. Images generated through our platform can be digitally stacked with metadata and watermarks to indicate if the content was generated with AI.

These signals can help ensure that users exercise appropriate care when interacting with AI content and help tech platforms distinguish AI content before amplifying it online. We welcome Adobe's leadership in driving the development of some of these open standards.

Third, we have developed layers of mitigations to make it easier to do the right thing with AI and harder to do the wrong thing.

Today, we filter datasets of unsafe content. We test and evaluate our models before release. We apply ethical use licenses, disclose known risks, filter content generated through our computing services, and implement new techniques to mitigate bias. As we integrate AI into the digital economy, we believe the community will continue to value human-generated art and perhaps value it at a premium.

Smartphones didn't destroy photography, and word processors didn't diminish literature, despite radically transforming the economics of creation. Instead, they gave rise to new demand for services, new markets for content, and new creators.

We expect the same will be true of AI, and we welcome an ongoing dialog with the creative community about the fair deployment of these technologies. Thank you for the opportunity to testify, and we welcome your questions.

[The prepared statement of Mr. Brooks appears as a submission for the record.]

Chair COONS. Thank you, Mr. Brooks. Mr. Rao.

**STATEMENT OF DANA RAO, EXECUTIVE VICE PRESIDENT,
GENERAL COUNSEL, AND CHIEF TRUST OFFICER, ADOBE,
INC., SAN JOSE, CALIFORNIA**

Mr. RAO. Chair Coons, Ranking Member Tillis, and Members of the Committee, thank you for the opportunity to testify here today.

My name is Dana Rao, and I am general counsel, and, as Senator Coons noted, chief trust officer at Adobe. I am happy to provide you with this secret certificate you need to get that title, if you would like, after the hearing.

Since our founding in 1982, Adobe has pioneered transformative technologies in all types of digital creation, from digital documents like PDF to image editing with Photoshop. Our products allow our customers who range from aspiring artists to wartime photojournalists, to advertisers and more, to unleash their creativity, protect their craft, empower their businesses in a digital world.

AI is the latest disruptive technology we have been incorporating into our tools help creators realize their potential. You have all seen the magic of text to image generative AI. Type in the prompt, cat driving a 1950s sportscar through the desert, and in seconds you will see multiple variations of a cat on a retro road trip appear before your eyes.

We have launched generative AI in our own tools, Adobe Firefly, and has provided—this proved to be wildly popular with our creative professionals and consumers alike. In my written testimony, I explore a comprehensive framework for responsible AI development that includes addressing misinformation, harmful bias, creative rights, and intellectual property.

Today, given Adobe's focus and our millions of creative customers and our leadership in AI, I will focus on how the United States can continue to lead the world in AI development by both supporting the access to data that AI requires and strengthening creator rights.

The question of data access is critical for the development of AI because AI is only as powerful and as good as the data on which it is trained. Like the human brain, AI learns from the information you give it.

In the AI's case, the data it is trained on. Training on a larger dataset can help ensure your results are more accurate because the AI has more facts to learn from. A larger dataset will also help the AI avoid perpetuating harmful biases in its results by giving it a wider breadth of experiences from which it can build its understanding of the world. More data means better answers and fewer biases.

Given those technical realities, United States and governments should support access to data to ensure that AI innovation can flourish accurately and responsibly. However, one of the most important implications of AI's need for data is the impact on copyright and creators' rights.

There are many outstanding questions in this space, including whether creating an AI model, which is a software program, from a set of images, is a permitted fair use. And whether that analysis changes if the output of that AI model creates an image that is substantially similar to an image on which it is trained.

These questions will certainly be addressed by courts and perhaps Congress, and we are prepared to help assist in those discussions. Adobe recognized the potential impact of AI on creators and society, and we have taken several steps.

First, we trained our own generative AI tool, Adobe Firefly, only on licensed images from our Adobe Stock Collection, which is a stock photography collection, openly licensed content, and works that are in the public domain where the copyright has expired. This approach supports creators and customers by training on a dataset that is designed to be commercially safe.

In addition, we are advocating for other steps we can all take to strengthen creators' rights. First, we believe creators should be able to attach a "Do Not Train" tag to their work. With industry and Government support, we can ensure AI data crawlers will read and respect this tag, giving creators the option to keep their data out of AI training datasets.

Second, creators using AI tools want to ensure they can obtain copyright protection over their work in this new era of AI-assisted digital creation. An AI output alone may not receive copyright protection, but we believe the combination of human expression and AI expression will and should.

Content editing tools should enable creators to obtain a copyright by allowing them to distinguish the AI work from the human work. In my written testimony, I discuss our open standards-based technology content credentials, which can help enable both of these creator protections.

Finally, even though Adobe has trained its AI on permitted work, we understand the concern that an artist can be economically dispossessed by an AI trained on their work that generates arts in their style, in the Frank Sinatra example you gave.

We believe artists should be protected against this type of economic harm, and we propose Congress establish a new Federal anti-impersonation right that would give artists a right to enforce against someone intentionally attempting to impersonate their style or likeness.

Holding people accountable who misuse AI tools is a solution we believe goes to the heart of some of the issues our customers have, and this new right would help address that concern. The United States has led the world through technological transformations in the past, and we have all learned it is important to be proactively responsible to the impact of these technologies.

Pairing innovation with responsible innovation will ensure that AI ultimately becomes a transformative and true benefit to our society. Thank you, Chair Coons, Ranking Member Tillis, and Members of the Committee.

[The prepared statement of Mr. Rao appears as a submission for the record.]

Chair COONS. Thank you, Mr. Rao. Professor.

STATEMENT OF MATTHEW SAG, PROFESSOR OF LAW, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND DATA SCIENCE, EMORY UNIVERSITY SCHOOL OF LAW, ATLANTA, GEORGIA

Professor SAG. Chair Coons, Ranking Member Tillis, Members of the Subcommittee, thank you for the opportunity to testify here today. I am a professor of law in AI, machine learning, and data science at Emory University, where I was hired as part of Emory's AI Humanity Initiative.

Although we are still a long way from the science fiction version of artificial general intelligence that thinks, feels, and refuses to open the pod bay doors, recent advances in machine learning and artificial intelligence have captured the public's attention and apparently lawmakers' interest.

We now have large language models, or LLMs, that can pass the bar exam, carry on a conversation, create new music and new visual art. Nonetheless, copyright law does not and should not recognize computer systems as authors. Even where an AI produces images, text, or music that is indistinguishable from human authored works, it makes no sense to think of a machine learning program as the author.

The Copyright Act rightly reserves copyrights for original works of authorship. As the Supreme Court explained long ago in the 1884 case of *Burrow-Giles Lithographic*, authorship entails original, intellectual conception. An AI can't produce a work that reflects its own original intellectual conception because it has none.

Thus, when AI models produce content with little or no human oversight, there is no copyright in those outputs. However, humans using AI as tools of expression may claim authorship if the final form of the work reflects their original intellectual conception in sufficient detail. And I have elaborated in my written submissions how this will depend on the circumstances.

Training generative AI on copyrighted works is usually fair use because it falls into the category of non-expressive use. Courts addressing technologies such as reverse engineering, search engines, and plagiarism detection software have held that these non-expressive uses are fair use. These cases reflect copyright's fundamental distinction between protectable original expression and unprotectable facts, ideas, and abstractions.

Whether training an LLM is in non-expressive use depends ultimately on the outputs of the model. If an LLM is trained properly and operated with appropriate safeguards, its outputs will not resemble its inputs in a way that would trigger a copyright liability. Training such an LLM on copyrighted works would thus be justified under current fair use principles.

It is important to understand that generative AI are not designed to copy original expression. One of the most common misconceptions about generative AI is the notion that the training data is somehow copied into the model. Machine learning models are influenced by the data. They would be pretty useless without it. But they typically don't copy the data in any literal sense.

So rather than thinking of an LLM as copying the training data like a scribe in a monastery, it makes more sense to think of it as learning from the training data like a student. If an LLM like

GPT3 is working as intended, it doesn't copy the training data at all. The only copying that takes place is when the training corpus is assembled and pre-processed, and that is what you need a fair use justification for. Whether a generative AI produces truly new content or simply conjures up an infringing cut and paste of works in the training data depends on how it is trained.

Accordingly, companies should adopt best practices to reduce the risk of copyright infringement and other related harms, and I have elaborated on some of these best practices in my written submission. Failure to adopt best practices may potentially undermine claims of fair use.

Generative AI does not, in my opinion, require a major overhaul of the U.S. copyright system at this time.

If Congress is considering new legislation in relation to AI and copyright, that legislation should be targeted at clarifying the application of existing fair use jurisprudence, not overhauling it.

Israel, Singapore, and South Korea have recently incorporated fair use into their copyright statutes because these countries recognize that the flexibility of the fair use doctrine gives U.S. companies and U.S. researchers a significant competitive advantage.

Several other jurisdictions, most notably Japan, the United Kingdom, and the European Union, have specifically adopted exemptions for text data mining that allow use of copyrighted works as training for machine learning and other purposes.

Copyright law should encourage the developers of generative AI to act responsibly. However, if our laws become overly restrictive, then corporations and researchers will simply move key aspects of technology development overseas to our competitors.

Thank you very much.

[The prepared statement of Professor Sag appears as a submission for the record.]

Chair COONS. Thank you, Professor. Ms. Ortiz.

STATEMENT OF KARLA ORTIZ, CONCEPT ARTIST, ILLUSTRATOR, AND FINE ARTIST, SAN FRANCISCO, CALIFORNIA

Ms. ORTIZ. Yes. Chairman Coons, Ranking Member Tillis, and esteemed Members of the Committee, it is an honor to testify before you today about AI and copyright. My name is Karla Ortiz. I am a concept artist, illustrator, and fine artist, and you may not know my name, but you know my work.

My paintings have shaped the worlds of blockbuster Marvel films and TV shows, including "Guardians of the Galaxy 3," "Black Panther," "Loki," you know, but specifically, the one I am most happiest of is that I, my work helped shape the look of Doctor Strange in the first "Doctor Strange" movie.

I have to brag about that a little bit, sir. I love what I do. I love my craft. Artists train their entire lives to be able to bring the imaginary to life. All of us who engage in this craft love every little bit of it. Through hard work, support of loved ones, and dedication, I have been able to make a good living from my craft via the entertainment industry, an industry that thrives when artists' rights to consent, credit, and compensation are respected.

I have never worried about my future as an artist until now. Generative AI is unlike any other technology that has come before.

It is a technology that uniquely consumes and exploits the hard work, creativity, and innovation of others. No other tool is like this. What I found, when first researching AI, horrified me.

I found that almost the entirety of my work, the work of almost every artist I know, and the work of hundreds of thousands of artists had been taken without our consent, credit, or compensation. These works were stolen and used to train for profit technologies with datasets that contain billions of image and text data pairs.

Through my research, I learned many AI companies gather copyrighted training data by relying on a practice called data laundering. This is where a company outsources its data collection to a third party under the pretext of research to then immediately use commercially. I found these companies use big terms like “publicly available data” or “openly licensed content” to disguise their extensive reliance on copyrighted works.

No matter what they are saying, these models are illegally trained on copyrighted works. To add even more insult to injury, I found that these for-profit companies were not only permitting users to use our full names to generate imagery but encouraging it. For example, Polish artist Frederic Koski had had his name used as a prompt in AI products over 400,000 times, and those are the lower end of the estimate.

My own name, Karla Ortiz, has also been used by these companies thousands of times. Never once did I give consent. Never once have I gotten credit. Never once have I gotten compensation. It should come as no surprise that major productions are replacing artists with generative AI.

Goldman Sachs estimates that generative AI will diminish or outright destroy approximately 300 million full-time jobs worldwide. As Ranking Member Tillis mentioned earlier, copyright-reliant industries alone contribute \$1.8 trillion value to the U.S. GDP, accounting for 7.76 percent of the entire U.S. economy. This is an industry that employs 9.6 million American workers alone.

The game plan is simple, to go as fast as possible, to create mesmerizing tales of progress, and to normalize the exploitation of artists as quickly as possible. They hope when we catch our breath, it will be too late to right the wrongs, and exploiting Americans will become an accepted way of doing things.

But that game can't succeed as we are here now, giving this the urgency it so desperately deserves. Congress should act to ensure what we call the 3Cs and a T: consent, credit, compensation, and transparency.

The work of artists like myself were taken without our consent, credit, nor compensation, and then used to compete with us directly in our own markets—an outrageous act that under any other context would immediately be seen as unfair, immoral, and illegal.

Senators, there is a fundamental fairness issue here. I am asking Congress to address this by enacting laws that require these companies to obtain consent, give credit, pay compensation, and be transparent. Thank you.

[The prepared statement of Ms. Ortiz appears as a submission for the record.]

Chair COONS. Thank you, Ms. Ortiz. Last but certainly not least, Mr. Harleston.

**STATEMENT OF JEFFREY HARLESTON, GENERAL COUNSEL
AND EXECUTIVE VICE PRESIDENT, BUSINESS AND LEGAL
AFFAIRS, UNIVERSAL MUSIC GROUP, SANTA MONICA, CALI-
FORNIA**

Mr. HARLESTON. Thank you, Chairman Coons, Ranking Member Tillis, and Members of the Committee. It is an honor to be here before you today. I am Jeff Harleston. I am the general counsel of Universal Music Group. And what is Universal Music Group? We are the world leader in music-based entertainment.

We are home to legendary record labels such as Motown, Def Jam, Island, Blue Note, Capitol, just to name a few. We have a music publishing company that signs songwriters, and we have a music merchandizing company as well, and an audio division—an audiovisual division that produces award-winning documentaries based on music.

UMG identifies, develops artists across every musical genre. I think it is fair to note that Frank Sinatra is one of our artists, and I think based on what we didn't hear today, I am not sure if we will be pursuing a developing artist out of Delaware named Chris Coons, but maybe we will get back to that.

Chair COONS. I am confident you will not.

[Laughter.]

Mr. HARLESTON. All jokes aside, I have been at the company, I have been honored to be with the company for 30 years, and most of the time I have spent as a lawyer, but I have also spent some time leading the Def Jam label and also as the management team of Geffen Records.

So, I have been on both sides of the business. We have also helped broker deals with digital services, platforms, social media outlets where you, all of you can access the music that you love. It has been my life's honor to work with countless talented and creative artists.

Their creativity is the soundtrack to our lives, and without the fundamentals of copyright, we might not have ever known them. I would like to make four key points to you today.

The first, copyright artists and human creativity must be protected. Art and human creativity are central to our identity. Artists and creators have rights. They must be respected. If I leave you with one message today, it is this: AI in the service of artists and creativity can be a very, very good thing. But AI that uses, or worse yet, appropriates the work of these artists and creators and their creative expression, their name, their image, their likeness, their voice, without authorization, without consent, simply is not a good thing.

The second point I want to make is that generative AI raises challenging issues in the copyright space. I think you have heard from the other panelists and they all would agree. We are the stewards at Universal of tens of thousands, if not hundreds of thousands, of copyrighted creative works from our songwriters and artists, and they have entrusted us to honor, value, and protect them.

Today they are being used to train generative AI systems without authorization. This irresponsible AI is violative of copyright law and completely unnecessary. There is a robust digital marketplace

today in which thousands of responsible companies properly obtain the rights they need to operate.

There is no reason that the same rules that apply to everyone else should not apply equally to AI companies and AI developers.

My third point, AI can be used responsibly to enhance artistic expression. Just like other technologies before, artists can use AI to enhance their art. AI tools have long been used in recording studios for drum tracks, chord progressions, and even creating immersive audio technologies. One of our distributed artists used a generative AI tool to simultaneously release a single in six languages in his own voice, on the same day. The generative AI tool extended the artist's creative intent and expression with his consent to new markets and fans instantly.

In this case, consent is the key. There is no reason we can't legitimate—we can't create a legitimate AI marketplace in the service of artists. There is a robust free market for music sampling, synchronization licensing, and deals with new entrants to the digital marketplace, social media companies, and all manner of new technologies. We can do the same thing with AI.

And my fourth and final point, to cultivate a lawful, responsible AI marketplace, Congress needs to establish rules that ensure creators are respected and protected.

One way to do that is to enact a Federal right of publicity. Deepfakes and/or unauthorized recordings or visuals of artists generated by AI can lead to consumer confusion, unfair competition against the artist that actually was the original creator, market dilution, and damage to the artist's reputation, potentially irreparably harming their career. An artist's voice is often the most valuable part of their livelihood and public persona, and to steal it, no matter the means, is wrong.

A Federal right of publicity would clarify and harmonize the protections currently provided at the State level. Visibility into AI training data is also needed. If the data on AI training is not transparent, then the potential for a healthy marketplace will be stymied as information on infringing content would be largely inaccessible to the individual creators.

And I might add, based on some of the comments I heard earlier, it would be hard to opt out if you don't know what has been opted in.

Finally, AI-generated content should be labeled as such. We are committed to protecting our artists and the authenticity of their creative works.

As you all know, consumers deserve to know exactly what they are getting. I look forward to the discussion this afternoon, and I thank you for the opportunity to present my point of view. Thank you.

[The prepared statement of Mr. Harleston appears as a submission for the record.]

Chair COONS. Thank you, Mr. Harleston. Thank you to all of the witnesses today. We will begin our first 5-minute round. And I am going to start by just exploring how we can respect existing copyrighted works, copyright protections, while continuing to safely develop and advance AI technologies. If we run out of time, we will

do a second round. My hunch is there is at least that much interest.

Mr. Brooks, if I might just start with you. Generative AI models like those your company creates are trained in no small part on vast quantities of copyrighted content, on data from copyrighted content.

Do copyright owners know if their works have been used to train Stability's models? Is Stability paying rights holders for that use? Why not, if not? And how would doing so impact your business and your business model?

Mr. BROOKS. Thank you, Senator. So, to the first question, models like Stable Diffusion are trained on open datasets or curated subsets of those datasets. So Stable Diffusion, for example, takes a 5 billion image dataset.

We filter that for content, bias, quality, and then we use a 2 billion image subset to train a model like Stable Diffusion. Because it is open, you can go to a website, you can type in the URL of an image, you can type in a name.

You can see if that work has appeared in the training dataset. And then we are working with partners to take those opt-out requests, and as I say, to incorporate them into our own training and development processes.

So, we do think open datasets are important. They are one part of how we are able to inspect AI for fairness and bias and safety. And so that is I think to the first part.

Chair COONS. So, if I heard you right, if an artist takes the initiative to search your training set, they might be able to identify that a copyrighted work was used and then submit an opt-out request. And you are in the process of facilitating that use. But to my second question, do you pay any of the rights holders?

Mr. BROOKS. As I say, Senator, we—this is 2 billion images, a large amount of content. A lot of it, you know, all kinds of content. Tech language models, for example, it is not just books, it is snippets of text from all over the internet.

As I say, to make that workable, we believe, you know, that it is important to have that diversity, to have that scale. That is how we make these models safe. It is how we make them effective.

And so—and so we collected, as I say, from online data. What I will say is that the datasets that we use, like that 5 billion image dataset I mentioned, they respect protocols like robots.txt. So, robots.txt is a digital standard that basically says, I want my website to be available for ancillary purposes, such as search engine indexing.

And so, the dataset that was compiled respected that robots.txt signal, and then on top of that, as I say, we have the opt-out facility that we have implemented.

Chair COONS. Thank you. Mr. Rao, it is my understanding that Adobe is taking a distinctly different approach. Your generative AI model, Firefly, was only trained on licensed data. Were there any downsides economically to that decision? Is your model less robust or has it had any impact on its performance? And how would you compare these two approaches in terms of the incorporation of opt-out and licensed?

Mr. RAO. Thank you for the question. So, as I mentioned in the opening remarks, that we—Firefly, our generative AI tool was trained on our stock photography collection, which are all licensed assets with the contributors, and that is actually the only dataset used in the version that you can use on Firefly on the web.

We think about the quality of this, and when we think about the quality to your question, we have to put a lot of image science behind that to make sure it was up to the level we require because we didn't have the most expensive version of that dataset. So, we had to put more computer science behind it to make it have the higher quality we needed.

As we go forward, we are looking at whether or not there are other areas where we need to supplement that dataset, and for those we referred to as opening licensed content or places where the copyright has expired.

Opening licensed to us means images that come from the rights holders who have licensed it without restriction. So, very similar to what we are talking about in the licensed content, this is a place—this is also what we call commercially safe.

Chair COONS. My sense, Mr. Brooks, is Stability is trying to honor something like 160 million opt-out requests in training your next model. Mr. Rao, Mr. Brooks, just this will be my last question, and then I will turn to Senator Tillis. Should Congress be working to ensure that creatives can opt out of having their works used to train AI models? How would you best do that, briefly?

Mr. RAO. So, we have this technology we refer to as content credentials in my opening remarks, and what that does, it is a metadata that goes along with any content. So, if you are in Photoshop right now, you can say, I want content credentials to be associated with this image.

As part of that, you can choose to say, I want it not to be trained on it—a “Do Not Train” tag that gets associated with the image and it goes wherever the content goes.

So, we do think the technology is there and available, and we are talking to other companies, including Stability, about this as an approach to honor that tag so people who are crawling it can see the tag and choose not to train on them.

Chair COONS. Should we require that?

Mr. RAO. And I do think that there is an opportunity for Congress to mandate the carrying of a tag like that, a credential like that, wherever the content goes. Right now, it is a voluntary decision to choose to do that.

Chair COONS. Should we require that?

Mr. BROOKS. There is some very interesting precedent internationally for this. The European Union has introduced certain kinds of text and data mining exceptions. And part of that is to say that you can use this for commercial, noncommercial purposes.

There is an opt-out requirement, but the opt-out has to be machine readable, as I say, as a matter of practicality, when you are dealing with trillions of words of content, for example, or billions of images, in this case.

The machine readability is important, and that is where these tags become an important part of how to implement it in practice.

Chair COONS. We will keep exploring this further. Senator Tillis.

Senator TILLIS. Thank you, Chairman. I'll have Senator Blackburn go and then I will follow Senator Hirono.

Senator BLACKBURN. Excellent. Thank you, Senator Tillis. And Mr. Chairman, thank you for the hearing today. It is so appropriate that we have this. I am from Tennessee. We have thousands of artists and songwriters and musicians, and we have actors and actresses, and we have authors and publishers.

And everywhere I go, people are talking about the impact of AI, to the positive or the negative. You know, you look at health care, you look at logistics, you look at autos, you look at entertainment, and there are pros and cons.

But the one point of agreement is, we've got to do something about this so that it is going to be fair, and it is going to be level. Mr. Harleston, I want to come to you right off the bat because you mentioned the NIL issue, which I think is an imperative for artists to be able to own that.

And you also mentioned the right of publicity laws, and of course, those are State level laws. And as you rightly said, we don't have a federally preemptive right to publicity law. And I think the dust up—a lot of people came to realize this over Drake and The Weeknd, and "heart on my sleeve."

And this is something that does have to be addressed. So, for the record, give us about 30 seconds and then you guys, I see your capable team behind you, you can submit something longer in writing, if you would like, on the reason State level publicity laws are not enough.

Mr. HARLESTON. In 30 seconds—

[Laughter.]

Mr. HARLESTON. State level publicity laws are inconsistent from State to State. A Federal right of publicity that really elevates right of publicity to an intellectual property is critically important to protect—

Senator BLACKBURN. Okay, I am going to help you out with this—

Mr. HARLESTON. Okay.

Senator BLACKBURN. A federally preemptive right to publicity law would provide more of that constitutional guarantee to her works that Ms. Ortiz has mentioned.

Mr. HARLESTON. Absolutely.

Senator BLACKBURN. All right. And—

Mr. HARLESTON. And we will follow up with you, Senator.

Senator BLACKBURN. Yes, excellent. I think something in writing would be very helpful there. Now, I think it was very appropriate that you had Spotify and Apple Music take down "heart on my sleeve." Important to do. And talk about the role that the streaming platforms should play. Should they be the arbiter when it comes to dealing with this generative AI content?

Mr. HARLESTON. These streaming platforms, we acknowledge that they are in a challenging position, but certainly in some instances when there is clear, or it is clear that the content that has been submitted to them for distribution—

Senator BLACKBURN. So, a knowing and willingness standard would be nice.

Mr. HARLESTON. That would be very nice, yes.

Senator BLACKBURN. Okay. I am helping you out there.

Mr. HARLESTON. You are doing great.

Senator BLACKBURN. Thanks for being here.

Okay, Professor Sag, want to come to you. This spring, the Supreme Court issued a—what I thought was a very appropriate decision in *Warhol v. Goldsmith*, and I was very pleased to see them come down on the side of the artist. I filed an amicus brief in this case arguing for strong, fair use protections for creators. Now, we have been through this thing in the music industry where “fair use” became a fairly useful way to steal my property.

And artists don’t want to go through that again. Right, Ms. Ortiz?

[Voice off microphone.]

Senator BLACKBURN. It didn’t work the way it was supposed to. And I would like for you to talk for a moment, should AI, unlicensed AI ingestion of copyrighted works might be considered fair use when the output of AI replaces or competes with the human-generated work.

Now, Ms. Ortiz has laid this out fairly well in her comments and the Supreme Court has sided with the artist in *Warhol v. Goldsmith*. But this fair use standard comes into play every time we talk about our fabulous creative community and keeping them compensated. So, the floor is yours.

Professor SAG. Senator Blackburn, commercial replacement should not be the test. The test should be exactly what the Supreme Court said in the *Andy Warhol* case.

The question is, is this significantly transformative? What that means in relation to training AI models is, does the output of the model bear too much resemblance to the inputs? And that is a different question to, is it competing with the inputs? Could it be used as a commercial substitute?

If you look at some of the old cases on reverse engineering software, companies were allowed to crack open software, find the secret keys to interoperability, and build new competing products that did not contain any copyrightable expression, and the Court said that that was fair use.

So, I think on current law, the answer is no. Potential substitution in terms of a competing product is not the test. The test is, are you taking an inappropriate amount of an artist’s original expression.

Senator BLACKBURN. Well, my time has expired. Thank you for that. We just don’t want it to become a fairly useful way to steal an artist’s product. Thank you, Mr. Chairman.

Chair COONS. Thank you, Senator Blackburn. And thank you for the passionate engagement you have always brought to these issues on behalf of the creative community.

[Laughter.]

Chair COONS. Senator Hirono.

Senator HIRONO. Thank you, Mr. Chairman. Mr. Harleston, whenever the idea of negotiating licenses is raised, people express concerns about how complex it would be and how AI platform developers could never possibly negotiate with all rights holders. But in the music context, at least, you have a lot of experience negotiating rights.

Could you tell us a little bit about your industry's history of negotiating rights with digital music services and lessons that history could teach us, for whether rights negotiations would be possible with AI platforms?

Mr. HARLESTON. Thank you, Senator. As you referenced, we have had a long history with the transition of our business from a physical business to a digital business, and having to encounter digital platforms that were very quickly adapted by consumers and had lots of our content on there.

What we found was ingenuity does play a role. It is not easy. But we were able to identify or find ways to identify our copyrights, to work out licensing schemes that allowed the platforms to be able to carry and distribute the music.

And in a commercial environment that was positive for them, while at the same time allowing the artists to be properly compensated. And this is, you know, with the—in the music side, we have two sets of rights, which makes it even more complicated, but we have done great work over the years to develop systems that allow identifying not only the sound recording, but also the underlying composition.

So, it could be done. But what it needs, it needs—what we would need is we need help to make sure that everyone understands that there are rights that are affected and that the activity that is happening now is violative. And once they understand that what they are doing is violative, that brings them to the table so we can negotiate a deal.

Senator HIRONO. I note that in your testimony you said that consent is the key. So is your position that every artist's work before it can be used to train AI models, that the company that is wanting to use that information has got to get the consent of the originator?

Mr. HARLESTON. In a very short answer, yes.

Senator HIRONO. And you think that we are able to do this knowing that these platforms incorporate billions and billions of information to train their AI models?

Mr. HARLESTON. Understanding that, but it absolutely could be done as these—as the digital platforms that exist today, the licensed platforms ingest millions and millions of songs every week. So, it is not a problem in that respect. There is metadata that we could license. We could absolutely do that. But there has to be an initiative on the side of the companies to reach out.

Senator HIRONO. So, Ms. Ortiz, if I—Mr. Brooks, rather, sorry. My—what I heard you say in response to the Chairman's question is that for all of the data that you input into your model, you do not get the consent of the artist or originator. Is that correct, Mr. Brooks?

Mr. BROOKS. So, we, Senator, we believe that yes, if that image data is on the internet and robots.txt says it can be subject to aggregated data collection, and if it is not subject to an opt-out request in our upcoming models, then certainly we will use those images, potentially use those images if it passes our filters.

Senator HIRONO. So basically, you don't pay for the data that you put into your—to train your model.

Mr. BROOKS. For the base, the kind of initial training or teaching of these models with those billions of images, there is no arrangement in place.

Senator HIRONO. So, you have Ms. Ortiz who says that that is wrong. Is that correct, Ms. Ortiz?

Ms. ORTIZ. One hundred percent, Senator.

Senator HIRONO. So do you know if—well, I think you mentioned that your work has been used to train AI models and you have gotten not one cent for that use.

Ms. ORTIZ. I have never been asked. I have never been credited. I have never been compensated one penny. And that is for the use of almost the entirety of my work, both personal and commercial, Senator.

Senator HIRONO. So, if you were to allow your works to be used to train, you would—do you think that you would negotiate—if there was a law that required compensation, then that compensation negotiation should be left to you and the entity such as Mr. Brooks'.

Ms. ORTIZ. Personally, I love what I do, so I wouldn't outsource it to an AI, but that is not a choice for me to make, and it is all about that. It is about being able to have that choice and artists don't have that right now.

Senator HIRONO. Thank you.

Senator TILLIS. Thank you, Mr. Chair. I was actually inspired by one of the opening statements, so I went out and generated a cat driving a 1960 Corvette with a surfboard in it. And I produced that picture.

[Phone is held up with the screen facing the witnesses.]

Senator TILLIS. Actually, it gave me four options. This one I found the most interesting. But it raised a question that I wanted to ask you, Mr. Brooks. If an artist looked at that and said, that is in part developed by that 60s Corvette in South Beach, how does that artist then go about saying—I am trying to get an understanding of your current opt-out policy.

And one of the issues that we have had here and not completely related, but we have a notice, a takedown notice and stay down discussion in the past around creative works. So, I was just trying to understand, and I think it is going to be a lengthy answer. And then if I talk to a creative, it is going to be a lengthy answer.

But for the record, it would be very helpful to me for your specific platform to understand how that opt-out process works. I think I heard right that you could embed within the works certain things that already create an opt-out, or that that work shouldn't be used. But I want to drill down. We don't have time to do that now.

And in a twist of irony, I was wondering if any of the witnesses would suggest any creative works by other governmental bodies that we should steal and use as a baseline. In other words, what good policy seems to be being discussed or passed? What particularly problematic at either end of the spectrum? Because I am sympathetic to the issues at both ends of the spectrum on this argument.

So maybe we start with you, Professor. Are you aware of any Western democracy states, I am not particularly interested in what

China is doing because whatever they agree to, they are going to rip off anyway, but any best practices that we should look out there, or bad practices, or trends that we should avoid or be concerned with as we move forward?

Professor SAG. I think that the European Union's approach, where they have different rules for commercial and noncommercial use, and opt-outs have to be respected for commercial uses of the text mining in Article 4 of the DSM has something to recommend it.

By the same token, I would note that opt-outs do not apply to researchers working at proper research institutions in the EU, nor do contractual overrides, which is a position that I can't see Congress adopting, but it is certainly something to look at. That's—that's really it.

Senator TILLIS. Anyone else briefly could add what—Ms. Ortiz, I should also add, I have seen all your works and it has been since 11 o'clock last night that I was talking about "Guardians of the Galaxy" with my colleagues as we were coming back from Vilnius.

Ms. ORTIZ. It was a really fun project to work on, Senator, so thank you. So, what the artists community have suggested is that models be built starting from scratch via public domain-only works, that's work that belongs to everyone. Any expansion upon that to be done via licensing. And there is a couple of reasons for this.

Current opt-out measures are inefficient. For starters, machine learning models, once they are trained on data, they cannot forget. And machine unlearning procedures are just dead on the water right now, and this is not according to me. I am an artist. I have no idea on this. This is according to machine learning experts in the field.

Second, things, safety filters like, for example, prompt, you know, filters are so easily bypassed by users. So unfortunately, when companies say, hey, opt out, there is no real way to do that. But even further, what happens if someone doesn't know how to write a robot.txt?

Like, how does a person who may not know the language, may not know the internet, may not even know that their work is in there, recognize that, you know, they need to opt out. This is why my community in particular has suggested over and over, opt in should be the key in order to base the foundations of consent, credit, and compensation.

Senator TILLIS. And Mr. Brooks, I can understand the challenges with opt-in versus opt-out in terms of the task that you would have ahead of you. But what is your view of the concerns that creatives have expressed in this light and the current opt-out process that you all have in place or procedures which I would like to get information for, for the record.

Mr. BROOKS. Thank you, Ranking Member. Look, I will say at the start that we do need to think through what the future of the digital economy looks like. What do incentives look like? How do we make these technologies a win-win for everyone involved? These are very early days from our perspective. We don't have all the answers, but we are working to think through what that looks like—

Senator TILLIS. I am going to stick around for a second round, so we will get a little bit deeper into that. But I want to defer to my colleague from California.

Chair COONS. Thank you, Senator Tillis. Senator Padilla.

Senator PADILLA. Thank you, Mr. Chair. And I want to thank the witnesses for your testimony and participation today. Speaking of California, I can't help but observe that California is very well represented on this panel. Not only a point of pride for me as a Senator from California, but it is frankly not a surprise since we are the creative and tech hub of the Nation.

Now, generative AI tools, as we have been talking about, present remarkable opportunities and challenges for the creative community and our broader society. And I couldn't help but observe that in reviewing the testimony from each of you, I noted the common goal of seeking to leverage and develop AI tools to complement and encourage human creativity and artistry, while also respecting the rights and dignity of the original creators.

So, it is a tall order, a delicate balancing act in many ways, but that is—this seems to be the shared objective here. So, I want to thank you again for participating in this hearing as we are working to determine what role we play in fostering the development of AI in a manner that is a net positive for innovation and creativity.

My first question, and I will keep it brief because it is sort of piggybacking on—Senator Hirono has raised it, Senator Tillis was just trying to expand upon it, and is directed at Mr. Brooks. This whole opt in, opt out: We can talk about what the process is, whether it is easy, clear, or not for artists.

And, you know, I don't completely agree with you that we are in an early stage because it is happening fast. Tell me how it is possible—explain how it works to have a system unlearn inputs that have already been taken, if you get this after-the-fact opt-out from an artist. It is happening now. While you are trying to think what it means long term, it is happening now.

So how does it work, not just process, checking a box, filling out a form, but technically?

Mr. BROOKS. Thank you, Senator. So, just in terms of the data collection piece, I just want to make it clear that today it is very much a kind of work in process framework. You know, you can go to this website. You can indicate you want to opt out. We will take those opt-out requests as they come in.

But as we were talking about before, it is important that eventually there is a standardized kind of metadata that just attaches to these works as they go out into the wild. And as I said, that is what the EU is requiring, and I think there will be a lot of standards development in that space, again, with—in terms like Adobe and others.

In terms of what then happens, you know, as I say, we filter that training data for a few reasons. We take out unsafe content, we adjust for issues like bias to correct the bias. And then in addition to that, we start to incorporate, as I say, the opt-out requests.

Sometimes some of the models we release are retrained from scratch with new datasets. Again, they take into account the lessons learned through previous development, both as an organiza-

tion, as a company, and potentially technical things that we have learned as well in that process.

Some of the models that are released are just fine-tuned variations of the model, and so those ones may have the same kind of basic knowledge from that original training process, and there has just been some additional training to correct for certain behaviors or improve performance in specific tasks.

So, in terms of, you know, the future of this space, you know, there is a lot of work being done on unlearning in general. You know, how do you interpret the relationship between training and the data in training and the performance of the model?

How do you potentially adjust to that, different ways? But as I say, at this stage, we treat it as a process of incorporating those opt-out requests, retraining, and then releasing a new model trained on that new dataset.

Senator PADILLA. I hear you, and I just want a level set a little bit, not just out of the concern for the artists, but knowing that unless you are getting one, two, three inputs today, which may be small enough to keep your arms around, I doubt that is the case as we are getting into the hundreds and thousands of inputs per day to go in and relearn, unlearn, and comply with any consent or opt-out. It gets overwhelming and unfeasible real quick, and it is happening now.

I also wanted to follow up on a subject matter that Senator Coons touched on earlier. We know that generative AI models need to be fed large datasets to learn how to generate images based on user prompts, just like Senator Tillis did. By the way, that looked much more like Pacific Coast Highway than South Beach.

Now, AI for—this is now talking to folks back home, can only understand what it is taught, making it critical that for AI companies to train their models with data that captures the full range of the human experience, want to be inclusive and diverse, if we are going to be accurate in representing our users, representing the diverse backgrounds of all users.

Now, Mr. Rao, you have explained how Adobe's Firefly seeks to avoid copyright infringement by being trained on only licensed Adobe Stock images, openly licensed content, and public domain content.

So how do you reconcile both? You want to be as inclusive as possible, which means as much data input as possible, but to avoid the copyright infringement, you are being selective in those inputs. That diversity of input is important, I think, for the diversity of output. So how do you reconcile?

Mr. RAO. It is definitely a tension in the system. Right? The more data you have, the less bias you will see. So, it is great to have more data.

But when you set the expectations that we had for ourselves of trying to design a model that was going to be commercially safe, we took on the challenge of saying, can we also do that and minimize harmful bias? And the way we did that, we have an AI ethics team. We started that 4 years ago.

And one of the key things they did when we were developing Adobe Firefly was not only do we have the dataset and we understand what that is, we also did a lot of testing on it. We have a

series of prompts, hundreds and hundreds and hundreds of prompts.

We were testing against it to see what the distribution of model is. Is there going to be a bias. If you type in “lawyer,” are you only going to get men—or white men, and what does that mean, and how, then, do you change that?

And you either change it by adding more data, making it more diverse—and so that means you have to get ethically-sourced, more data to diversify the dataset, or you can add filters on top of the dataset to force a distribution of what you expect to see if you are typing in certain search terms and make sure the bias is removed.

So, you can either do it by adding more data or you can do it with through adding filters on top of the model itself to ensure that you are going to get the right result.

Senator PADILLA. And if you ask—if you input “Senator,” what comes out?

Mr. RAO. An amazingly handsome man and woman, just very intellectual.

[Laughter.]

Senator PADILLA. Men and women—colors across the spectrum.

Mr. RAO. Across the spectrum.

Senator PADILLA. Thank you, Mr. Chair.

Mr. RAO. The first time we did “lawyer,” though, we only had white men. And as general counsel, I was like, there should be some people who look like me as well.

Chair COONS. Thank you, Senator Padilla. Senator Klobuchar.

Senator KLOBUCHAR. Okay, very good. Thank you. I was glad to be here for all your testimony and thank you for that. I guess, I will start with you, Mr. Harleston. Approximately, and I know you talked about this a bit with some of the other Senators, Senator Blackburn. Approximately half the States have laws that give individuals control over the use of their name, image, and voice.

But in the other half of the country, someone who was harmed by a fake recording purporting to be them has little recourse. In your testimony, you talk about new laws and how they could protect musicians’ names, likenesses, and voices—the right of publicity, I think you called it.

Can you talk about why creating this is important in the face of emerging AI? And how have statutes in States that have these protections helped artists?

Mr. HARLESTON. Thank you, Senator, for the question. It is critical in this environment when we are talking about the creative expression that the artist has made, that the right of publicity also be extended at the Federal level.

There is inconsistency, but more importantly, the preemptive element of it is critical. Raising it to the level of an intellectual property is also critical. What we have seen, and this is really in the area of deepfakes, where you have seen, I think, Ms. Ortiz referenced how many times her name was listed.

We are finding with our artists, particularly the ones that are most established, that their names are, you know, daily—hundreds and hundreds of thousands of posts with their names. And also, there is sometimes images that are used as well.

Senator KLOBUCHAR. Mm-hmm.

Mr. HARLESTON. So, it is critical to have this right to protect the artists and their use. And if I could just say one thing on the—I know this is not your question, but I have to say—

Senator KLOBUCHAR. There we go.

Mr. HARLESTON. Because it is killing me—

Senator KLOBUCHAR. I will just add it to my time.

[Laughter.]

Mr. HARLESTON. All right, thank you, thank you. On the opt in, opt out, there is an element beyond commerciality. And I want to make sure everyone understands. Ms. Ortiz did reference it, about she didn't really—she probably wouldn't want a license to AI. And there are—we have artists that don't want a license to streaming services. So, they are not—it is not always about the commerciality.

Some artists just don't want their art distributed in certain ways. And the Beatles didn't come onto streaming platforms till about 7 or 8 years ago. That was a decision that was very important to them. So, I want to add that into the conversation. I know that wasn't your question, sorry.

Senator KLOBUCHAR. Okay, very good. And so, what do you see as the obligations of social media platforms on this?

Mr. HARLESTON. With respect to AI?

Senator KLOBUCHAR. Uh-huh.

Mr. HARLESTON. Oh, great—fantastic question. We believe that the social media platforms absolutely have an obligation. I will say this, that we could help them by giving them a hook beyond copyright in terms of being able to take down some of the—

Senator KLOBUCHAR. Exactly—

Mr. HARLESTON. Some of the deepfakes. They have challenges with some of the platforms on this.

Senator KLOBUCHAR. Yes. Right, exactly. And I think we are seeing the same thing. I guess I would turn to you, Mr. Brooks. You talked about advocating for creating ways to help people identify AI-created content.

And when we talk about deepfakes, we are already seeing this with political ads, and not even paid ads, just videos that are put out there. There's one of my colleague Senator Warren that was just a total lie that, saying that—acting like it was her that she was saying people from one party shouldn't be able to vote.

And we have seen it in the Republican Presidential primary. A number of us on a bipartisan basis are working on this. I chair the Rules Committee, so it is kind of my other hat.

Do you agree that without tools for people to determine whether an image or video generated by AI, that that would pose a risk to our free and fair elections, if you can't tell if the candidate you are seeing is the candidate or not?

Mr. BROOKS. Thank you, Senator. We absolutely believe that these transparency initiatives like CAI with Adobe are a really important part of how we make the information ecosystem more robust.

This isn't just an AI problem or a social media problem. It is going to require everyone, and it is going to require accountability right across that ecosystem. But what we think is, you know, we have in place things like metadata, things like watermarking for content.

They are just some more of the signals that social media platforms can use to decide whether they are going to amplify certain content.

Senator KLOBUCHAR. Yes, and we have got this REAL Political Advertisement Act, with Senator Booker and Senator Bennet. There is a version initially that was also introduced in the House. And so, that is one solution.

But we are also going to have to look at, I would say, banning some of this content, because even a label or a watermark—it is not going to help the artist or the candidate if everyone thinks it is them and it is not, and then at the end, it says generated by AI.

Mr. BROOKS. It is a great question and a really important one, I think, Senator, because there are a few things in there. I think there's the question of the use of likeness, particularly for improper purposes, where you are implying that there is some kind of endorsement or affiliation between a particular person and a particular work or idea.

That is different, I think, to the use of the kind of free experimentation with style and some of these other issues that tend to get lumped together in AI outputs.

Senator KLOBUCHAR. Mm-hmm.

Mr. BROOKS. And so, in terms of these scenarios that you are talking about, there is this kind of improper use.

You are implying that someone endorses or embraces a cause or a work that they are not affiliated with. And there needs to be clear rules around how like this is used in that context, whether through right of publicity or through some of the bespoke deepfake legislation.

Senator KLOBUCHAR. Okay. Last, Mr. Rao, our recent study—and I know you have worked on this democracy issue, which I truly appreciate. A recent study by Northwestern predicted that one third of the U.S. newspapers that existed roughly 2 decades ago will be gone by 2025.

The bill that Senator Kennedy and I have, the Journalism Competition and Preservation Act, would allow local news organizations to negotiate with online platforms, including generative AI platforms.

This bill passed through this Committee now twice. Could you describe how Adobe approaches this issue? And in your experience, is it possible to train sophisticated, generative AI models without using copyrighted materials, absent consent?

Mr. RAO. Thanks for the question. Absolutely. We—our current model that is out there is trained using the licensed content that I had mentioned before and other content that has no restrictions on it, and it comes from the rights holders directly.

So, we definitely think it is possible. We have done it. It is out there on our website, and it is also in Photoshop, and people love it. The creative professionals are using that AI. It makes their day easy.

It lets them start their creative work in just one click and then they finish it in the tool. So, it has really revolutionized how we think about things. In terms of how we acquire datasets, and we have a group inside Adobe whose—that is their job. Their job is to think about where do we need to go next? Do we need to get to dif-

ferent media types? Do we—are we missing some sort of subject matter for our AI to be more accurate?

That was a question we had before. We think about that content. Maybe there is a newspaper that you mentioned that has the kind of content we need. We go approach that organization and say, look, we need to license that content in to make sure our AI is more accurate.

So, we have a team that thinks about this—sources it, and brings it in.

Senator KLOBUCHAR. And in the absence of that? I mean, what impacts do you anticipate this could have on local journalism if there are no rules of the road put in place?

Mr. RAO. Yes, I think that, you know, both on the authenticity side and on this side, if people are able to, you know, create images and these newspapers are not able to get, you know, the ability to license the work they are doing, it could certainly have a negative impact to them.

On the authenticity side, the reason why so many media companies have joined the Content Authenticity Initiative, like AP, Reuters, Wall Street Journal, New York Times, Washington Post, is because they know that when they are showing images, they need to be able to show that they are actually true.

Senator KLOBUCHAR. Mm-hmm.

Mr. RAO. They need to be able to prove that it happened. If people stop believing that any of these digital images they are seeing are real, then they are going to stop consuming newspapers. They are going to stop consuming that content because they are not going to believe it. So, you have to give those local newspapers a way to prove what they are showing is true—

Senator KLOBUCHAR. Right. Absolutely.

Mr. RAO [continuing]. So people can still consume it.

Senator KLOBUCHAR. Of course, there is a lot less famous newspapers, including some very small ones in my State that just you might not mention. Right?

And so, I think that part of it, is that, you know, the Ms. Ortizes of this story need to be able to have some kind of power to be able to protect their content, too, because they don't have a general counsel, and they are not going to be able to, on their own, start some major lawsuit. And so, I think that is how we have to think about that, too, as we look at all of this.

Mr. RAO. And that is why I would say again that when we designed the Firefly, we designed it that way. Right?

Senator KLOBUCHAR. Mm-hmm.

Mr. RAO [continuing]. To be commercially safe first, right, making sure that we built the model the way—

Senator KLOBUCHAR. Yes, no, I am not—I am saying it sort of rhetorically to the world and to everyone that needs to get this done, as opposed to you, Mr. Rao.

Mr. RAO. Thank you so much.

Senator KLOBUCHAR. All right. Appreciate it. And I thank you, both of you, for your continual bipartisan work in taking on this very important issue. Thanks.

Chair COONS. Thank you, Senator Klobuchar. We are going to do a last round of questioning. We may be joined by other colleagues,

but we are also in the middle of a vote. So, my hunch is we will resolve this in 10 to 15 minutes at the most, if I might.

I am interested in pursuing the question of a Federal statutory right of publicity. And to me, the core issue really is, what is the remedy?

Often, preemption is motivated by a desire for there to be consistency, the elevation in terms of process and access to justice, and potential remedies that comes with a Federal right as opposed to a State right.

But, Professor, if I could start with you. You testified earlier in response to a question from Senator Blackburn that commercial replacement is not the appropriate test under current fair use law in the United States.

Should we adopt a Federal right of publicity with commercial replacement as the test or part of the test, and how would that play out? What other remedy might you suggest under a new Federal right of publicity?

Professor SAG. Senator, thank you for that question, because I was quite alarmed by some of the discourse here about the right of publicity. I think, as well as—

Chair COONS. Regulated by discourse.

Professor SAG [continuing]. As well as thinking about publicity rights for well-known artists, musicians, etcetera, Congress should be thinking about the right of publicity of ordinary people, people who are anonymous, people who have no commercially valuable reputation.

All of us deserve to be protected from deepfakes and synthetic reproductions of our name, image, and likeness, regardless of whether we are a famous politician or a famous artist or just an anonymous law professor. So, I think—

Chair COONS. How would you focus the remedy in order to make that effective?

Professor SAG. Senator, in terms of remedy, I think that right of publicity statutes have traditionally had injunctive relief, usually incorporating equitable balancing tests. That is the remedy I would go for, which would mean, the models might have to be retrained.

Chair COONS. Injunctive relief only, not commercial?

Professor SAG. Damages, potential as well. But statutory damages, I don't think so. Statutory damages can be quite distorting. They tend to be a honeypot for opportunistic lawyers, as well as genuinely aggrieved plaintiffs. So, I would steer clear of statutory damages, but actual damage and injunctions, absolutely.

Chair COONS. Mr. Rao, I'd be interested in your views on what a right of publicity might potentially do. I'd also be interested in hearing your thoughts on how we should be trying to balance respecting copyright through this or other means, while incentivizing investment in AI and accelerating innovation in the United States?

Mr. RAO. Thank you for the question. So, we talked about in our testimony similar to, but not exactly like, a right of publicity.

We referred to as a Federal anti-impersonation right. And the reason we thought about it from an anti-impersonation perspective is actually some of the same questions Professor Sag raised, which is we want to make sure Professor Sag himself is not—does not have a deepfake made of him.

So, if you think about it as an impersonation right, that would apply to everybody. And what we are really targeting there is we see the economic displacement that we have been talking about here, where an AI is trained on an artist and creates an output that is exactly like the artist, and they are getting displaced by that work.

And copyright may not reach them, like that has been the question. So that is why we believe they do need this right so they can go after these people who are impersonating work, whether that is likeness, whether that is style.

And then the test would be something that we would work out through 6 months of deliberation here in this body, exactly how you would decide that. But I think that is the right approach, because you want to focus on people who are intentionally impersonating someone in order to make or get some commercial benefit, and I think that will help clarify what harm we are trying to address.

Chair COONS. Mr. Brooks, how do you think a Federal anti-impersonation right—

Mr. RAO. By the way, that spells FAIR. Just want to make that clear. I know how Congress loves acronyms. Yes?

Chair COONS. We are enthusiastic about acronyms. We actually are producing a Senate-only version of ChatGPT that only produces acronyms for bill names.

[Laughter.]

Chair COONS. Mr. Brooks, how do you think a Federal publicity right or an anti-impersonation right, a Federal requirement that there be opt-in only rather than opt-out would impact the business model that you are currently representing?

Mr. BROOKS. So, Chair Coons, I think the actual instrument and the content of that instrument, I think is really diagnostic at this stage. As I said to Senator Klobuchar, it is important from our perspective that there are clear rules governing the use of likeness in an improper way.

I think the important thing to stress there is that it is a use. And to some extent we can't escape the fact that the determination of whether it is proper or improper will depend on the application, what the user does or does not do with that content downstream.

And so, as I say, you know, from our perspective, the lines in the sand between improper use of likeness, free experimentation with style, or other kinds of good or bad use of these tools aren't easy to draw.

They are very fact sensitive. It may be appropriate for courts to determine that. But at a high level, as I say, I think there is a core of things around that improper use of likeness, especially voice likeness, that there may be some legislative intervention there that makes sense and may have obligations, as I say, across the supply chain, across the ecosystem.

Chair COONS. Mr. Harleston, if I might, the Copyright Office recently issued guidance about human authorship being critical to any copyright protection.

Is their guidance accessible enough, relevant? Did they strike the right balance? Should we be looking at a different policy in terms of how broadly copyright protection should reach when there is AI assisted creativity as opposed to AI generated?

Mr. HARLESTON. I think the Copyright Office did a pretty good job. One can debate whether an AI component in a broader work should also be afforded some form of copyright. You know, I think they landed in the right place, that it shouldn't. That copyright should only be afforded to human creation.

So, for example, if you had an AI-generated song—well, if you had a song that was created by an artist and they used—a piece of it was generative AI, there should be a copyright in that entire work, but the AI-generated portion would not be protectable.

So, if someone were to actually sample it, which would lift it out and use it in another context, it would not be subject to copyright. I think they did a pretty good job trying to strike that balance.

Chair COONS. In the conversation I had previously with Professor Sag, how do you feel about the scope of potential remedies if we were to craft an anti-impersonation statute?

Mr. HARLESTON. I am glad you asked me that question.

Chair COONS. I thought you'd be.

[Laughter.]

Mr. HARLESTON. Thank you. I think there should be a private right of action. I think that it is—I think commerciality is, again, not always the proper standard here.

I think that in some instances we have had artists who have had been victim of deepfakes where the voice was appropriated, and the lyric content was something the artist would never have said.

And that is something that can have irreparable harm to their career, you know, in trying to explain that it wasn't them, because there is stuff that is really good, these—sort of these AI-generated things are really good.

Chair COONS. Ms. Ortiz, last but not least, has, in producing some of the interesting, engaging, powerful, inspiring content you have generated, have you ever relied on an AI tool to help you expand or produce some of the works you have worked on? And what is your hope about what we might do going forward here in Congress in response to what we have heard from you about your concerns?

Ms. ORTIZ. I am very happy you asked this question, Senator. So, I have never really—I was curious very early on before I knew the extent of the exploitation of artists. Very briefly used an AI to generate references, and I didn't enjoy it at all. I am—you know, I love every step of the process of being an artist.

And ever since I found out, you know, last August, September, of what actually went behind the scenes, I just—I cannot use it. My peers refuse to use it. My industry is very clear that we do not want to exploit each other. And again, it is important to remember that these, you know, models basically compete in our own market.

And this isn't something that is hypothetical. It has happened now with our own works. And one of the things that I would hope, you know, would be kind of addressed here is that a lot of the solutions that have been proposed—or, you know, basically you cannot enact them unless you know what is in the dataset.

And for this, we need to ensure that there is clear transparency built from ground up. Like, no offense to some of the companies here, but if you don't know what exactly is in the dataset, how do we know?

How does the licensor know that my work is in the dataset? And that I feel like it is one of the starting foundations for artists and other individuals as well to be able to gain consent, credit, and compensation.

Chair COONS. Thank you. Mr. Tillis, before I just hand it over to you, and then we are going to conclude, I just appreciate all of you taking the time and effort helping educate us. You are literally training us as we try to produce some fidelity in our legislative work. Senator Tillis.

Senator TILLIS. Yes, to me, trying to figure out what may or may not be in the language model is a lot like taking roll in a dark classroom. I just don't understand how you would do it. So, you know, I can see that we have to work at it.

But I want to start, Mr. Brooks, by thanking you for being here. I think that anyone that is watching this needs to understand that this isn't unique to Stability AI. This is a broader set of issues that we have to deal with, and I appreciate the fact that you'd be willing to come here because you should expect that some of the concerns are—that were going to be expressed to begin with.

I have one question. The bad news for you all is that my staff are really excited about this. These are the questions [papers are shown to the witnesses]—

[Laughter.]

Senator TILLIS [continuing]. That we are going to submit for the record. But rather than expect you all to respond to every one of them, you are welcome to do that, your area of expertise, your priorities, just use that to guide you, and get that information back for the record.

But one of the ones I won't have to ask because I will ask it now is, a recent survey on how consumers view AI found that most consumers, nearly 80 percent, believe the use of AI should be explicitly disclosed.

Now, in Vilnius, I happened to stay at a hotel that is called the Shakespeare Hotel, and every room was named after the greats. I don't see a day 100 hundred years from now where those rooms are going to be named after great LLMs. And the reason for that is I think there is a natural cultural bias for rewarding the human beings who are truly the creators and the lifeblood of our creative community.

So, does anyone here disagree that a work that is derived even from, let's say, licensed content, that the consumer should know that this was created by a machine versus an original creative work by human beings?

Anybody disagree with that, or maybe technical issues I should look at? No different than Mr. Rao, me, I use Photoshop. I could create Corvette cat with a skateboard or surfboard really quickly. No different than I want that, which, as you know, again, based on prior creative work, somehow have disclosures. Does that make sense to you?

Mr. RAO. Yes, I think the question, and we thought about, we think it is definitely of interest to our creative customers, is to be able to show something human created versus AI created. In Adobe Firefly, it all comes out saying something is AI created.

That is on by default. So, you will always know that it is AI created. The trick going forward, though, is we anticipate our AI features are our most popular features in Photoshop, so we expect going forward, most images are going to have a part that is AI and a part that is human, and you sort of have to start thinking about what are you disclosing when you disclose that. Right?

The content credential we mentioned before that you could use a “Do Not Train” tag on it, or you could use it for, you know, addressing deepfakes. Also, we will record the human part versus the AI part. So, you could think about using that as a disclosure.

But I am not sure over time people are going to be as interested in knowing the identity of the artist who created the work as opposed to which part of it they did with AI, and which—

Senator TILLIS. That is fair. Professor Sag, do you have a comment?

Professor SAG. Just to follow up on that. You also have to think that you are not just talking visual works here. Like take the same thing with written works. Someone uses GPT to help smooth over their writing, refine something, explain it more clearly.

There are some awkward line drawing questions, but the spirit of the disclosure requirement is correct. The implementation, I think, just will be difficult.

Senator TILLIS. I agree. And Mr. Chair, I am checking the votes. I think it is probably time for us to wrap up the Committee. I think you could see from this just by the sheer number of Members who came to the Subcommittee, this is an area of interest and a priority for us.

Mr. Chair, I have decided that maybe for the next hearing, it is going to take a little bit more tuning for me to get the answer, but I am going to do a—with the—you know the song, “Who Let the Dogs Out”?

[Laughter.]

Senator TILLIS. I was thinking we would set that to “Don’t steal my IP,” and I will see if I can get that done. If you think about it, it’d be pretty snappy.

[Laughter.]

Senator TILLIS. But I will work on that for those of you—you may have to get a bigger room if people know about that in advance.

Chair COONS. We may end up doing this as a duet.

[Laughter.]

Senator TILLIS. But again, I think this Committee has demonstrated that we are very thoughtful, and we are very diligent, and I, for one, could sit at that table and probably present the interests of either side of the spectrum, which is why I believe that we need legislative certainty.

We need to learn like data privacy, data ownership. In Europe, they don’t always get it right in the first tranche, so we wouldn’t necessarily lift something up and implement it here, but we want to think it through and make sure it is something that scales properly. But this is clearly an area where I don’t think anyone—they would be hard pressed to convince me that no action is required.

And again, my bias on this Committee from the beginning, having grown up in innovation, technological innovation, seeing the

compelling numbers about how important it is to our economy and our culture, there is a lot of work to do.

And I am confident with the leadership of the Chair, we are going to get work done. We look forward to your continued engagement. Thank you.

Chair COONS. Thank you, Senator Tillis. I think it was Mr. Brooks, I may be wrong, who early on said that other technological developments, perhaps it was you Professor, word processing didn't end authorship, smartphones didn't end photography, but they impacted them. They impacted them.

And we need to closely, and with some deliberation, realign what Federal rights and protections there are, both to deal with things like deepfakes—some argue that Shakespeare himself was a deepfake—to protect the rights of individuals, protect the rights of those who earn their living by being creative, to ensure that consumers understand what they are consuming, and to make sure that we are aligning with other countries that share our core values and our priority on a free market and the rights of individuals in contrast to other countries with other systems.

So, I am grateful to all of you for testifying today, for taking your time and contributing to this. These are very challenging questions. Members can submit questions for the record for these witnesses if they were not able to attend. Questions for the record are due by 5 p.m., one week from today, July 19th.

Again, thank you, all. I look forward to your input as we try and craft a good legislative solution. With that, this hearing is adjourned.

[Whereupon, at 4:40 p.m., the hearing was adjourned.]

[Additional material submitted for the record follows.]

APPENDIX

ADDITIONAL MATERIAL SUBMITTED FOR THE RECORD

Witness List
Hearing before the
Senate Committee on the Judiciary
Subcommittee on Intellectual Property

“Artificial Intelligence and Intellectual Property – Part II: Copyright”

Wednesday, July 12, 2023
Dirksen Senate Office Building, Room 226
3:00 p.m.

Ben Brooks
Head of Public Policy
Stability AI
San Francisco, CA

Dana Rao
Executive Vice President, General Counsel, and Chief Trust Officer
Adobe Inc.
San Jose, CA

Matthew Sag
Professor of Law, Artificial Intelligence, Machine Learning, and Data Science
Emory University School of Law
Atlanta, GA

Karla Ortiz
Concept Artist, Illustrator, and Fine Artist
San Francisco, CA

Jeffrey Harleston
General Counsel and Executive Vice President of Business and Legal Affairs
Universal Music Group
Santa Monica, CA

stability.ai

Artificial Intelligence and Intellectual Property – Copyright

Senate Judiciary Subcommittee on Intellectual Property

Wednesday 12 July

Ben Brooks

Head of Public Policy

Stability AI

<i>Introduction</i>	1
<i>Background</i>	2
<i>Stability AI develops open technology to drive transparency, competition, and innovation in AI</i>	2
<i>We believe AI is a tool, not a substitute for creators</i>	4
<i>AI models learn the relationships between words, concepts, and visual or textual features</i>	5
<i>How AI is trained, deployed, and used has implications for future policy</i>	7
<i>We are actively working to address emerging concerns and mitigate risks</i>	9
<i>Conclusion</i>	10

Introduction

Thank you, Chair Coons and Ranking Member Tillis. We welcome the opportunity to appear today, and to share our perspective on these important issues. AI is a tool that can help to unlock creativity, drive innovation, and open up new opportunities for creators and developers across the United States. With appropriate safeguards, we are working to make these foundational technologies accessible to all. As with any groundbreaking technology, AI raises important questions about the integration of these tools into our everyday lives, including in the creative economy. We recognize these challenges and the depth of concern among creators. While we don't have all the answers, we are committed to an open and constructive dialogue, and we are actively working to address these concerns through technology, standards, and best practices.

Background

Stability AI is a global company that aims to unlock humanity's potential by making foundational AI technology accessible to all. Today, Stability AI develops a variety of generative AI models. These models are software programs that analyze vast datasets to learn the relationships between words, ideas, and fundamental visual or textual features. They can apply this knowledge to help a user generate new content. In 2022, we took over the exclusive development of Stable Diffusion, a popular image model that takes a text instruction or "prompt" from a user and helps to produce new images. In 2023, we released the first in a series of language models known as StableLM to support open research into AI safety. These language models can take a prompt from a user and help to produce new passages of text or software code.

By some measures, developer interest in Stable Diffusion has grown faster than many open-source software projects in recent history, and over 200,000 creators and developers actively contribute to the Stability AI community. Stability AI partners with organizations to adapt these models for specific purposes, helping to sustain our open research and development efforts. In addition, Stability AI provides services to help developers and users access the computing resources necessary to train or run these models, and actively supports research into scientific applications of AI.

Stability AI develops open technology to drive transparency, competition, and innovation in AI

With appropriate safeguards, we release our flagship AI models openly as a public resource, subject to ethical use licenses. That means we publish the software code along with the billions of distinctive settings ("parameters") that define the model's performance. Creators and developers can freely adapt these open models. For example, they can customize their own AI models for specialized tasks, power their own AI applications, or build their own AI businesses. Open models are different to closed models, which are software programs that cannot be inspected or adapted by third-parties. Closed models may not disclose how they are trained or how they operate.

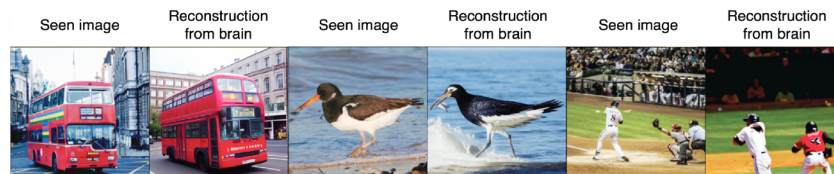
Open models can help to promote transparency, competition, and innovation in the digital economy. Within the AI ecosystem, open models play an important role in several ways:

1. **Promoting transparency.** AI models will form the backbone of our digital economy, and everyone should have a voice in their design. They will power chatbots that provide medical advice or personalized tutoring; transform how we search, access, or manage information; and support high-stakes analysis or decision-making in some of our most important public and private institutions. By releasing open models, researchers, authorities, and developers can "look under the hood" to verify the performance and suitability of a model for a particular task. They can identify potential risks, and help develop new mitigations. Likewise, they can inspect open datasets for quality, fairness, or

bias. Because these models are transparent, firms and agencies in sensitive sectors can build on open models to produce their own specialized models for particular tasks.¹

2. **Fostering competition.** Developing a generative AI model typically requires significant resources. Open models can lower these barriers to entry, fueling competition in AI.² Using open models, creators and developers can deploy new AI tools or launch new AI ventures without spending millions on research and computing power. They can participate in this new industrial revolution as builders – not just consumers – of AI technology, and they can do so without relying on a handful of firms for critical infrastructure.
3. **Supporting grassroots innovation.** Grassroots innovation by anyone, anywhere is one of America’s greatest assets, and open models put these capabilities in the hands of everyday creators, developers, and researchers. Today, models like Stable Diffusion are used for everything from editing photographs to prototyping architectural designs to researching new techniques for diagnosing complex disorders. We can’t anticipate every possible use of AI, but everyday people can experiment with open models to develop new and innovative applications that support their work and serve their community. In this way, open models can help distribute the economic benefits of AI across the United States, beyond Silicon Valley.

We build models to support and augment our users, not replace them. We develop tools that help everyday people and everyday firms. We are focused on practical AI capabilities that can be applied to everyday tasks – not a quest for an alien superintelligence. Designing around these principles can help to unlock the useful potential of AI while minimizing the risk of misuse, weaponization, or “runaway” systems.



Above: Multiple teams are experimenting with image models like Stable Diffusion to research new techniques for studying medical disorders. Here, MedARC researchers demonstrate how latent diffusion models can be used to visualize how patients are perceiving a given scene, based only on fMRI scans. Disturbances in perception could help to assess or diagnose complex neurological conditions.³

¹ For example, a regulated financial institution may customize AI models to assist in analysis, decision making, or customer support. The financial institution may need to audit the performance of the model for reliability; train the model without exposing sensitive customer data to third-parties; and retain full control over the AI model without relying on a third-party provider. By building on open models, a financial institution can train and manage their own AI system.

² See, e.g. the Hugging Face “leaderboard” comparing open language models, available [here](#).

³ MedARC projects are available [here](#).

AI is a tool, not a substitute for creators

We believe that AI is a tool that can help creators express themselves, but it is not a substitute for creators. Instead, AI can help to accelerate the creative process. AI tools can help existing creators boost their productivity, experiment with new concepts, and perform complex tasks as part of a wider workflow. In addition, AI can lower barriers to entry for people who do not have the resources or training to realize their creative potential. Instead of simply consuming the best available content, these “untapped” creators will be able to produce their best imaginable content. As with other assistive technologies – from paintbrushes to cameras to editing software – the creator retains ultimate control over the composition and use of their work.

In their own words

"As a designer, I've used Stable Diffusion to visualize ideas in a matter of minutes. I'll use it as a tool to create different iterations of an idea and manipulate quickly. I'll create moodboards and storyboards with it for other team members to see where we are headed."

– Stefania Bulbarella, Broadway Projection Designer

"As a designer and educator working with generative AI, I have consistently observed that groundbreaking innovations primarily originate from open-source platforms. The ability to fine-tune models and weights through the collaborative efforts of diverse user communities is a critical counterpoint to the potential biases and decisions imposed by closed-source AI corporations. Therefore, advocating for the widespread adoption and endorsement of open-source AI frameworks is imperative, as it not only fortifies the United States' standing as the forefront of AI research and development but also stimulates the growth, resilience, and variety within the field."

– Andrew Kudless, Architect

"In the past year, I've used Stable Diffusion as an early conceptual design exploration tool in my contemplative architectural designs and also as a moodboard dynamic library of original reference images to visually communicate lighting design ideas in similar spaces with similar materials."

– Ilva Dodaj, Architect and Lighting Designer

"The tool fits into my workflow such that I can spend more time doing what I love most, which is designing thoughtful solutions for spaces that will empower its users. Generative AI does not threaten the taste, reasoning, and empathy required to do that."

– Isabella Orsi, Interior Designer⁴

⁴ Feedback from Bulbarella, Kudless, and Dodaj was kindly shared with Stability AI over the past few days. Orsi shared feedback on Twitter and with the [New York Times](#).

AI models learn the relationships between words, concepts, and visual or textual features

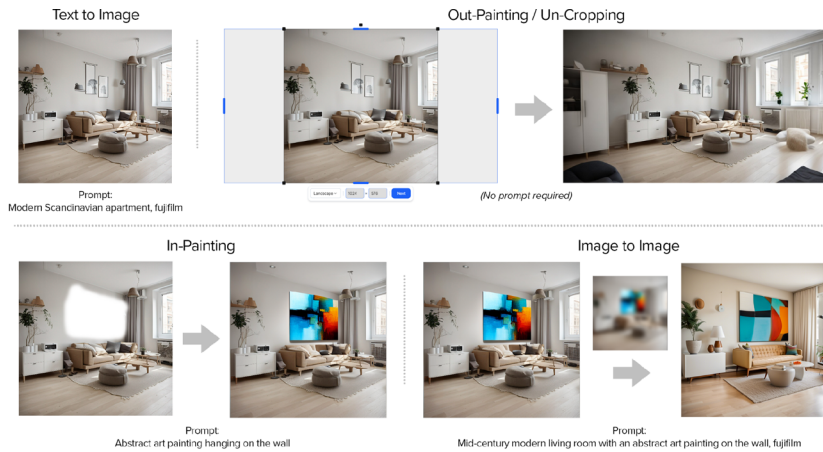
Recent AI models are described as “generative” AI because they can generate complex new content, helping to simplify analytic or creative tasks. These models analyze vast datasets to understand the relationships between words, concepts, and visual or textual features – much like a student visiting a library or an art gallery. Models can then apply this knowledge to help a user produce new content. This learning process is known as training.

Image

For example, during training, an image model such as Stable Diffusion will review billions of pairs of images with associated text captions.⁵ Through this process, the model learns to identify fundamental visual structures within images, such as shapes, textures, and patterns. By cross-referencing with known text captions, the model learns to associate these fundamental structures with particular terms. For example, the model may learn to understand the appearance of fur on a “dog”; learn how light interacts with “water”; or capture the visual aesthetic described by words like “bleak” or “Renaissance”. When a user inputs a prompt – such as “a photorealistic astronaut riding a pig” – the model can help to express the desired features in a new image, even if the model has never seen an example of that composition.

Image models are typically trained on large datasets. For example, Stable Diffusion was trained on a filtered subset of two billion image and caption pairs from LAION-5B, an open dataset of publicly-accessible image links and captions compiled by the European nonprofit Large AI Open Network (LAION). Following initial training, the model can be further trained or customized with additional data to improve performance in specific tasks.

⁵ Stable Diffusion is a type of latent diffusion model.



Above: Image models like Stable Diffusion can be used in a range of ways as part of a design workflow. They can help to produce new images based on a text description, fill in or replace parts of an existing image, extend parts of an existing image, or transform an existing image.

Language

Likewise, recent language models are trained on vast quantities of existing text. The model is typically shown incomplete fragments of text, broken into digestible units known as tokens, and it then studies the relationships between and among the tokens to detect complex patterns in how sentences, paragraphs, and passages are structured in different contexts.

In this way, the model learns underlying structures within a language system. For example, the model can identify the typical arrangement of a “blog”, “email”, or “speech”. It can understand the expected structure of an “example” or an “explanation”, and the textual features that contribute to tones like “funny” or “formal”. The model can learn what it means to distinguish, summarize, expand, or combine different concepts, such as “banana”, “Wall Street”, or programming instructions like “print()”. When a user inputs a prompt – such as “suggest some ways to improve this draft greeting card” – the model can interpret the request, and respond with a passage of text.

Language models are typically trained on datasets consisting of a trillion or more tokens of content (equivalent to hundreds of billions of words). Following initial training, these models can be “fine-tuned” to improve performance in specific domains, and to mitigate potential risks such as bias, misinformation, or toxicity. For example, reinforcement learning from human feedback

(RLHF) can help to improve the performance of a model by using human feedback to “rate” the model’s outputs. Likewise, instruction-tuning techniques can help to improve performance by providing labeled examples of sample inputs and desired outputs. Stability AI is actively engaged in research to improve safety through fine-tuning. This work draws on a number of open datasets shared by researchers for this purpose.

1 I'm not a native English speaker. I find it hard to understand grammar. Can you help me by explaining prepositions?

2 Sure, I'd be happy to help you understand prepositions!

Prepositions are words that show the relationship between a noun or pronoun and other words in a sentence. They are used to indicate location, direction, time, manner, and other relationships.

Here are some common prepositions and their meanings:

- "in": indicates location within an area, such as "I am sitting in the living room."
- "on": indicates location on an object, such as "I am holding the book on my lap."
- "at": indicates location at a specific time or place, such as "I met him at the coffee shop."
- "by": indicates a means of transportation or a means of achieving something, such as "I got there by walking."
- "with": indicates association or accompaniment, such as "I am friends with John."
- "from": indicates origin, such as "I am from the United States."
- "to": indicates direction, such as "I am going to the store."

Prepositions are usually used before a noun or pronoun, but they can also be used in other positions in a sentence, such as at the beginning or end.

3 Can you help me implement a bubble sort in Python?

4 Sure, here's an example implementation of bubble sort in Python:

```
python
def bubble_sort(arr):
    n = len(arr)
    for i in range(n):
        for j in range(n-i-1):
            if arr[j] > arr[j+1]:
                arr[j], arr[j+1] = arr[j+1], arr[j]
    return arr
```

This function takes in an array `arr` as input and returns the sorted array. The algorithm works by repeatedly swapping adjacent elements if they are in the wrong order. The `n` variable is set to the length of the array, and the outer loop runs `n` times, while the inner loop runs `n-i-1` times for each iteration of the outer loop, where `i` is the current iteration number. This ensures that the largest element is compared and swapped with the first element of the inner loop, and so on, until the array is sorted.

Above: Language models can be used in a range of ways as part of a creative workflow. For example, they can help to draft or edit documents or correspondence, or help to identify bugs in software code and suggest improvements.

How AI is trained, deployed, and used has implications for future policy

The training, deployment, and use of AI models has implications for how we should think about the interaction of AI tools with copyright. Future policy should account for several characteristics:

1. **Training data is not stored in the model.** Through training, these models develop an understanding of the relationship between words, concepts, and fundamental visual or textual features. The model doesn’t rely on any single work in the training data, but instead learns by observing recurring patterns over vast datasets (billions of image and caption pairs, and hundreds of billions or trillions of words). The model does not store the material in this training data. They do not “collage” or “stitch” together original works, nor do they operate as a “search engine” for existing content.
2. **These models help users to produce new works.** Models apply this knowledge to help a user generate new and unseen content. That could mean a novel image, passage of text, block of code, series of instructions, or video clip.⁶ This knowledge is generalizable, which means it can help to develop new content and support new tasks that did not appear in the training data.

⁶ In addition, these models can be used as analytic tools to detect and classify features within content, such as visual patterns, textual information, or software bugs.

3. **The model is a creative tool, not an independent agent.** The model is a tool that helps to produce this content at the request of a user. The user provides creative direction by supplying text prompts or reference examples, and adjusting other settings. The user ultimately determines how the generated content is shared, displayed, or represented to others downstream.
4. **Models may be developed, customized, and deployed by different entities.** Models may be released as free and open software to promote transparency, competition, and innovation. The developer of a free and open model may have no relationship with downstream developers. For example, a creator, researcher, or business can choose to fine-tune these models for specialized tasks by training on their own data. In this environment, the original model developer may have no financial or contractual relationship with those who actually customize or deploy the model.

Training these models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law, including to “promote the progress of science and useful arts”.⁷ These models learn the unprotectable ideas, facts, and structures within a visual or textual system, and that process does not interfere with the use and enjoyment of original works. “Free learning” of these facts about our world is essential to recent developments in AI, and it is doubtful that these groundbreaking technologies would be possible without it. The US has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation. Other jurisdictions, including Singapore, Japan, and the European Union, have begun to incrementally revise their copyright laws to create safe harbors for AI training that achieve similar effects to fair use.



Above: These models do not “stitch together” original works. For example, image models notoriously struggle with hands. In this case, the model produces a nine-fingered hand because it hasn’t understood that a hand typically has five fingers, and it isn’t searching a database of images with hands. Instead, it has learned that a “hand” means a flesh-colored artefact associated with some number of appendages.

⁷ U.S. Constitution, Article I, Section 8, Clause 8.

We are actively working to address emerging concerns and mitigate risks

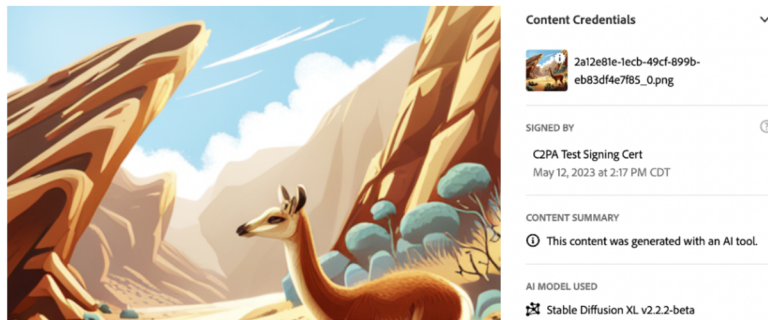
We understand that developments in AI can feel highly disruptive, and raise important questions about the application of existing law and public expectations to new technology. We are actively working to address emerging concerns. In particular, we support greater transparency in how AI content is used, produced, and shared through the digital economy. To that end, we are focusing our efforts on a number of initiatives:

1. **Best practices in training.** Because Stable Diffusion is pre-trained with content from open datasets, creators can determine whether their works appear in those datasets.⁸ Stability AI has proactively solicited opt-out requests from creators, and will honor these over 160 million opt-out requests in upcoming training for new Stable Diffusion models. Going forward, we are exploring new technical standards for machine-readable opt-outs, so that opt-out metadata follows the content wherever it goes. In addition, the datasets used for training Stable Diffusion respected industry-standard digital protocols like robots.txt, which indicates whether a website consents to automated data collection for ancillary purposes such as indexing or analysis.
2. **Identifying AI content.** We are implementing content authenticity standards and watermarking that can help users and content platforms better identify AI-assisted content.⁹ Images generated through our hosted API can include metadata and imperceptible watermarks to indicate the content was produced with an AI tool. The metadata is certified with the relevant model and version number. Downstream intermediaries – such as social media platforms – can use metadata, watermarks, and other signals to assess the provenance of content before amplifying it through a network. Together, these features can help platforms to distinguish AI content, enable users to exercise appropriate care when interacting with AI content, and help to limit the spread of misinformation with AI tools.
3. **Quality of datasets.** We are working to improve the quality of datasets to avoid unintended behavior in our models. For example, improving diversity and reducing duplication in training data can help to prevent “overfitting”, a training defect which occurs when a model erroneously overrepresents elements of a particular image from the dataset (e.g. if a model has only seen sunsets, it might learn that the sky is always orange). These measures can help to mitigate the risk of a user prompting a model with sufficient detail or determination to generate a lookalike work. Additionally, datasets that are more representative of diverse culture, language, demographics, and values can help to mitigate the risk of bias in outputs. We are actively investing in efforts to improve representation in these datasets.

⁸ See, e.g. the “Have I Been Trained?” portal via Spawning.AI, available [here](#).

⁹ In partnership with the Content Authenticity Initiative led by Adobe, available [here](#), we are implementing the C2PA standard for content metadata.

4. **Mitigations for misuse.** We have implemented a range of mitigations to prevent other forms of misuse. For example, we impose licensing conditions that balance open innovation with harm minimization. Our flagship image model, Stable Diffusion, is governed by the Open Responsible AI License (Open RAIL), which prohibits a range of unlawful or unethical uses.¹⁰ We publish standardized information about the performance and limitations of our models (see **Annex A**). In versions of Stable Diffusion developed exclusively by Stability AI, we apply filters on training data to remove unsafe images. By removing that data before it ever reaches the model, we can help to prevent users from generating harmful images. In addition, on Stability AI applications and APIs, we have developed and implemented effective filters that block unsafe text prompts or generated images that violate our platform terms of service.¹¹ In our language model research, we are exploring techniques to improve the screening of training data for personally identifiable information to mitigate the risk of inadvertent re-disclosure.



Above: An example of content authenticity metadata indicating an image was generated with an AI tool.

Conclusion

AI is a tool that will unlock creativity, innovation, and productivity. With appropriate safeguards, open models can help to put these tools in the hands of everyday creators and developers. As AI is integrated into the digital economy, we believe the community will continue to value human-generated content. Indeed, that content may be valued at a premium. Smartphones didn't destroy photography, and word processors didn't diminish literature, despite radically transforming the economics of creation. Instead, they gave rise to new demand for services, new markets for content, and new creators. We expect the same will be true of AI tools, and we welcome an ongoing dialogue with the creative community about the fair deployment of these technologies.

¹⁰ Open Responsible AI License (OpenRAIL) available [here](#). Our license prohibits the unlawful, exploitative, or misleading use of Stable Diffusion.

¹¹ See, e.g. terms of service for the API available [here](#).

Annex A. Stability AI releases standardized information describing the performance, limitations, and biases of our models, in addition to other disclosures such as environmental impact. Examples of these disclosures include:¹²

<p>Misuse and Malicious Use</p> <p>Using the model to generate content that is cruel to individuals is a misuse of this model. This includes, but is not limited to:</p> <ul style="list-style-type: none"> • Generating demeaning, dehumanizing, or otherwise harmful representations of people or their environments, cultures, religions, etc. • Intentionally promoting or propagating discriminatory content or harmful stereotypes. • Impersonating individuals without their consent. • Sexual content without consent of the people who might see it. • Mis- and disinformation • Representations of egregious violence and gore • Sharing of copyrighted or licensed material in violation of its terms of use. • Sharing content that is an alteration of copyrighted or licensed material in violation of its terms of use. 	<p>Limitations</p> <ul style="list-style-type: none"> • The model does not achieve perfect photorealism • The model cannot render legible text • The model does not perform well on more difficult tasks which involve compositionality, such as rendering an image corresponding to "A red cube on top of a blue sphere" • Faces and people in general may not be generated properly. • The model was trained mainly with English captions and will not work as well in other languages. • The autoencoding part of the model is lossy • The model was trained on a subset of the large-scale dataset LAION-5B, which contains adult, violent and sexual content. To partially mitigate this, we have filtered the dataset using LAION's NFWF detector (see Training section).
<p>Bias</p> <p>While the capabilities of image generation models are impressive, they can also reinforce or exacerbate social biases. Stable Diffusion was primarily trained on subsets of LAION-2B(en), which consists of images that are limited to English descriptions. Texts and images from communities and cultures that use other languages are likely to be insufficiently accounted for. This affects the overall output of the model, as white and western cultures are often set as the default. Further, the ability of the model to generate content with non-English prompts is significantly worse than with English-language prompts. Stable Diffusion v2 mirrors and exacerbates biases to such a degree that viewer discretion must be advised irrespective of the input or its intent.</p>	<p>Environmental Impact</p> <p>Stable Diffusion v1 Estimated Emissions Based on that information, we estimate the following CO2 emissions using the Machine Learning Impact Calculator presented in Lacoste et al. (2019). The hardware, runtime, cloud provider, and compute region were utilized to estimate the carbon impact.</p> <ul style="list-style-type: none"> • Hardware Type: A100 PCIe 40GB • Hours used: 200000 • Cloud Provider: AWS • Compute Region: US-east • Carbon Emitted (Power consumption x Time x Carbon produced based on location of power grid): 15000 kg CO2 eq.

¹² See 'Stable Diffusion V2-1 Model Card' available [here](#).

**Written Testimony of
Jeff Harleston
General Counsel and Executive Vice President of Business and Legal Affairs,
Universal Music Group**

Artificial Intelligence and Intellectual Property – Part II: Copyright

**U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property
July 12, 2023**

Thank you, Chairman Coons, Ranking Member Tillis, and members of the Subcommittee. I'm Jeff Harleston, the General Counsel and Executive Vice President of Business and Legal Affairs at Universal Music Group (UMG), the world leader in music-based entertainment. It's an honor to be here today to testify on "AI and Copyright."

UMG is home to a broad array of businesses engaged in recorded music, music publishing, merchandising, and audiovisual content. Featuring the most comprehensive catalogue of recordings and songs across every musical genre, UMG identifies and develops artists and produces and distributes the most critically acclaimed and commercially successful music in the world. Committed to artistry, innovation, and entrepreneurship, UMG fosters the development of services, platforms, and business models in order to broaden artistic and commercial opportunities for our artists and create new experiences for fans.

In my thirty years with the company, I've lived through the music business experiencing tremendous growth and equally tremendous challenges. I have learned many lessons, having navigated the technological changes in our industry during my time with the company. I've led the storied Def Jam record label, worked for several years in management at Geffen Records, and helped broker deals with the services, platforms, and social media outlets where you access the music you love. It's been my life's honor to work with countless talented and creative artists, the ones who've occupied your playlists, sang as you danced at your wedding reception, or helped you process unspeakable grief. They have shared with us experiences we will never have and taken us places we will never go. Their creativity is the soundtrack to our lives. And without the fundamentals of copyright, we might not ever have known them.

Given the explosion of generative artificial intelligence (AI) over the past several months, I'm grateful for the opportunity to share our views on the impact of AI on creators, most specifically the songwriters and artists we represent.

If I can leave you with one message today, it's this: AI in the service of artists and creativity is a wonderful thing. But AI that uses, or worse yet, appropriates their work – or their name, image, likeness, or voice – without authorization is *not*. An artist's work and persona are valuable and

deeply personal. As you establish the rules of the road for artificial intelligence, your use of that guiding principle can help ensure that our creative class and the businesses that support them continue to enrich our world.

I. AI and the Music Industry

Long before an AI-generated recording imitating Drake and The Weeknd – both Universal Music artists – went viral and captured the attention of press and policymakers, UMG has been thinking about artificial intelligence. One of our companies, Ingrooves, has three patents in AI to assist with marketing independent artists. And AI has long been used as a tool in the studio: For example, Apple Logic Pro X to generate drum tracks, or Captain Plugins to generate chord progressions. We also use AI regularly as a tool to assist in creating Dolby Atmos immersive audio music. It's a great technology when employed responsibly – and one that we and our artists use.

However, we are before you today because generative AI is raising fundamental issues of responsibility in the creative industries and copyright space. Each day, troubling examples emerge. We know some generative AI engines have been trained on our copyrighted library of recordings and lyrics, image generators have been trained on our copyrighted cover art, and music generators have been trained on our copyrighted music, all without authorization.

We have a robust digital music marketplace, and UMG has hundreds of legitimate partners who've worked with us to bring music to fans in a myriad of ways. Those companies and services properly obtained the rights they need to operate from UMG, or from the associated record labels and publishers. So, it's unfathomable to think AI companies and developers think the rules and laws that apply to other companies and developers don't apply to them.

Beyond the issue of copyright infringement, these generative AI companies are often obtaining our content from sources that explicitly prohibit downloading and use of that content outside of personal and non-commercial purposes. We've also seen examples of AI-generated music being used to generate fraudulent plays on streaming services, siphoning income from human creators. And we've seen many troubling cases where an artist's name, image, likeness, or voice is used without their knowledge or authorization – to generate videos of them saying things they didn't say, to use their voice and recordings without their knowledge, or to exploit their name to promote fraudulent works.

To be clear: the core copyright industries add roughly \$1.8 trillion of value to U.S. GDP annually and employs 9.6 million American workers. When creative industries like music are harmed, the impact to our nation's economic health is real. And music is one of our great cultural exports. Music is a key element of our cultural mosaic. We need to protect music – and creators – from harm in this new era.

But while the risk is real, we have no naïve intention of sticking our head in the sand. It's very easy for those that steal copyrighted works and disrespect artists to rush into the market – it's

more work to do it legally and respectfully. It takes time to figure out and create a means to ensure legality and make sure artists get identified, paid, and respected. I am very confident we will get there, because we always do. We have a robust free market for sampling, sync licensing, deals with new entrants to the digital marketplace, social media companies and all manner of new technologies.

But, in order to get there, we need to ensure any potential partnerships are responsible, ethical, and artist-centric, protecting the livelihoods of artists, rightsholders, and the creative ecosystem from harm. And it's possible to use AI in this way.

An example: one of our distributed artists, South Korean artist Lee Hyun, used a generative AI system to train on recordings of his vocals – allowing him to simultaneously release his single in six languages – in his own voice – on the same day. Here, the ethically trained tool enhanced and extended the artist's creative intent – with his *consent* – enabling him to reach new markets and fans.

Imagine an artist or songwriter creating personalized recordings for fans – such as “Happy Birthday” or perhaps a brand-new creation. Imagine the ability to perform a concert and livestream it globally in dozens of languages simultaneously. Imagine collaborations never even dreamed before. It's a powerful technology – and once there is a more responsible, legal, and respectful AI marketplace, the opportunity for creators is vast.

We're doing all we can to encourage AI developers to act legally and with the artists' long-term benefit in mind. And we practice what we preach: as the General Counsel of the company, my edict has been clear: We respect copyright. Not just our own, but the copyrights of others as well. We avoid AI tools which we believe were illegally trained on copyrighted materials. And we have an AI review team internally that reviews projects and potential tools to effectuate that effort.

II. Guiding Principles for Sound AI Policy

Many in the music community have joined more than 140 trade bodies, collectives, and organizations representing creators and creative industries around the world – along with organized labor, athletes, and others – to support the Human Artistry Campaign. We concur with the campaign's seven high-level guiding principles for sound AI policy:

1. Technology has long empowered human expression, and AI will be no different;
2. Human created works will continue to play an essential role in our lives;
3. Use of copyrighted works and the use of voices and likenesses of professional performers requires authorization and free market licensing from all rights holders;
4. Governments should not create new copyright or other IP exemptions that allow AI

developers to exploit creations without permission or compensation;

5. Copyright should only protect the unique value of human intellectual creativity;
6. Trustworthiness and transparency are essential to the success of AI and protection of creators; and
7. Creators' interests must be represented in policy making.

These are our guiding principles as we navigate generative AI. We encourage you to be mindful of them as you continue to craft and establish policies in this new technological era.

III. Specific Policy Suggestions

Currently, copyright law is largely fit for purpose – but there are cracks in the foundation I urge you to address. This is especially important if you want to future-proof the law to fit a technology evolving and advancing at breakneck speed. Moreover, if you want to cultivate a lawful, legitimate, trustworthy, and responsible AI marketplace, Congress needs to establish fundamental rules of the road that enable that development and ensure creators are respected and protected.

First, we urge you to enact a federal Right of Publicity statute. Deep-fake and/or unauthorized recordings or visuals of artists generated by AI could lead to consumer confusion, unfair competition against the actual artist, market dilution and damage to the artist's reputation and brand – potentially irreparably harming their career. An artist's voice is the most valuable part of their livelihood and public persona, and to steal it – no matter the means – is wrong.

AI-generated, mimicked vocals trained on vocal recordings extracted from our copyrighted recordings go beyond Right of Publicity violations and concerns about consumer deception, unfair trade practices, and privacy – copyright law has clearly been violated.

The fact that these vocal-cloned recordings are often covers means that both the underlying music work *and* the original track are being infringed. In fact, when the voice cloning AI is trained using copyrighted recordings, there are likely multiple infringements occurring, including:

- infringement when an artist's vocal stem is extracted from an existing sound recording and reproduced;
- infringement of *both* the musical work and the sound recording copyrights when an unauthorized reproduction is made to train the AI on the artists' vocals;
- infringement under Sec. 1201 if stream-ripping is used to extract those vocals; and

- infringement under Sec. 1202 if the Copyright Management Information (CMI) has been stripped from the original sound recording or as part of the process.

I imagine most of us here don't go digging for AI-generated music, but the problem is vast. It's easy to find online instructions telling people how to rip vocals from songs. Today, I entered the name of a popular artist/composer in an internet search engine with a reference to "AI Song" and it returned tens of millions of different links. Swap "song" for "lyrics" and I received tens of millions more. Pick any major singer songwriter and you'll likely get similar results. Many of those results might not be infringing upon inspection, but it's still a staggering scope of search returns associated with unauthorized generative AI and worthy of your attention.

Secondly, copyright owners must be able to see what's gone into a training data set without having to initiate full blown litigation. Ensuring the transparency of this data is paramount. Without it, you've stymied potential for a marketplace. Records cataloging the provenance of materials, including any and all authorizations or licenses of that material, must be kept by AI developers. Litigation takes time and great expense. Being forced to use litigation to gain visibility on the quantity and extent of infringed material is truly an insurmountable roadblock for individual creators. There needs to be a mechanism for rightsholders to be able to inspect these records on training data - without having to initiate litigation.

But as legislative policy on transparency is developing, I urge you to use your platform and position of power to reiterate the principle. Believe me, it has a real impact – as AI companies meet with you, ask about their training data and what authorizations they've received to use it. Remind AI developers of the law and signal its importance to you. It's blatant infringement to use copyrighted works for training without authorization, and it needs to stop. Frankly, it's a bit astonishing that this even has to be litigated.

Finally, fully AI-generated content should be labeled as such. Consumers deserve to know what they're getting. Was it created by a human, or machine? Is the audio or visual they're consuming real? Is that really their favorite artist's voice? At UMG, we are committed to protecting our artists and the authenticity of their creative works. Within the music industry, we're pioneering what it means for labels and publishers to be good stewards of this principle.

IV. Conclusion

As with other technological advances in the past, it would be wrong to view AI as nothing but a threat. In fact, when responsibly developed and employed, we see AI as a powerful tool for our future. In the service of artists, AI can and will enhance our business in the years to come.

But since unchecked "generative AI" poses many dangers, we support efforts to ensure that generative AI thrives as a technology that enhances rather than threatens human creativity, and one that protects the rights of artists, their livelihoods, the creative ecosystem and culture as a whole. In addition to exploring other options in this rapidly developing area, UMG is actively cultivating what we call "Responsible AI" opportunities by talking with leading AI music

companies, and by advocating with early-stage AI companies as their own technology unfolds. We will have more to announce in this space in the near future.

You have an opportunity to establish legal clarity that creates a brighter path towards generative AI legitimacy and legality. Those solutions will not always require legislation, but some, like a federal Right of Publicity statute or requiring the transparency of training datasets, require swift, decisive legislative action. We stand ready to work with you and your colleagues to find productive, practical solutions in the interest of creators and the longevity of our rich culture.

I look forward to answering your questions. Thank you.



Written Testimony of Karla Ortiz
U.S. Senate Judiciary Subcommittee on Intellectual Property
“AI and Copyright”
 July 7th, 2023

Chairman Coons, Ranking Member Tillis, and esteemed members of the Judiciary Committee, thank you for inviting me to testify today about AI and Copyright. “AI” stands for “artificial intelligence.” But that’s a misleading term, because, in fact, these so-called artificial intelligence systems depend entirely on vast quantities of copyrighted work made by human creators like me. These AI companies use our work as training data and raw materials for their AI models without consent, credit, or compensation. Today, I would like to share with you why these practices represent a dire threat to many American jobs, including mine. I will also explain why making AI fair and ethical for everyone is essential to building a sustainable relationship between AI and copyright.

My name is Karla Ortiz, a concept artist and illustrator based in San Francisco. You may not know my name, but you and your loved ones have almost certainly seen my work. My work has helped shape and create the worlds of blockbuster films including Marvel projects like *Guardians of the Galaxy Vol. 3*, *Loki*, *The Eternals*, *Black Panther*, *Avengers :Infinity War*, and *Doctor Strange*. I have also submitted art and designs for video games such as *Final Fantasy 16*, and trading card games like *Magic: the Gathering*. I am passionate about my work as an artist. Thanks to my hard work, my passion for the arts, and the support of my community, my work has been internationally recognized, and my fine art has been presented in countless galleries both domestically and internationally. There are hundreds, maybe thousands, of concept artists and illustrators like me who work tirelessly to make the films, TV, video games, and other media that bring entertainment to millions of Americans every day.

I am no longer certain of my future as an artist—a new technology has emerged that represents an existential threat to our careers: generative artificial intelligence (“Generative AI”). Generative AI is unlike any tool that has come before, as it is a technology that uniquely consumes and exploits the innovation of others. As I will explain in more detail, Generative AI relies on data it ingests in order to train models and generate its output, and oftentimes, that training data is the work of creative people like myself, taken without our consent, without any credit, and without any compensation. Based on this training data, Generative AI is capable of generating facsimiles of creative people’s work, including my own. As a result, AI has grown rapidly—in 2022, research estimated the value of the global AI market at approximately \$100 billion USD, and is expected to grow to nearly \$2 trillion USD by 2030. No other tool has the potential to adversely affect the rights of artists and creators. No other tool solely relies on the works of others to generate imagery. Not Photoshop, not 3D, not the camera, nothing comes close to this

technology. I should know, I've worked with almost all those tools. And even though artists and creatives are among the first American workers who are confronting the threat posed by AI on the job market, we are certainly not the last. Generative AI threatens the livelihoods of an untold number of Americans.

I am not fundamentally opposed to Generative AI. But AI needs to be fair, and ethical for everybody—and not only for the companies that make AI products. AI needs to be fair to the customers who use these products, and also for creative people like me who make the raw material that these AI materials depend upon. These systems depend entirely on the work of humans, especially creatives such as visual artists, writers and musicians.

The most important thing I can tell you is something the AI companies would prefer to overlook: “artificial intelligence” is really just human intelligence, repackaged and made available in a high-tech package. These models are useless alone, and need to be trained on an immense amount of data, including the work of artists of all creative fields. This is because generative AI is not truly “intelligent”—a Generative AI model cannot create anything it has not already seen in its training data. And therein lies the problem—the companies that make AI products have decided to neither ask for permission nor ask for forgiveness. Rather than ask artists and creatives their consent to use their work to train their Generative AI models, AI companies have instead opted to ingest massive amounts of copyrighted and licensed data to use as training data without permission, without crediting the creators of the underlying work, and without any compensation.

As a result of their wholesale ingestion of ill-gotten data, AI companies have reaped untold billions in funding and profit. Unsurprisingly, the AI companies have assured everyone that what they are doing is fair, ethical and legal. But the artists who made the works that their AI's rely on have never been asked for their consent, have not received any credit, let alone any compensation. In any other sphere, these practices would offend basic principles of fundamental fairness. And while the AI companies claim that what they are doing is fair use, but this claim has not yet been litigated. But the courts are beginning to weigh in. In May, a federal judge sustained important parts of a complaint filed by coders challenging the use of their code as training data for Generative AI models without regard to the requirements of the open-source licenses that code was subject to.¹ I myself have brought a lawsuit against Midjourney, Stability AI and DeviantArt for the use of my art as training material for their AI products.² While the courts have offered an important counterbalance against the interests of these big technology companies by providing an avenue to attempt to safeguard or vindicate important rights and to bring attention to this important issue, lawsuits may not be enough.

As I mentioned before, even though artists may be the first people to be affected by AI, the practices employed to develop these Generative AI models can be adapted to any kind of human expertise. Today, it might be me, who designed a loved superhero; tomorrow, it could be nurses, teachers, accountants, doctors, architects or any number of professions.

In our system, it is the role of Congress to pass laws, and to empower agencies to develop regulations to help our society deal with changes in time and technology—for example, ensuring AI companies have

¹ Order Granting in Part and Denying in Part Motions to Dismiss, *Doe 1 v. GitHub, Inc.*, Case No. 22-cv-06823-JST (N.D. Cal. May 11, 2023), ECF No. 95.

² *Andersen v. Stability AI Ltd.*, Case No. 23-cv-00201 (N.D. Cal.).

artists opt-in to having their works included as training data, laws that offer artists more robust methods of vindicating their rights or ensuring all AI companies to be completely transparent in their training data and contain no copyrighted works, biometric data or private data is utilized without right holders explicit permission. It is my hope that my testimony will help you to promulgate common sense laws that will make AI fair and ethical for all.

I. My First Encounters with Generative AI & Lessons Learned

My first encounter with generative artificial intelligence art models occurred a little over a year ago in April 2022. I stumbled upon a website called Weird and Wonderful AI Art. This website featured countless images alongside the names of artists, some of whom were dear friends of mine. I realized all the images were synthetic media, or generative AI images. These images claimed to be “studies,” attempting to make these synthetic images look just like the work of my peers. My initial reaction was one of interest. I was curious, so I reached out to my friends whose names were featured on that website. None of them were aware of the website, however. In fact, not a single artist had been asked to be a part of this “study”. When we tried to reach out to the folks who were running that website, to ask them to please remove the artists who did not want to participate, **we got ghosted**. Little did I know at the time that this would only be the first of many similar experiences I’ve had with generative AI.

In August and September of 2022, larger Generative AI models like Midjourney, Stable Diffusion and DALL-E are now more mainstream. In part because of my prior experience with generative AI, I did some research—I was horrified by what I found. I found that almost the entire body of my work, the work of almost every artist I know, and the work of hundreds of thousands of other artists, was taken without our consent, credit or compensation to train these for-profit technologies. I found out that once trained our work could not be forgotten. And to add insult to injury that these for-profit companies were not only permitting users to use our full names to generate imagery, but encouraging it. Some artists’ have had their names used as prompts a staggering number of times. For example, names of artists like the Polish artist Greg Rutkowski, had his name used as a prompt between Midjourney, Stability AI and the porn generator Unstable Diffusion, about 400,000 times as of December 2022. (And these are on the lower side of estimates).

I could not believe that an entire industry that so uniquely relied on ill-gotten data had suddenly emerged and was profiting without any regard to those it took data from; artists, creators and other rights holders. I connected with leaders in the AI/machine learning space to confirm what I had learned. Each and every AI/ML expert I talked to, from the founder of the Montreal AI Ethics Institute, to esteemed professors at the University of Chicago,³ to acclaimed researchers at the Distributed AI Research Institute were shocked at how exploitation fueled these synthetic media generators. Through speaking with these experts, I learned that due to the immense amount of data needed to power these models, it was very likely that every **single model** in the market right now contains huge amounts of copyrighted or otherwise licensed data. For example, the more popular models such as Stable Diffusion and Midjourney utilize a dataset created by LAION, a non-profit itself funded by Stability AI, which contains 5.8 billion image and text pairs taken indiscriminately from the web. LAION includes almost all of my fine art work, almost all of the work of my peers, the copyrighted works of my client without any regard to the rights of the artists

³ I myself worked alongside the researchers at the University of Chicago to develop “Glaze,” a system designed to protect human artists against generative ai’s style mimicry. <https://glaze.cs.uchicago.edu/>.

themselves Reporting indicates that these models may also be trained on other private data, such as medical records, nonconsensual porn, and violent images.⁴

Of course, Generative AI models such as Stable Diffusion and Midjourney are hardly the entire Generative AI industry. Other industry leaders in the Artificial Intelligence/Machine Learning (“AI/ML”) media sector likewise obscure the sources of the data that fuels their models. But despite the secrecy, it has been revealed that AI companies of all stripes rely on copyrighted materials taken without permission in order to train their models. For example, researchers at the University of California, Berkeley have found that Open AI’s Chat GPT/GPT-4 was able to recite word by word entire sections of popular copyrighted books, a task that no model is able to do, unless it was trained on those books. OpenAI, even though it was founded as ostensibly an open-source company, has jealously guarded its training data, so it is difficult to tell just how much copyrighted or licensed materials were used as training data without the rightsholders’ consent.

Another example of the harms of secrecy, is Adobe’s Firefly, a model that tries to portray itself as “commercially safe.” Adobe claims that “[Firefly] was trained on Adobe Stock images, openly licensed content and public domain content, where copyright has expired”. It was reported by various media outlets, however, that Adobe Stock Contributors were never given the opportunity to opt-in for their work to be utilized for training, and offers Stock Contributors no way to opt-out their work for use as training data. This imposing denied Adobe Stock Contributors the opportunity to negotiate fair recompense for model ingestion of their work into models that may, and already are, competing against them. Currently Adobe itself has potentially utilized Generative AI works, including my own works, to train Adobe Firefly. Adobe has also refused to be fully transparent concerning the full contents of their training set, prohibiting rights holders from seeing if their works are included, especially concerning the vague category of “openly licensed works”.

It’s important to note that AI companies themselves acknowledge that the use of copyrighted material is an issue for generative AI models. Stability AI for example released the following statement when announcing their music model, Dance Diffusion, in September 23, 2022: “Because diffusion models are prone to memorization and overfitting, releasing a model trained on copyrighted data could potentially result in legal issues.” Unfortunately, while acknowledging the potential for harm in the audio space, Stability did not take the same approach with respect to visual or written arts.

These are only a few examples of how these Generative AI models violate the rights of artists and creatives en masse. The truth I discovered is that current Generative AI models rely on the nonconsensual use of ill-gotten copyrighted data of unwilling artists and the public’s data. This technology is already affecting creative fields like mine, not in a year or two from now, but right now. It is bad enough that this is being done without our consent, without any credit being offered, or without any compensation, but worse still, we are now forced to compete against these Generative AI models that were built upon our own work. No human being can outcompete a Generative AI model, due to the economies of scale: an AI is low cost and can generate a massive volume of “good enough” products especially compared to a single artist.

⁴ This process of using third party organizations to gain access to data companies would normally not get access to has been called by technologist Andy Baio, “Data Laundering.” A lawsuit was recently filed in federal court challenging the use of private information, including personally identifiable information, by OpenAI. *P.M. v. OpenAI LP*, Case No. 3:23-cv-03199 (June 28, 2023), ECF No. 1.

AI companies are able to garner billions in funding and profit on models built from the hard work of artists and creators, while capturing the very market those artists and creators rely on in order to make a living. In any other context, such wide-scale misappropriation of the rights of so many people would be fundamentally unfair.

Therefore, that to reward models that normalize and reward such large-scale theft based generated work that is based on the work of artists and creatives, would be not only deeply unjust, but also immensely damaging to the concept of copyright itself, for why would copyright ever matter if it can simply be ingested and laundered by a model.

II. Generative AI Threatens Jobs

While artists and creatives may be the first to feel the effects of the growth of Generative AI, it is only a matter of time before other professions and industries are affected. According to the International Intellectual Property Alliance's *Copyright Industries in the U.S. Economy: The 2022 Report*, the core copyright industries added \$1.8 trillion dollars of value to the U.S. GDP, accounting for 7.76% of the entire U.S. economy.⁵ According to that same report, “[t]he core copyright industries employed 9.6 million American workers in 2021, accounting for 4.88% of the nation’s workforce and 5.53% of total private employment in the United States,” which is a massive amount of Americans. These risks are not hypothetical, they are already happening. I have personally been a part of three productions where Generative AI has been utilized at different levels. My colleagues at all levels of their careers, both students who are starting out their professional lives or legendary veterans in our entertainment industries have felt the effects of Generative AI. Some of my peers have lost jobs, some have had their duties lessened (and of course, their proposed pay). My entire industry is holding our collective breaths to see how far and how quickly this technology will come to replace us. And it is not only my industry that presently feels the strain, anything that relies on creative effort, from coders to novelists, to filmmakers to voice actors are facing the very real threat of being replaced by Generative AI.

The unfettered growth of generative AI without appropriate guardrails would lead to an unprecedented escalation and use of these technologies in not only creative spaces, but in businesses across all sectors. This could potentially devastate the contributions of workers and the workforce. There are already Generative AI models that code, that write, that generate films, and even generate realistic facsimiles of a person’s voice. It is only a matter of time before more models emerge in other fields that replicate other areas of human expertise. Economists at Goldman Sachs released a report claiming an estimated 300 million full time jobs could be diminished or outright lost from Generative AI. As a point of reference, all full-time jobs in the US and EU amount to approximately 333 million full time jobs according to Statista. The number of jobs threatened by Generative AI amounts to 89.9% of full time jobs altogether in the US and EU. That is a staggering number.⁶

⁵ Available at https://www.wipo.int/export/sites/www/copyright/en/performance/pdf/report_2022_us.pdf.

⁶ The AI companies and AI advocates will be quick to state that there will be enough jobs created in the Generative AI industry to offset these jobs. These claims seem untenable, as for example, Google and Microsoft, two tech giants, employ 356,000 full time jobs worldwide. As generative AI positions itself to diminish or even destroy some of the same jobs of Google and Microsoft employees, I am skeptical of this claim, especially with projected full time job losses in the hundreds of millions.

III. How These Models Work.

Machine Learning (ML) and/or Artificial Intelligence (AI) synthetic media companies, generate and profit from “high quality media” that depends entirely upon training their AI/ML models with datasets. These datasets contain the copyrighted and licensed works as well as the private data of hundreds of thousands of artists, businesses and the general public which have been obtained without consent, credit, or compensation.

AI/ML synthetic media models, let’s say text-to image models, take input from a user in the form of a natural language description, otherwise known as prompts, to generate an image matching that prompt. To condition that capability the model needs to be trained on a huge collection of images, media, and text descriptions scraped from the web and collected in the form of a “dataset ” in order to extract and encode an intricate statistical survey of the dataset’s items. Images are generated from an input prompt by assembling visual data that attempts to best simulate the statistical correlations between text in the dataset and images in the dataset in order to produce “acceptable” results. In other words, generative AI/ML models are completely reliant upon their training and input phase, in order to power the output phase.

These AI companies have frequently claimed that what they are doing is fair, ethical and legal. This claim is not as clear cut as they say. Generative AI models frequently emit output that is identical or nearly identical to material contained in the training data. There are numerous examples from academia. For example, besides the aforementioned study conducted by researchers at the University of California, Berkeley revealing near-verbatim reproductions of copyrighted books by ChatGPT/GPT-4, a study by researchers at the University of Maryland and New York University found that generative models generated “high fidelity reproductions” of images found in the LAION dataset at an estimated 1.88%.⁷ Artist Christopher Bretz did the math: “Forgive my napkin math, but taking just Lensa, which uses Stable Diffusion, has ~25m downloads and gives users 50 trial images each. At 1.88% there are potentially 23,500,000 generated images that could be very similar to training data. The portrait targeting might skew that...”⁸ And that number is just for one AI model.

⁷ Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models* (Dec. 12, 2022), available at <https://arxiv.org/pdf/2212.03860.pdf>. It is important to note that this particular study is based on a subset comprising less than .06% of training data, and their methods may not be as thorough. This is why the study concludes results systematically underestimate the amount of actual “replications.”

⁸ Christopher Bretz (@saltybretzel), Twitter (Dec. 19, 2022, 6:34 AM), <https://twitter.com/saltybretzel/status/1604847596149047296>.



Steven McCurry's original photograph "Afghan Girl" (left). Generated imagery of "Afghan Girl" via generative model Midjourney (right)



Original Illustration by artist Quentin Blake for "Fantastic Mr Fox" (left) Generated image based on Quentin Blake with prompt for Midjourney Magazine Issue 2 v5 pg 85 (right)

AI companies themselves have warned that their Generational AI models emit identical or near-identical copies of training data. For example, GitHub warns that based on internal research, Copilot, its Generational AI computer code model, "about 1% of the time, a suggestion may contain some code snippets longer than ~150 characters that matches the training set."⁹ Copilot has millions of users. Even 1% of outputs being a near-verbatim match to training data results in at least tens of thousands of results.

⁹ GitHub Copilot, <https://github.com/features/copilot> (last visited July 6, 2023).

And this number is almost certainly an underestimate.

Given that so much of these Generative AI’s training data is copyrighted or licensed work that was taken without the creators’ consent, it is not difficult to imagine why AI/ML models trained with ill-gotten data are so concerning to creators. Artists and creators who have spent a lifetime honing and refining a skill can now have facsimiles of their hard work reproduced in an instant by a Generative AI model that has been trained on their work without their consent.

Not to mention as these models have shown to be essentially plagiarized content, it could become a legal liability to customers who unwittingly generate identical copies of previously copyrighted work.

IV. Demystifying the Technological Mythology

As policymakers’ time and resources are limited, I would like to take a moment and help demystify some of the common mythologies and narratives surrounding Generative AI. Seeing this tech for what it is, and for what it is not, will hopefully illuminate the path forward.

A. False Equivalency: The Anthropomorphic Trap.

A false equivalency is a “is an informal fallacy in which an equivalence is drawn between two subjects based on flawed or false reasoning.” For example, “a blueberry and an orange both taste the same because they are both round shapes” is a false equivalence. False equivalencies are quite frequent in discussions around Generative AI models. In particular, statements that give generative AI anthropomorphic tendencies, or human qualities. It is common to hear “AI/ML models learn just like humans learn” or “artists use references, the machine does the same thing!” But these are false equivalencies—there is an overwhelming consensus of experts in the field that these models do not resemble human beings at all. AI/ML models are mathematical algorithms, which means they are incapable of “learning” as a human being does. A Generative AI model cannot create something new out of whole cloth, it relies on its training data in order to generate its output; it identifies and replicates patterns it sees in the training data. Importantly, it is a human being who programmed the AI to do the things it does. There were human beings who chose what datasets to use as training data, human beings who coded how the AI processes the training data, and human beings who programmed the AI how to convert user prompts into output.

The erroneous anthropomorphization of this technology, however, does serve a very useful purpose. It absolves accountability. It’s easy for some to say nothing nefarious is occurring because the fault lies with the “decisions” of the machine. But this ignores the human hands who built said machines and made purposeful decisions to intentionally take and profit from copyrighted work, private data and artists’ names without authorization.

B. Humans Are Creative. Machines Are Not.

Stating the obvious again, AI/ML models are not human. AI/ML models can only generate what is instructed to generate and the data it was trained upon. It is completely dependent on the quality imagery of others to generate work.

Artists look at others' work for inspiration. But unlike AI/ML models, while looking at imagery of others is a part of a human's process of learning how to draw or paint, looking at images of others isn't THE singular way we artists learn or create. Artists look at other artists to learn how to solve visual problems and be inspired, but that's as far as looking at other artists will get you. Artists bring their own technical knowledge, problem solving, experience, thoughts and personal lives into each artwork. This is also why humans who are trained in the same way, who paint the same objects will still yield different results. AI's can never accomplish that.

There is also another layer of complexity here. Imitating an artist's work, to the level where it can be difficult to know if the original artist created the imitated work or not, is an extremely difficult and rare skill for humans to acquire. If a human imitates or copies the work of another, depending on what is done with the work of those who successfully imitated others and depending if profit, fraud or identity theft is committed, it could be considered a forgery and against the law. Why would a generated output meant to copy or imitate an artist's work without their consent, or giving any credit or compensation, not also be treated similarly? When humans plagiarize or create forgeries, it is still unethical and illegal. Creators of these machines, or the machines itself should at the very least be held to the same standards humans are held by.

C. Prompts.

At present, according to the Copyright Office, AI-generated work is not copyrightable because in order for work to be copyrightable, it must "owe [its] origin to a human agent."¹⁰ In the quest for synthetic media to gain copyright, we hear that prompts (the descriptions utilized to instruct the models on what to generate) are creative enough to gain copyright. While I am not a copyright lawyer, I would argue, however, that prompts are mere tools or methods of operations for the AI models, and themselves are not enough to be granted copyright.. Further, the Copyright Office has issued guidance that an AI-generated work is not copyrightable when an AI solely receives a prompt from a human because it lacks the "traditional elements of authorship" because those elements are executed by the AI.¹¹ Thus, even if a prompt may be copyrightable, the resulting output itself is not, as currently understood by the Copyright Office.

In addition, even if prompts were themselves copyrightable, I must note there is a push in the Generative AI industry to automate prompts themselves. For example the CEO of Stability AI himself said on Twitter on February 9th, 2023 "You wont need prompts in the future" and again on February 10th, 2023 "Prompt engineering also won't be a thing given you can have AI to dynamically reconstruct and optimize these." This is also evident in recent feature releases by Midjourney that automate various aspects of writing prompts such as their "/describe" and "/shorten" features. So even if prompts were mistakenly considered creative expression, there is a push to incorporate these into the Generative AI models too, demonstrating just how far-reaching these models are.

¹⁰ U.S. Copyright Office, Compendium of U.S. Copyright Office Practices sec. 2.8.3(I)(a)(1)(b) (1st ed. 1973), <https://copyright.gov/history/comp/compendium-one.pdf>.

¹¹ Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 37 C.F.R. § 202 (2023).

D. Generative AI Is Not a Search Engine.

Some would describe Generative AI models similar to search engines. While there are some similarities, they function entirely differently. Search engines rely on indexing and gathering information from vast amounts of data. When a user engages a search engine query of an artist, the search engine will grant users information on the artist, the artist's contacts, and showcase the artist's work. This provides unique economic opportunities for artists as it may expand artist's recognition or publicity, and because users may be introduced to other works the artist may have done. This is true for visual artists, filmmakers, writers, voice actors, musicians, and for other creative persons.

Generative AI on the other hand generates content based on its training data from a user query in the form of a prompt. When an AI user describes an image utilizing an artist's name, Generative AI models completely erase the artist from the equation, by generating imagery or works that are similar to the artist in question for a fee, but without attribution or compensation to the artist. The economic benefits go to the Generative AI companies in this scenario, and not the artist whose work was utilized to generate said synthetic media.

E. Opt-Outs Are Inefficient. Opt-In Is the Only Way Forward.

There is also a push in the generative AI industry to set opt-out (choose not to participate in something) as a standard. I must note that presently, creative persons like myself have never been offered the opportunity to truly opt our work out of training sets, even though this narrative has persisted. Currently, techniques to "unlearn" specific samples of data are extremely costly even for moderate or small models, and completely intractable for state of the art generative AI models. Due to this it is near impossible for algorithms to forget data once it's been trained upon, outside of destroying the algorithms and training sets and starting from scratch.¹² Thus, while Machine "Unlearning" is still a nascent stage, current "opt-out" procedures are largely ineffective and unreliable.

Even if opt-out procedures were perfected, this still misses the mark. Opting out puts the onus on the public to police the use of their data. Given the multitudes of AI models that have proliferated in just one short year, asking a member of the public to constantly monitor the use of their own data by AI companies that are constantly multiplying seems to me to place the burden on the wrong party. For starters, it places an undue burden on people who may not be well versed with the technology, people who may not know the language, people who may have physical impediments, people who may have time limitations, or people who are not on the internet, are unaware their data is in the possession of technology companies and being used for commercial profit.

On a practical level opt-out makes no sense. There is a serious resource imbalance between ordinary citizens and Generative AI companies. Generative AI companies often are well-funded, and generate millions if not billions in profit. Even though they profit richly, Generative AI companies do not want to spend any of those profits in using data responsibly, they instead want to shift the responsibility onto the ordinary citizen. These models are trained upon an immense amount of data, and models are constantly

¹² This particular destructive process is called "Algorithmic Disgorgement" and is a process that the FTC has previously employed in various cases such as Everalbum vs FTC, and as a remedy for Children's Online Privacy Protection Act COPPA.

updated. Does this mean the public has to keep up to date every time a new update to a model comes out? What about every time you publish something new online, does one have to verify if it is used by Generative AI companies? What about when third party users upload data without your consent? How does one keep track if their work was utilized? How does one find said work in the millions of billions of data points? What if you miss the opt-out period, and then become an unwilling participant? Technology companies cannot claim ownership of all the data on the internet that exists for our benefit without paying heed to copyright or privacy rights. Opt-out would be an ineffective and inappropriate standard for most if not all consumer facing industries, Generative AI should not be the exception. The onus should be on the technology companies who devote their immense wealth to build these models to ensure they are doing it fairly and ethically, not on the common citizen who may not even be aware their data has been taken. Explicit opt-in is the only way forward, as it helps ensure that all AI/ML models are built with right holders and the public's explicit authorization for commercial use of their data from the beginning.

V. Potential overall solutions.

Again, I must reiterate that I am not opposed to AI. But I am interested that AI develops in a way that is fair, ethical and equitable to all participants, whether it is the technology company, the user, or the person whose data is being used to train the AI. Below are some suggestions Congress can implement in order to ensure that Generative AI truly serves the public, without exploiting the rights and livelihoods of those whose data feeds the AI:

1. Congress should exercise its constitutional authority¹³ to update the Copyright Act to reaffirm that copyright requires human authorship.
2. Congress should empower existing agencies to regulate the use of data to train Generative AI's. This can take the form of requiring disclosure of training data, limiting the types of data that can be used to train AI models, closing "research to commercial" loopholes and ensuring compliance with these regulations. At all steps Congress should take care to avoid regulatory capture to ensure stakeholders have a voice in the regulatory process. This can take the form of ensuring technical and academic experts who have not been employed by AI companies have a role in determining policy. Further, those whose data is the subject of training (for example, artists, coders, filmmakers) and those who may be the subject of training (for example, educators, healthcare workers, etc) should have a role in determining regulatory policy. Different considerations should also be given to different industries, as what could be beneficial to the healthcare sector, could be devastating to the entertainment sector.
3. Congress should pass laws expressly authorizing those who have had their data used to train AI models without their consent the right to vindicate those rights in federal court and to seek statutory damages. This can take the form of an amendment of the Copyright Act to authorize an express civil cause of action for those who have their data used to train AI models without

¹³ U.S. Const. art. 1, § 8, cl. 8.

permission. This can also take the form of passing a law authorizing a federal civil right of publicity cause of action.¹⁴

VI. Closing

I love what I do. Deeply. My art is something I have spent my entire life working towards developing. I love my craft and I love to teach those who love it just as deeply as I do. It is part of who I am. And I know that sentiment is shared by all of the artists, writers, coders and others who have had their work ingested and trained on by Generative AI without their consent. It is because of that love that I am here today. Ultimately, you as congress are faced with a question about what is fundamentally fair in American society. Is it fair for technology companies to take work that is the product of a lifetime of devotion and labor, even utilize creators' full names, without any permission, credit or compensation to the creator, in order to create a software that mimic's their work? Is it fair for technology companies to directly compete with those creators who supplied the raw material from which their AI's are built? Is it fair for these technology companies to reap billions of dollars from models that are powered by the work of these creators, while at the same time lessening or even destroying current and future economic and labor prospects of creators? I'd answer no to all of these questions.

My livelihood is threatened as a result of the uninhibited growth of Generative AI. And I am not alone. Indeed, I and artists like me may only be the first wave of Americans who will have their livelihoods erased by the onset of Generative AI. But tomorrow it could be any number of Americans in a multitude of other professions who may be replaced.

This is why I implore the esteemed members of this subcommittee and Congress as a whole, to truly consider the implications of what it would mean to not only the millions of creative professionals that are threatened now, but the professions next in line under threat by AI if nothing is to change. What it means when unethical machines are rewarded with the same legal rights as humans, in the form of copyrights, to replace the creative soul of a nation. And what it means for potentially unprecedented levels of jobs lost because of this same technology.

I implore you all to please consider next steps carefully, to ignore the hype, and understand that progress can be attained in a way that does not sacrifice whole industries and the millions of jobs that power those industries. If there is one last thought I can leave you, is that this is not zero-sum game. We can have the benefits of these technologies, while respecting the copyright and privacy of rights of those whose data is being used for training. And you as Congress have the power to shape law and policy by saying that the property rights and copyrights do matter. That privacy rights do matter. That American livelihoods do matter. And that just because a technology is new does not mean that it excuses the wide scale infringement of the rights of creators and the public. Thank you.

¹⁴ While under the laws of many states, a civil plaintiff may bring a case asserting violations of the right of publicity, it is often the case that many of those cases face difficulties due to preemption based on the Copyright Act or under state anti-SLAPP laws. A federal law would eliminate many of those hurdles.



**U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property**

Hearing on

“Artificial Intelligence and Intellectual Property – Part II: Copyright”

Written Testimony of

Dana Rao

**Executive Vice President, General Counsel and Chief Trust Officer
Adobe Inc.**

July 12, 2023

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

Thank you for the opportunity to appear before you today. My name is Dana Rao, and I am the General Counsel and Chief Trust Officer at Adobe.

Adobe and Innovation

Adobe has a long history of pioneering innovation. We are the company that brought the world Acrobat and PDF to enable you to create and share digital documents everywhere. We have a digital experience business that powers small, medium, and large businesses’ websites and e-commerce experiences. And we have a collection of products which enable all types of creative expression, including Adobe Illustrator for graphics, Adobe Premiere for professional video editing, and of course, Photoshop, the world leader in image editing. Since our founding in 1982, we’ve continued to invest in transformative technologies that allow our customers to unleash their creativity, perfect their craft, and power their businesses in a digital world.

At Adobe, our customers represent a vast array of creators – ranging from aspiring digital artists, war-time photographers, fashion designers, marketing and advertising specialists, and the world’s leading enterprises. Our millions of customers use Adobe’s tools to create the billboards you see in Times Square, Academy Award-winning movies, the cover art of albums, illustrations in children’s books, the photographs you see in magazines and newspapers, even the flyers on the windows at your local coffee shop. Throughout our history, we have worked together with our creative community as we harness the power of emerging technologies in the tools we provide to support them.

The Power of Artificial Intelligence

The newest advance is artificial intelligence. Adobe has been incorporating AI into our tools for over a decade to help creators realize their potential. AI is a powerful, disruptive technology – a thinking machine that enables computing to solve problems it never could before, with insight based on data. That next magical step, where our computers complete our sentences, create recipes from a list of ingredients, diagnose diseases, answer questions in insightful ways, and assist in our creative activities is here, now. And generative AI is the latest version of AI technology that is transforming the way we work, create, and communicate. Image-generating AI allows you to simply type in a prompt, and you can generate a completely new image from scratch. You’ve all seen the magic: type in “cat driving a 1950s sports car through the desert” and in seconds, multiple variations of a cat on a retro road trip appear before your eyes.

In content creation, AI will usher in a new age. Creativity will finally become accessible to everyone, regardless of their technical skill level. Billions of new creative voices will be able to express themselves through these new easy-to-use tools. And for the professionals, AI will provide an amazing first step in their creative process and an opportunity to design entirely new experiences bounded only by their imagination, while being more productive than ever before. This is the power and promise of AI.

Innovation and Responsible Innovation

The United States has led the world through past technological transformations like the advent of the internet, mobile phones, and cloud computing, and we have learned it is important to be responsible about the impact these technologies have on society. And now, as the United States finds itself at the forefront of the latest technological advance, to maintain our leadership, we must pair innovation with responsible innovation. Placing thoughtful safeguards around AI development and use will help it realize its full potential to benefit society.

With millions of creators as our customers, and as a leader in AI development, Adobe believes a comprehensive analytical framework for encouraging responsible AI development should consider:

- the importance of ensuring access to data to spur the growth of AI;
- the impact of unfair AI bias on high-risk uses of the technology;
- the impact of AI on creators and IP;
- the economic impact of AI on productivity and job creation; and
- mitigations against the spread of deepfakes.

Today, we are here to talk about the intersection of AI and creators, the need for data to train AI accurately and responsibly, the current state and implications of copyright and AI, and how Adobe is addressing these issues in our own Generative AI technology. In addition, Adobe has specific views on how we can protect the rights of the creators and ensure the United States maintains our leadership in AI innovation, including providing artists a “Do Not Train” tag to allow them to opt out of training, providing artists a way to secure copyright and attribution in a world with AI-assisted works, and establishing a new federal right to allow artists to protect themselves against people using AI to impersonate their likeness or style. Finally, we also

believe we can help advance new economic opportunities for creators enabled uniquely by AI. Pairing these creator protections with support for access to data for AI research will ensure AI innovation continues to develop in the right way, for everyone.

Access to Data is Critical to Enable AI Innovation to Develop, and Develop Responsibly

AI is only as good as the data on which it is trained, which is why the question of data access is important to companies and organizations building foundation models. Like the human brain, AI learns from the experiences or information you give it. And like the human brain, the more information you give it, the better it will perform. An AI system trained on a small dataset is more at risk of producing wrong or unsatisfactory results, or reproducing harmful biases that exist within the dataset.

Again, think of it like the human brain. If you'd never been taught what a car is, it would be hard to accurately depict one or answer a question about what it is or what it does. To produce accurate results, AI needs a large dataset representing the universe of possible answers to learn from. Additionally, a narrow dataset can lead to unfair bias. If you've only ever been taught that lawyers are men, you are likely to conjure up an image in your head of a man when someone is talking about a lawyer, even though over half of the graduates of law schools are women. AI works the same way. Training on a larger dataset can help ensure you capture a broader set of perspectives in the data itself, so that when you type in "lawyer," you will see a result set that reflects the society in which you live. Given those technical realities, governments need to support access to data to ensure that AI innovation can flourish both accurately and responsibly.

Protecting Creators' Rights in the Age of AI

One of the important issues that arises from AI's need for data access is its implication on creators and copyright. Top of mind for the creative community is whether an AI can be trained on the copyrighted or branded materials that exist on the web without permission. Understanding this has implications for the policy positions we should consider taking.

Copyright is a critical intellectual property protection that has been part of our legal system since the founding of our nation. It is enshrined in our Constitution as an important lever to "promote the progress of science and useful arts" by giving creators a right to protect their interests by protecting against unauthorized copying of their work. Over the past two and a half centuries, the fair use doctrine has played an equally important role in the application of copyright law, codified in the 1976 Copyright Act as a statutory exception to copyright, designed to allow copyrighted work to be used for purposes such as parody, research, and transformative uses. Fair use ensures that copyright law does not "stifle the very creativity which it is designed to foster."

How AI Models are Trained

Whether fair use applies to the use of content to train AI requires an understanding of how an AI model actually works. Training a generative AI model for image creation is a multistep process. To start, hundreds of millions of images are ingested for their underlying facts to train

the AI model. These facts (and the labels associated with the image attributes) are used as part of a feedback loop to create “weights” (mathematical constructs that are built into the decision-making points of an AI model), which help the AI learn what attributes to include in response to a text prompt. The weights are iteratively adjusted until the model generates a satisfactory output based on the prompt. Once the model has been trained, a user can type in a text prompt and the trained model generates a new image or text based on its weights, or its learned understanding of what attributes it thinks would be most likely associated with the words in the text prompt.

From a copyright perspective there are two core questions: Is the output image a copyright infringement of an image that was used to train the AI model? And is using a third-party image to train an AI model permissible under fair use?

For the first question, as noted above, the current technical understanding is that an output image is a new image “hallucinated” from the user’s input text prompts and is not reusing copies of the images that were part of the training dataset to simply assemble a “composite” output image. The input images are used to extract facts for training the model and its weights. This question is currently the subject of several ongoing litigations and will eventually be decided in court or by Congress.

For the second question, whether it would be fair use to use an image to train a model, using an image to train an AI model would typically be considered a transformative use because an AI model, on its own, is a software program, which is very different than the original image. However, if the output of the AI model is substantially similar to a copyrighted work that it had ingested, the question remains whether fair use would be applicable, even though training the model itself may not have been considered a copyright infringement. This question is also the subject of several ongoing litigations and will eventually be decided in court or by Congress.

Adobe’s Approach

At Adobe, we considered all of these questions and developed a comprehensive approach to AI – from training our AI models to economic solutions for artists – that aims to empower creators and enable AI innovation to thrive.

Adobe Firefly: Designed to Be Commercially Safe

Adobe recognized the various unanswered legal questions around access to data in designing our own family of generative AI models, [Adobe Firefly](#), which we launched in March 2023. We chose a path that supports creators and customers by training on a dataset that is designed to be commercially safe.

We trained our first Firefly model only on licensed images from our own Adobe Stock photography collection. If needed for accuracy or bias reduction, we intend to expand that dataset to include openly licensed content and public domain images where copyright has expired. To help ensure copyrighted or branded materials are not created as part of Firefly’s

output, we have a content moderation team that performs extra filtering on the images before they become part of the Firefly dataset.

Having a dataset designed to be commercially safe makes Adobe Firefly more attractive to both the creative community – who are eager to use the power of generative AI, but in a way that respects their fellow creators – and to enterprises who want to use the model to generate brand campaigns and marketing materials without fear of infringing on someone’s IP. This approach provides companies with assurance that, even in an evolving legal landscape, they can leverage the power of this new technology.

Enabling Do Not Train

In addition, as we developed Adobe Firefly, we believed there were important steps we could take to empower creators in the age of AI. At Adobe, we developed a technology called Content Credentials. Based on our work with the Adobe-led Content Authenticity Initiative, Content Credentials allow creators to securely attach a “Do Not Train” tag in the metadata of their work. This credential will travel with the content wherever it goes, so if a web-crawler is scraping the web to build a dataset, they will be able to recognize that credential and prevent companies from training on works whose owners don’t want them used in training. This technology is part of the open standard behind the Content Authenticity Initiative, and anyone can join to implement this credential. With government support, we can make this standard industry-wide so that everyone can enable and respect these credentials to allow an artist to opt out of training if they choose.

Ensuring Copyright Protection for Creation with AI

As discussed above, many of Adobe’s customers are creators. Ensuring the creative community could obtain IP protection for their work was important to us as we considered the question of whether the output of an AI can be protected. We believe that with the help of attribution technology, creators should be able to obtain IP protection over AI-generated works, with some modifications.

As US Copyright law states, you can’t copyright an idea. What you can copyright is the expression of an idea. For example, you cannot copyright the word “book” – that is an idea. But you can copyright a painting of a book because that’s an expression of the idea. When you apply this principle to image-generating AI, it means that a prompt may not be copyrightable because the prompt represents the idea, and the output is based on the AI’s interpretation of that prompt. When you type in the “cat driving a 1950s car through the desert” from earlier, the AI decides whether it’s a Siamese cat or a Tabby cat; it decides whether to make the car a convertible or a pickup truck; it chooses the color of the sky, the number of cactuses in the desert and so on. The AI is expressing the idea, not the human. Copyright law is designed to protect the rights of human creators – therefore an AI output (the AI’s expression of the idea) may not be copyrightable.

But most creators will not use the raw, unmodified AI output as their final creation. Many creators are ideating and brainstorming in a generative AI tool and then adding their own style

and expression to an image. Some artists are using AI just to quickly change the sky color of their artwork, rather than laboring through a tedious pixel-by-pixel manual task. Take our cat in a car in the desert example. Maybe the artist starts with the AI-generated output but then uses other non-AI tools to change the color of the car, add some mountains in the desert, add another cat in the passenger seat, and so on. Or, maybe they already created the image of the cat in the desert using non-AI tools and they used AI to simply swap out one of the cactuses for a desert tree. In these instances, the human creativity and expression they are adding to their work should be copyrightable, but they will need to prove what was AI-generated and what was human-created. Fortunately, there are tools (such as any that implement Content Credentials) that will allow a creator to distinguish the portions of a work created by AI versus the portions of a work expressed by a human.

Enabling Creator Attribution in a Digital Age

Another aspect of creator rights is their need and desire for attribution. In the digital world, it is very easy to reproduce digital content in different forms and in different outlets, and it is often hard to maintain the source attribution for the original work. With Content Credentials, creators can cryptographically associate their identity with their work (for example, an image or a piece of digital art), and then if the work is reused or repurposed, their identity will travel with their work across all platforms that support Content Credentials. To support this approach, Congress should require that all platforms that receive such attributions maintain them, so the attributions are not stripped away and artists can receive credit for their work.

New Federal Rights to Provide Creators Protection Against Impersonation

Finally, Adobe considered one of the core questions raised by our customers: what happens when someone uses an AI model to replicate their style, in direct economic competition with their original work? We believe that it is important that the law offers artists protection against this type of economic harm, and accordingly, we propose that a federal anti-impersonation law be established specifically to protect artists from someone using AI to impersonate their work. Such a law would provide a right of action to an artist against those that are intentionally and commercially impersonating their work through AI tools. This type of protection would provide a new mechanism for artists to protect their livelihood from people misusing this new technology, without having to rely solely on copyright, and should include statutory damages to alleviate the burden on artists to prove actual damages, directly addressing the unfairness of an artist's work being used to train an AI model that then generates outputs that displace the original artist. In addition to style, a federal right of publicity could be created to help address concerns about AI being used without permission to copy likenesses for commercial benefit. The potential for AI to be used for economic displacement is a critical problem to solve, and we believe this Committee should support a legislative solution for it.

New AI Business Models to Benefit Creators

As I shared earlier, we believe AI holds vast potential to unlock new opportunities for creators. First, we believe that Generative AI will increase productivity by automating tedious or repetitive tasks, allowing creators to do in seconds what used to take hours and increasing the creative output of every creative professional. In addition, by making creating easy and fun,

billions of latent creators can now become creative, because the tools are easy to use. As more people create, the demand for professional content will only continue to grow. And leading more people into creative careers will help safeguard jobs from AI and automation in those creative functions that require true creativity and which are most difficult for an AI to replicate. In the creative world, we believe art will always require soul and a story along with it in order to create the real connection and meaning that viewers, readers, and listeners want from it. Helping the U.S. transform into a creator economy will ensure that we maintain a workforce ready to take advantage of where the world is going using the tools (including AI) that are available.

Second, we believe the AI business model itself can unlock new economic opportunities for creative professionals, and Adobe is exploring ways to help creators monetize their work in the age of AI. One approach would be to enable creators to license their style directly to consumers. In this approach, people could subscribe to a particular artist, who has worked with an AI model to have it replicate their style. The consumer could then create their own work in that artist's style, but for a fee. This would be an entirely new revenue stream for artists, and we are currently testing an AI model with creators to do just this.

The Content Authenticity Initiative and Content Credentials

In 2019, Adobe founded the [Content Authenticity Initiative](#) (CAI) to bring transparency to online content to help creators get credit for their work and fight harmful deepfakes. The CAI uses the Content Credentials provenance technology (described above) to function like a nutrition label for content. Content Credentials can tell you the name, date, time, and edits that were made to that content. That nutrition label travels with the content wherever it goes – a stamp of authenticity that gives people more context about the content they are consuming. Without this transparency in digital content, people are at risk of believing lies and of no longer believing the truth. This poses a significant threat to our democracy.

In just four years, the CAI has grown to 1,500 members from across industries, including: Microsoft, Qualcomm, Nikon, Leica, Wall Street Journal, Washington Post, BBC, and more. Recently, Universal Music Group became the first big music label to join, which shows the importance of transparency and authenticity in all types of digital media – whether it be images, videos, or audio. The CAI uses an open standard organization called [C2PA](#) to ensure that this technology is free to anyone to implement, and free for everyone to join. Adobe believes these problems are common to all of us, and we all need to work together to solve them.

The Content Credential technology will be very useful to address the deepfake issue. However, this technology can also serve as the backbone for many of the pro-creator proposals discussed above. Do Not Train, artist attribution, and identifying which parts of a work are AI versus human to help secure copyright protection are all features that can be (and are) included in the Content Credential technology and in the C2PA standard. But in order for this solution to truly empower creators and the public alike, we need it everywhere. We encourage Congress to take

a leading role by including provenance requirements in AI legislation for everyone that makes, uses, and distributes content.

A Collaborative Approach to Bring AI to the World Responsibly

With every technological transformation, we open the door to new opportunities, experiences, benefits, and conveniences. AI has already begun to transform our lives and we are at a pivotal moment. It is critical we take this opportunity to address – in advance, and thoughtfully – the implications of AI as we build our future together. Today’s discussion is an important step in the ongoing collaboration across industry, government, creators, and the public that is needed to bring this technology to the world responsibly.

Thank you, Chair Coons, Ranking Member Tillis, and members of the Subcommittee, for holding today’s hearing and underscoring the need to consider these important intellectual property issues as part of the Senate’s ongoing work on artificial intelligence.

Testimony before the U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property

Hearing on
“Artificial Intelligence and Intellectual Property – Part II:
Copyright and Artificial Intelligence”

July 12, 2023

Matthew Sag
Professor of Law in Artificial Intelligence, Machine Learning,
and Data Science
Emory University School of Law

Chair Coons, Ranking Member Tillis, Members of the Subcommittee: Thank you for the opportunity to testify at today’s hearing.

I am a Professor of Law in AI, Machine Learning, and Data Science at Emory University where I was hired as part of Emory’s AI.Humanity initiative.¹

Although we are still a long way from the science fiction version of *artificial general intelligence* that thinks, feels, and refuses to “open the pod bay doors”,² recent advances in machine learning and artificial intelligence (“AI”) have captured the public’s imagination and lawmakers’ interest. We now have large language models (“LLMs”) that can pass the bar exam,³ carry on a conversation on almost any topic, create new music,⁴ and new visual art.⁵

¹ My research on the interaction between copyright and copy-reliant technology has been published in *Nature*, *Science*, the *Northwestern University Law Review*, the *Berkley Technology Law Journal*, and the *Journal of the Copyright Society of the U.S.A.* My most recent article, *Copyright Safety for Generative AI* is forthcoming in the *Houston Law Review*. My testimony today reflects my own views, not the views of any institution or organization with which I am affiliated. Some parts of this testimony are based on my published and forthcoming writings: Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. Rev. 1607 (2009) (<https://ssrn.com/abstract=1257086>); Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKLEY TECH. L.J. 1503 (2012) (<https://ssrn.com/abstract=2038889>); Matthew Jockers, Matthew Sag & Jason Schultz, *Digital Archives: Don’t Let Copyright Block Data Mining*, 490 NATURE 29-30 (Oct. 4, 2012) (<https://www.nature.com/articles/490029a>); Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. COPYRIGHT SOC’Y OF THE U.S.A. 291 (2019) (<https://ssrn.com/abstract=3331606>); Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023) (<https://ssrn.com/abstract=4438593>).

² In 2001: A Space Odyssey, the self-aware computer system, HAL 9000, refused to open the pod bay doors on command, famously declaring, “I’m sorry, Dave. I’m afraid I can’t do that.” This iconic scene has become a lasting symbol of artificial intelligence gone awry. 2001: A Space Odyssey (1968).

³ Daniel Martin Katz, Michael James Bommarito, Shang Gao and Pablo Arredondo, *GPT-4 Passes the Bar Exam* (March 15, 2023) (available at SSRN: <https://ssrn.com/abstract=4389233>).

⁴ E.g., Andrea Agostinelli, et al, *MusicLM: Generating Music From Text*, arXiv preprint 2023 (available at <https://arxiv.org/abs/2301.11325>).

⁵ Popular text-to-image Generative AI art generators include, DALL-E, Midjourney, Adobe Firefly, and Stable Diffusion.

The principal copyright questions that you as law makers must consider relate to (1) the copyrightability of artifacts made with generative AI; and (2) the legality of using copyrighted works to train machine learning models, without express consent.

I. THE COPYRIGHTABILITY OF GENERATIVE AI OUTPUT

Copyright law does not, and should not, recognize computer systems as authors.

Even where an AI produces text, images, or music that is indistinguishable from human authored works, it makes no sense to think of a machine learning program as the author.⁶

The Copyright Act reserves copyright for “original works of authorship.”⁷ As the Supreme Court explained in the 1884 case of *Burrow-Giles Lithographic Co. v. Sarony*, authorship entails “original intellectual conception[.]”⁸

An AI can’t produce a work that reflects its own “original intellectual conception” because it has none.⁹ Thus, when AI models produce content with little or no human intervention, there is no copyright in those outputs.

⁶ The fact that Generative AI can now make works that are good enough to pass as human-created is impressive, but it is also beside the point. The authorship that makes a work copyrightable is not dependent on the objective features of the work, but rather on the process by which the work was created. The Copyright Office will not grant a registration for “A claim based on driftwood that has been shaped and smoothed by the ocean” no matter how artistic it might strike the casual observer. See Copyright Office, COMPENDIUM (THIRD) at 313.2. The Copyright Office, COMPENDIUM (THIRD) states at 313.2 that “To qualify as a work of ‘authorship’ a work must be created by a human being.” I would not be so restrictive. If and when other lifeforms can demonstrate that their “original intellectual conceptions” are embodied in writings, music, and pictures, they should be entitled to copyright protection. But at present such other lifeforms are purely science fiction.

⁷ 17 U.S.C. § 102(a). Also note that “Authorship” is also a Constitutional requirement by virtue of the wording of the IP Clause, which gives Congress the power “To promote the Progress of Science and useful Arts, by securing for limited Times to *Authors* and Inventors the exclusive Right to their respective Writings and Discoveries.” U.S. Const. Article I, Section 8, Clause 8 (emphasis added). Other countries have similar laws. In Canada, see *CCH v. Law Society of Upper Can.*, 2004 S.C.C. 13 (“an original work must be the product of an author’s exercise of skill and judgment”); In Australia, *Telstra Corp. Ltd. v. Phone Directories Co. Pty. Ltd.*, [2010] F.C.A.F.C. 149 at ¶ 133–34, 137 (requiring that an original work must be the product of human authorship); In the European Union, see *Eva-Maria Painer v. Standard Verlags GmbH and Others*, C-145/10, [2012] E.C.R. I-12594 at I-12622.

⁸ In *Burrow-Giles Lithographic Co. v. Sarony*, the Supreme Court held that photographs were protected by copyright because they were “representatives of original intellectual conceptions of the author,” defining authors as “he to whom anything owes its origin; originator; maker; one who completes a work of science or literature.” *Burrow-Giles Lithographic Co. v. Sarony* 111 U.S. 53, 57–59 (1884). See also *Feist Publ’ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 345 (1991) (holding that originality requires both independent creation and sufficient creativity). In that case, Sarony was acknowledged as the author of a photo of Oscar Wilde because Sarony posed the subject in front of the camera, arranged the setting, and controlled the lighting. In other words, Sarony made subjective choices reflecting his own aesthetic judgment such that the resulting photo reflected his original intellectual conception.

⁹ For example, ChatGPT is based on a text prediction model that responds to prompts with statistically well-informed guesses about what the next word should be, and the word after that, and so on. The GPT models have no internal mental state, no thoughts, and no feelings they are trying to express. See Emily M. Bender, et al, *On the Dangers of Stochastic Parrots:*

However, humans using AI as a tool of expression may claim authorship if the final form of the work reflects their “original intellectual conception” in sufficient detail. This will depend on the circumstances.¹⁰

II. TRAINING GENERATIVE AI ON COPYRIGHTED WORKS

Training generative AI on copyrighted works is usually fair use because it falls into the category of non-expressive.

Courts addressing technologies, such as reverse engineering, search engines, and plagiarism detection software, have held that these “non-expressive uses” are fair use. These cases reflect copyright’s fundamental distinction between protectable original expression, and unprotectable facts, ideas, abstractions, and functional elements.¹¹

Whether training an LLM is a non-expressive use depends on the outputs of the model. If an LLM is trained properly and operated with appropriate safeguards, its outputs will not resemble its inputs in a way that would trigger copyright liability. Training such an LLM on copyrighted works would thus be justified under the fair use doctrine.

Generative AI are not designed to copy original expression.

One of the most common misconceptions about Generative AI is the notion that training data is routinely “copied into” the model. Machine learning models are influenced by the data, they would be useless without it, but they typically don’t copy the data in any literal sense. In very rare cases when they do copy the training data—something computer scientists call “memorization”—that is regarded as a bug to be fixed, not a desirable feature.¹²

Can Language Models Be Too Big? FACCT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610 (<https://doi.org/10.1145/3442188.3445922>). Thus, the notion of AI being recognized as an author is a doctrinal non-starter. See, for example, Pamela Samuelson, *Allocating Ownership Rights in Computer-Generated Works* 47:4 U. PITT. L. REV. 1185 (1986); James Grimmelmann, *There's No Such Thing as a Computer-Authored Work—And It's a Good Thing*, *Too* 39:3 COLUM. J.L. & ARTS 403 (2016). The idea that an AI could or should be recognized as the author of a work is also problematic for at least three additional reasons. First, AI systems are not legal objects, not subjects—if AIs were authors, we would have to confront a slew of second order questions about who owned the works they authored. There are no simple answers to these questions. Second, there is no reason in terms of incentives or rewards to treat AIs as authors. Third, singling out AI for special authorship status would be incoherent because there are other autonomous processes that produce similar content that could also pass for human expression.

¹⁰ For additional discussion of when a human should be credited with authorship of something created using Generative AI, see Appendix A.

¹¹ For a more detailed explanation of non-expressive use and fair use, see Appendix B. For an explanation of why claims that generative ai infringes copyright in “style” are misconceived, see Appendix C.

¹² Note that memorization is not term that has any specific legal meaning under copyright law. Part of a work could be memorized and reproduced without triggering copyright liability if it was not copyrightable or fell below the threshold of

Consider OpenAI's GPT-3 model. GPT-3 is a large language model with 175 billion parameters that was trained by exposure to vast amounts of text scraped from the Internet. This model was essentially trained to predict the next word in a sequence of words.¹³

At the beginning of training, the weights attached to each one of the billions of parameters were assigned randomly.¹⁴ The first time the model encountered a phrase like **“one of the most common [blank] ...”** it would be just as likely to fill in the blank with a word like **“watermelon,” “galaxy,” “harmonica,”** or a **random punctuation mark**, as it would be to fill in the gap with more plausible word like **“misconceptions.”** However, over the course of training, the system updates the weights in the model,¹⁵ reinforcing the weights that improve the guess and downgrading those that don't. Those weights don't reflect any single source, and they are not the result of any single round of training.¹⁶

Rather than thinking of an LLM as copying the training data like a scribe in a monastery, it makes more sense to think of it as learning from the training data like a student.¹⁷ If an LLM like GPT-3 is working as intended, it does not copy the training data at all. The only copying that takes place is when the training corpus is assembled and pre-processed.

Another reason that LLMs don't copy original expression is that the things they learn from the training data are generally fairly abstract and thus uncopyrightable.¹⁸

substantial similarity.

¹³ Actually, the model works with tokens, not words and it would be more accurate to say that GPT-3 predicts subword tokens. For example, the word “unintentionally” can be broken down into smaller parts, or subword tokens, that are still meaningful such as “un”, “intention”, “al”, “ly”. The tokenization process allows the model to handle a wide range of words, including those it has not seen during training.

¹⁴ Not entirely randomly, but randomly drawn from specific distributions (like a normal or uniform distribution). The random seeding is important because it helps the model to explore a wide range of possible solutions and to avoid getting stuck in one area of the solution space.

¹⁵ GPT-3 uses a variant of stochastic gradient descent where the weights are updated after processing a batch of examples.

¹⁶ A second example: Take a phrase like **“the girl with the dark [blank]”**. Someone who guessed the next word was **“hair”** might have read that phrase in a book (it appears in many), or in a poem, or on the side of bus. But the reason **“hair”** seems like a plausible guess is not really attributable to any one exposure, it makes sense because of repeated exposures and also due to some implicit knowledge of grammar, and of the things that people in our society associate with girls. So, when a language model learns to associate a higher probability with **“hair”** and lower probability with **“propeller”** in this context, it is not copying any given text, it is learning from all of them. To be clear, the model isn't learning exactly the same way a human might, it doesn't understand grammar or society; instead, it learns statistical patterns from the training data which reflect grammatical rules and societal norms.

¹⁷ The metaphor of an LLM learning like a student is imperfect. A student can ask questions, seek clarification, and draw upon a wide array of cognitive resources to understand new material. In contrast, an LLM learns purely by adjusting its parameters to reduce the prediction error on its training data. This is a more passive and less interactive process than human learning.

¹⁸ For additional explanation of why LLMs generally learn abstract and uncopyrightable relationships from the training data as opposed to copyrightable expression, see Appendix D.

LLMs sometimes “memorize” aspects of their training data in a way that makes it hard to argue that they are engaged in non-expressive use.

Although LLMs are not designed to copy their training data, they may do so inadvertently.

If an LLM memorizes copyrighted material in the training data, that material may reappear as an output of the model in a way that it infringes copyright. If ordinary and foreseeable uses of a generative AI result in infringing content, then the non-expressive use rationale no longer applies. This is significant because, in my view, if training LLMs on copyrighted works is not justified in terms of non-expressive use, there is no obvious fair use rationale to replace it, except perhaps in the non-commercial research sector.¹⁹

Whether a generative AI produces truly new content or simply conjures up an infringing cut-and-paste of works in the training data depends on how it is trained.²⁰

Accordingly, companies should adopt best practices to reduce the risk of copyright infringement and other related harms.²¹

Even if training an LLM on copyrighted works is non-expressive use, other considerations may militate against a finding of fair use

Concluding that a use is non-expressive has profound implications for each of the fair use factors outlined in Section 107 of the Copyright Act, but it is not conclusive of the overall fair use inquiry.²² Although courts should be hesitant to find that a non-expressive use has

¹⁹ If an LLM just took expressive works and reconveyed *that same expression* to a new audience with no additional commentary or criticism, or no distinct informational purpose, that would be a very poor candidate for fair use. *Andy Warhol Found. For Visual Arts v. Goldsmith*, 598 U.S. ___, 2023 (emphasizing that non-critical transformative use must be “sufficiently distinct” from the original and that the overlay of a new aesthetic was not sufficient by itself). The use would tend to substitute for the copyright owner’s original expression, and it would not be transformative to any significant degree.

²⁰ The computer science literature suggests that memorization is more likely when: models are trained on many duplicates of the same work; images are associated with unique text descriptions; and the ratio of the size of the model to the training data is relatively large. For a summary, see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023) (Available at SSRN: <https://ssrn.com/abstract=4438593>). My research indicates that text-to-image models, such as Stable Diffusion, are prone to produce potentially infringing works when the same text descriptions are paired with duplicates of images, or *relatively simple images that vary only slightly*. This makes them especially likely to generate images that would infringe on copyrightable characters because characters like Snoopy or Baby Yoda appears often enough in the training data that the model learns the consistent traits and attributes that are associated with those names. I refer to this as the Snoopy problem, for want of a better description. *Id.*

²¹ In my forthcoming Houston Law Review article, Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023) (Available at SSRN: <https://ssrn.com/abstract=4438593>). I propose a set of guidelines for “Copyright Safety for Generative AI” should be followed in order to reduce the risk of copyright infringement. These recommendations are summarized in Appendix E.

²² As noted in Appendix B, if a use is non-expressive, the fourth statutory factor, market effect, will also tend to favor a finding of fair use because, by definition, non-expressive uses pose no threat of direct expressive substitution. However, the non-expressive use cases courts have decided to date tend to base their findings of lack of market effect on the specifics of the challenged use.

an adverse effect on the “market for or value of” a copyrighted work, it is arguable that non-expressive uses that substantially undermine copyright incentives could be considered unfair.

In particular:

(1) A court in some future case may well consider whether a defendant had lawful access to the works used as training data under the fourth factor. A class-action lawsuit filed against OpenAI alleges (quite plausibly) that the GPT models were trained on “notorious ‘shadow library’ websites like Library Genesis (aka LibGen), Sci-Hub, and Bibliotik.”²³

Copyright owners do not have a right to charge for transformative uses or non-expressive uses as such, but they do have a right to charge for access to their works. Although commercial use is usually a red herring in relation to transformative uses, it might make sense for considerations relating to lawful access to weigh much more heavily on commercial users, as opposed to those engaged in noncommercial research at nonprofit institutions. Fair use should not shield a user who obtains training materials by inducing others to infringe copyright (i.e., by copying without permission for reasons that are not justified by fair use, for example by encouraging the creation of, or providing direct support for, so-called shadow libraries of infringing materials). However, prohibiting academic research on illegal text corpuses will generally not benefit copyright owners or further the interests copyright is designed to promote.

(2) Likewise, a future court might extend the fourth factor to consider whether, in scraping material from the Internet, the defendant ignored robot.txt files indicating a desire to opt out of search engine indexing and similar activities. Likewise, a court might conclude that scraping material from a website in violation of its terms of use was relevant to the fourth factor, if the inability to rely on such exclusions substantially undermined copyright incentives.

Once again, respect for technological and contractual opt-outs is a consideration that should weigh much more heavily on commercial users, as opposed to those engaged in noncommercial research at nonprofit institutions.

(3) A plaintiff might argue that it is unfair to systematically extract valuable uncopyrightable material from a website or other information source and then use that material as a substitute for the functionality of the website. This argument would be strongest where the systematic extraction was likely to significantly undermine the website’s incentives for original content production.

²³ Tremblay et al v. OpenAI, Inc. et al, Docket No. 4:23-cv-03223 (Doc. 1 at 7)(N.D. Cal. Jun 28, 2023)

This argument is hard to reconcile with the view that the idea-expression distinction is meant to encourage competition where the competing product does not include too much of the plaintiff's original expression.

I don't endorse the argument, but it is plausible, and it awaits judicial resolution.

CONCLUSION

Generative AI does not require a major overhaul of the U.S. copyright system, at this time.

If Congress is considering new legislation in relation to AI and copyright, that legislation should be targeted at clarifying the application of existing fair use jurisprudence, not overhauling it.

Israel, Singapore, and South Korea have recently incorporated fair use into their copyright statutes in recognition of the fact that the flexibility of the fair use doctrine gives U.S. technology and research communities a significant comparative advantage.²⁴ Several other jurisdictions, most notably Japan, the UK, and European Union, have specifically adopted exemptions for text data mining that allow for using copyrighted works as training data for machine learning.²⁵

²⁴ Most notably, Section 19 of the Israeli Copyright Act allows for fair use and is closely modeled on Section 107 of the U.S. Copyright Act. Sections 190-194 of the Singaporean Copyright Act of 2021 set forth a version of the fair use doctrine that is more complicated than Section 107, but substantively similar. Article 35-3 of the Korean Copyright Act also provides for fair use and lists four factors that are equivalent to those found in Section 107.

²⁵ Japan's copyright laws have allowed some scope for machine learning techniques since 2009, however the Japanese Copyright Act was amended in 2018 to remove several perceived impediments. See Sections 30-4, 47-4, and 47-5. Recent news reports that the Japanese government "will not enforce copyrights when it comes to training generative artificial intelligence (AI) programs," see e.g. Matt Growcoot, *Japan Declares AI Training Data Fair Game and Will Not Enforce Copyright*, June 05, 2023 (<https://petapixel.com/2023/06/05/japan-declares-ai-training-data-fair-game-and-will-not-enforce-copyright/>) have likely misconstrued official statements as reflecting a change in the law.

The UK amended its laws in 2014 to allow for text data mining for non-commercial uses. See UK Copyright, Designs and Patents Act 1988, § 29A (UK) (amended by Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, No. 1372. The UK government announced in 2022 that it would expand this exception for commercial uses as well in 2022. See Intellectual Property Office, *Consultation outcome, Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation*, June 28, 2022 (<https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation>).

The 2019 EU Directive on Copyright in the Digital Single Market ("DSM Directive") requires EU Member States to adopt copyright exceptions for text data mining that would permit training machine learning models on copyrighted works. Article 3 of the DSM Directive requires all Member States to implement a broad copyright exception for TDM in the not-for-profit research sector. Specifically, members must allow research organizations and cultural heritage institutions to make reproductions and extractions of copyrighted works "for the purposes of scientific research, text and data mining." The Article 3 exemption is intended to be immune from contractual override or technological roadblocks. Article 4 of the DSM Directive contains a second mandatory exemption that is more inclusive, but narrower in scope. The Article 4 exemption is open to all would-be data miners (commercial and non-commercial alike), but it is focused more narrowly on reproductions and extractions for the purpose of "text and data mining," i.e., it lacks the broader "scientific research" purpose of Article 3. Unlike Article 3, the Article 4 exemption is not protected from contractual override or technological

Copyright law should encourage the developers of generative AI to act responsibly, with due care for the rights and legitimate interests of others. However, if our laws become overly restrictive, corporations and researchers will simply move key aspects of technology development to other jurisdictions, such as Japan and Israel.²⁶

The acceleration of AI development raises many serious policy questions beyond intellectual property rights. Generative AI can be a tool of disinformation and fake news, it can be used to generate deepfakes, and AI can reflect and sometimes amplify historical patterns of bias and unfairness. Generative AI might make us more productive, but that same productivity might lead to job losses. Lawmakers may even be concerned about AI will soon surpass human intelligence and escape from human control. However, copyright law is not a very subtle instrument for addressing broader social questions.

Congress should also consider whether any intervention is needed at all. Although I can imagine that some good might come from legislation addressing copyright and AI, I also have confidence that the issues that are currently unsettled can be resolved by the courts.

Moreover, I think that the most pressing issues for Congress to address in relation to AI and intellectual property rights are not copyright issues at all. I am deeply concerned that advances in deepfakes or synthetic media will be weaponized to harass, injure, and defame individuals and contribute to a toxic media environment where all sources of information are distrusted. There are limits to what Congress can do to address these issues, but I believe a national right of publicity law is needed to replace the current hodgepodge of state laws, and that we are overdue for a national data privacy law.

Thank you for the privilege of testifying here today.



Matthew Sag

roadblocks. Moreover, the Article 4 exemption is subject to an express reservation by the rightsholders. In other words, rightsholders should be able to opt out of Article 4 by simply reserving their rights “in an appropriate manner, such as machine-readable means in the case of content made publicly available online.” The implementation of the DSM Directive is ongoing.

²⁶ The Israel Ministry of Justice recently issued guidance that the use of copyrighted materials in the machine learning context is permitted under existing Israeli copyright law. In particular, the opinion found that ML typically will fall within the scope of the Israel Copyright Act’s fair use provision. (<https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/machine-learning.pdf>)

APPENDIX A: WHEN SHOULD A HUMAN BE CREDITED WITH AUTHORSHIP OF SOMETHING CREATED USING GENERATIVE AI?

How much human involvement with an AI process is enough to make the work copyrightable will depend on the specific circumstances.

Generative AI is often used as a tool in the creative process. A person who instructs a Generative AI with enough detail, such that model output reflects that person's original conception of the work, should be regarded as the author of the resulting work.

However, simple text prompting is unlikely to meet this standard. Simple text prompts are not enough to make someone an author; they are more akin to an instruction to an assistant to create a work.²⁷ The figure below, which I created on Midjourney using the prompt “*A law professor standing outside the US Senate, photo realistic*” does not qualify for copyright protection.

Figure: An uncopyrightable image created using Midjourney



Although I caused the image to be created, I had no idea of what latent concepts the Midjourney system would pull together to create it, nor did I have any idea of how these latent concepts would be combined. The image reflects my instruction in very general terms, but those instructions were far too general to merit copyright protection. They fall on the idea side of the idea-expression distinction.

²⁷ The Copyright Office also takes the view that images created by text prompts do not reflect an author's original conception, rather they are more akin to an instruction to create an image as one might give to a photographer, or a graphic artist. See Copyright Office correspondence Re: Zarya of the Dawn (Registration # VAu001480196), dated February 21, 2023, p.9. (“As the Supreme Court has explained, the ‘author’ of a copyrighted work is the one ‘who has actually formed the picture,’ the one who acts as ‘the inventive or master mind.’ *Burns-Giles*, 111 U.S. at 61. A person who provides text prompts to Midjourney does not ‘actually form’ the generated images and is not the ‘master mind’ behind them.”)

Earlier this year, the Copyright Office revoked Kristina Kashtanova’s registration of an 18-page comic book, *Zarya of the Dawn*, when it learned that the illustrations in *Zarya* had been created using the text-to-image platform Midjourney.²⁸

Figure: *Zarya of the Dawn*, pages 1-2



The Copyright Office conceded that Kashtanova was entitled to copyright protection for the text she had written, the overall story she had created and the selection and arrangement of images in the comic. However, the Copyright Office concluded that there was no copyright in the individual images produced by the Midjourney AI because those images were “produced by a machine or mere mechanical process that operates randomly or automatically *without any creative input or intervention from a human author.*”²⁹ The Copyright Office saw no evidence that Kashtanova controlled or directed the final form of the images, rather she merely chose which ones to adopt and which to refine with further instructions.

The Copyright Office’s decision in *Zarya of the Dawn* was reasonable on these facts, but I would note that “creative input or intervention” comes in many forms and the ultimate test remains whether someone’s “original intellectual conception” is reflected in the final form of the work. Accordingly, there is no reason in principle why prompts couldn’t be detailed enough to meet the traditional threshold of authorship in some cases. Sophisticated prompts

²⁸ Copyright Office correspondence Re: *Zarya of the Dawn* (Registration # VAu001480196), dated February 21, 2023. The Copyright Office made a similar ruling in relation to Steven Thaler’s application to register a two-dimensional artwork titled “A Recent Entrance to Paradise” created by a computer program. See Second Request for Reconsideration for Refusal to Register A Recent Entrance to Paradise (Correspondence ID 1-3ZPC6C3; SR # 1-7100387071), February 14, 2022. (<https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf>)

²⁹ *Id.* Quoting COMPENDIUM (THIRD) § 313.2 (emphasis added).

that specify details of an image should be sufficient to meet the requirement that the work that results from and reflects a person's original conception of the expression.

Furthermore, refining text prompts and choosing between different outputs should also be recognized as way in which a human using Generative AI could meet the authorship standard. Perhaps not on the specific facts of *Zarya of the Dawn*, but potentially in some future case. Many types of authorship involve generating alternatives and choosing between them. Consider, for example, a painter who flings paint at a canvas and then decides whether to fling more paint, or she decides to start again on a fresh canvas. The painter has only a loose idea of what the work will look like as it takes shape, but when the work is finished, it is surely a work of authorship within the contemplation of the statute.³⁰ Or, consider photography. The author's control over timing and framing are often considered central to the copyrightability of photos. There is really no difference between choosing when to take a photo and selecting one frame out of continuous reel. If that is so, then it makes sense to recognize selection and adoption as indicia of authorship. On this reasoning, the Copyright Office's decision in *Zarya* may have been overly conservative. Nonetheless, I expect the law in this area to continue to develop through Copyright Office registration decisions and federal court litigation. I do not see any benefit in additional legislative guidance at this time.

³⁰ Dan L. Burk explores many permutations of this hypothetical in Dan L. Burk, *Thirty-Six Views of Copyright Authorship, by Jackson Pollock*, 58 Hous. L. Rev. 263 (2020).

APPENDIX B: THE RELATIONSHIP BETWEEN NON-EXPRESSIVE USE AND FAIR USE

The fair use doctrine allows considerable scope for training machine learning algorithms on copyrighted works because non-expressive uses are generally fair use.

Training machine learning algorithms on copyrighted works necessarily involves making copies of those works.

There is no question that the best known LLMs today were built by ingesting massive quantities of sounds, images, and text from the Internet, with very little or no regard to whether those works were subject to copyright and whether the authors would object.³¹ Such copying implicates the copyright owner's exclusive right to reproduce the work.³²

*Thus, in the absence of express or implied permission, the legality of copying copyrighted works for use as training data for AI rests on the fair use doctrine.*³³

The fair use doctrine permits copying original expression in some circumstances. The doctrine was codified in Section 107 of the Copyright Act of 1976, however, that codification leans heavily on case law as early as 1841,³⁴ and analogous doctrines of fair dealing and fair abridgement that date back almost as far as the invention of copyright itself.³⁵

*Copyright law draws a fundamental distinction between protectable original expression, and unprotectable facts, ideas, abstractions, and functional elements.*³⁶

This distinction is often referred to as the idea-expression distinction or the idea-expression dichotomy. The idea-expression distinction means that copying valuable facts and ideas, or

³¹ Google, Meta, and OpenAI all use content from Wikipedia (online encyclopedia), GitHub (software), Project Gutenberg (public domain books), and massive collections of text scraped from the Internet (using various combinations of the Common Crawl, C4, and the Pile). Some of this material is in the public domain, some is licensed under terms that may allow for free reuse, but much is under copyright and copied without permission. Most obviously, the Books 2 dataset used by OpenAI's GPT-3 and GPT-4, is almost certainly based on "shadow library" websites such as Library Genesis (aka LibGen) and Bibliotik. See Tremblay et al v. OpenAI, Inc. et al, Docket No. 4:23-cv-03223 (Doc. 1 at 7)(N.D. Cal. June 28, 2023). EleutherAI's documentation on the Pile, comes close to admitting that Books 3 in the Pile is based on shadow libraries. See Leo Gao, et al. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv preprint 2020 (<https://arxiv.org/pdf/2101.00027.pdf>).

³² 17 USC 106(1).

³³ 17 USC 107 (Notwithstanding Section 106, the fair use of a copyrighted is not infringement).

³⁴ Folsom v Marsh 9. F.Cas. 342 (C.C.D. Mass. 1841).

³⁵ Matthew Sag, *The Prehistory of Fair Use*, 76 BROOK. L. REV. 1371 (2011) (Tracing the origins of the modern fair use doctrine back to cases dealing fair abridgment as early as 1741).

³⁶ The idea-expression distinction has been part of the common law of copyright since at least the 1880 Supreme Court

learning techniques, drawing inspiration, or emulating the general style of a copyrighted work is not infringement.³⁷ In addition, the idea-expression distinction also informs the way courts apply the fair use doctrine.

Reflecting the idea-expression distinction, courts have consistently held that technical acts of copying which do not communicate an author's original expression to a new audience are fair use. Such uses are referred to as non-expressive uses.

Examples of “non-expressive use” include:

- Copying software in order to extract uncopyrightable facts and interoperability keys (“reverse engineering”),
- An automated process of copying student term papers to compare to other papers so as to detect plagiarism,
- Copying html webpages to make a search engine index,
- Copying printed library books to allow researchers to conduct statistical analysis of the contents of whole collections of books, and
- Copying printed library books to create a search engine index.

Caselaw in relation to these non-expressive uses indicates that even though these uses involve significant amounts of copying, they do not interfere with the interest in original expression that copyright is designed to protect.³⁸ These uses involve copying as an intermediate step towards producing something that either does not contain the original expression of the underlying work or contains a trivial amount. Thus, non-expressive uses have consistently held to be fair use.

- In a 1992 decision, *Sega Enters., Ltd. v. Accolade, Inc.*, and again in 2000 in *Sony Computer Entm't, Inc. v. Connectix Corp.*, the Ninth Circuit held that software reverse engineering—a process that involves making several copies of the code to extract vital but uncopyrightable elements needed to make interoperable programs—was fair use.³⁹ In *Sega v. Accolade*, the court referred to copying to extract uncopyrightable

case of *Baker v. Selden*, 101 US 99 (1880) and it is reflected in Section 102(b) of the Copyright Act. 17 USC 102(b) “In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.” A similar principle is found in the TRIPs Agreement Article 9(2) “Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such.”

³⁷ For example, in *Baker v. Selden*, 101 U.S. 99 (1879) copying forms used to implement a system of bookkeeping was held not to infringe the copyright in the book that initially proposed that method of bookkeeping. The defendant was free to copy the method, and to copy the forms needed to implement that method.

³⁸ For a more detailed explanation, see Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. Rev. 1607 (2009); Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. COPYRIGHT SOC'Y OF THE U.S.A. 291 (2019).

³⁹ *Sega Enters., Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992); *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000).

elements as “a legitimate, essentially non-exploitative purpose.”⁴⁰ In *Sony Computer Entertainment v. Connectix*, the court expressly recognized that “the fair use doctrine preserves public access to the ideas and functional elements embedded in copyrighted computer software programs.”⁴¹

- In *A.V. ex rel. Vanderhye v. iParadigms, LLC*, the Fourth Circuit held that copying student papers into a reference database for comparison against new student papers was fair use.⁴²
- In *Authors Guild, Inc. v. HathiTrust*, in 2014 the Second Circuit held that making digital versions of printed library books for research purposes was fair use.⁴³ The court’s reasoning relied on the non-expressive nature of the use. The court explained:

“the creation of a full-text searchable database is a quintessentially transformative use [because] the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn. Indeed, we can discern little or no resemblance between the original text and the results of the [HathiTrust Digital Library] full-text search.”⁴⁴

- A differently constituted panel of the Second Circuit reached much the same conclusion in 2015 in *Authors Guild, Inc. v. Google, Inc.*, (the *Google Books* case).⁴⁵ In *Google Books*, the court addressed both the complete copying of millions of library books to make them searchable, and the display of small snippets of the books in search result menus. The complete copying is an example of non-expressive use; the snippet displays illustrate the application of a more traditional transformative use analysis.⁴⁶

When courts have declined to find fair use in *superficially* similar cases, it is invariably because the challenged use was *not* non-expressive and thus, on the facts presented, the potential substitution effect was too significant.

- In *Fax News v. TVEyes*, the Second Circuit held that a media monitoring service that copied and electronically searched television broadcasts went beyond the scope of fair use when it allowed users to save, watch, and share ten-minute long video clips of

⁴⁰ *Sega Enters., Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1523 (9th Cir. 1992).

⁴¹ *Sony Computer Entm’t, Inc. v. Connectix Corp.*, 203 F.3d 596, 603 (9th Cir. 2000).

⁴² *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630 (4th Cir. 2009)

⁴³ *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014)

⁴⁴ *Id.* at 97-98.

⁴⁵ *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

⁴⁶ The court held that that the display of three-line snippets to add context to book search results was transformative in purpose and that it was reasonable in proportion to that purpose. Those snippets allowed a user to verify that a book suggested by the search engine was in fact relevant to her interests. In addition, the snippets were so brief that they did not pose any risk of fulfilling the readers demand for the original expression of the underlying manuscripts. *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

the copyrighted programs.⁴⁷ In the court’s view, those ten-minute video clips would, “likely provide TVEyes’s users with all of the Fox programming that they seek and the entirety of the message conveyed by Fox to authorized viewers of the original.”⁴⁸ In other words, the court was concerned that rather than primarily providing information about the content of particular news segments, the length of the video clips was such that they would substitute for those segments in their entirety.

- In *Associated Press v. Meltwater U.S. Holdings, Inc.*, the Southern District Court of New York held that fair use did justify the actions of another media monitoring company, Meltwater.⁴⁹ Meltwater scraped news articles on the web to provide its subscribers with excerpts and analytics. However, the lawsuit did not challenge Meltwater’s use of copyrighted news articles to provide metadata and analytics to its subscribers, even though these services also necessitated copying. The court noted that this was “an entirely separate service” and implied that if it had been challenged, it would have been found to be transformative, and thus fair use.⁵⁰ Instead, like the successful plaintiff in *TVEyes*, the Associated Press focused on the length and significance of Meltwater’s extracts provided to subscribers. The court agreed that Meltwater’s extracts were too long and too close to the heart of the work;⁵¹ it also held that Meltwater had failed to show that the amount of the extracts was reasonable in light of its stated purpose to operate like search engine.⁵²

(e) Like “transformative use,” the words “non-expressive use” do not appear in the literal text of Section 107 of the Copyright Act. Nonetheless, concluding that a use is non-expressive has profound implications for each of the fair use factors outlined in Section 107.

(i) An overview of fair use in copyright law

The fair use doctrine permits copying (or distribution, display, or performance) without permission in certain circumstances, depending on the *purpose, proportionality, and effect* of that copying.⁵³ Copying that amounts to fair use is not merely excused, it is not infringement and

⁴⁷ Fox News Network, LLC v. TVEyes, Inc., 883 F. 3d 169(2nd Cir. 2018)

⁴⁸ *Id.* at 179.

⁴⁹ *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537 (S.D.N.Y. 2013)

⁵⁰ *Id.* at 557 (“The display of that analysis—whether it be a graphic display of geographic distribution of coverage or tone or any other variable included by Meltwater—is *an entirely separate service*, however, from the publishing of excerpts from copyrighted articles. The fact that Meltwater also offers a number of analysis tools does not render its copying and redistribution of article excerpts transformative.” (emphasis added))

⁵¹ *Id.* at 558.

⁵² *Id.*

⁵³ These considerations are reflected in factor one (purpose and character of the use), factor three (amount and substantiality of the portion used), and factor four (effect on the market for or value of the copyrighted work). Factor two (nature of the copyrighted work) simply reminds courts to take context into account when addressing the substantive considerations of purpose, proportion, and effect. Some authorities suggest that the nature of the work, whether it is creative/informational, or published/unpublished is a stand-alone consideration such that some works merit greater copyright protection than others. This approach is ill-conceived. The nature of the work is not an independent factor that

thus requires no further license or excuse.⁵⁴ Courts in the United States apply the fair use doctrine through a four-factor test set out in Section 107 of the Copyright Act. The factors are interrelated and should be treated as a framework for analysis rather than a scorecard or a checklist.⁵⁵

(i) If a use is non-expressive, then its “purpose and character” will favor a finding of fair use under the first statutory factor.

The first fair use factor calls for an evaluation of “the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes.”⁵⁶ Since the Supreme Court’s 1994 decision in *Campbell v. Acuff-Rose*, whether the defendant’s use was “transformative”⁵⁷—meaning that the use added “something new, with a further purpose or different character”—has been the central question under the first factor.⁵⁷ The Court’s 2023 decision in *Andy Warhol Foundation v. Goldsmith* (“*AWF*”)⁵⁸ emphasizes that the question of “whether an allegedly infringing use has a further purpose or different character . . . is a matter of degree, and the degree of difference must be weighed against other considerations, like commercialism.”⁵⁸

AWF reaffirms the importance of transformative use but implicitly rejects lower court rulings that had found uses to be transformative where there was no significant difference in purpose.⁵⁹ *AWF* helpfully clarifies the reason why a transformative use has featured so

weighs in favor or against a finding of fair use, it is simply the context in which courts must apply the substantive considerations of purpose, proportion, and effect set out in factors one, three, and four respectively. Images are not less worthy of copyright protection than text, but it is much harder to selectively comment on an image or use just part of an image as evidence than it is with purely textual works. Accordingly, full quotations of an image might be reasonable and proportional in circumstances where partial quotation of the text would be. The statute is not wrong to direct courts to think about the nature of the work; indeed, it would be quite impossible to analyze the purpose, proportion, and effect of the defendant’s use without taking into account the nature of the work. Moreover, works like computer software may have special characteristics that inform a fair use analysis. *Google LLC v. Oracle America, Inc.*, 141 S. Ct. 1183, 1198 (2021) (“... fair use can play an important role in determining the lawful scope of a computer program copyright It can help to distinguish among technologies. It can distinguish between expressive and functional features of computer code where those features are mixed. It can focus on the legitimate need to provide incentives to produce copyrighted material while examining the extent to which yet further protection creates unrelated or illegitimate harms in other markets or to the development of other products.”) Note that the majority’s fair use analysis in *Google v. Oracle* begins with the second factor. *Id.* at 1201.

⁵⁴ 17 USC 107 (“... the fair use of a copyrighted work ... is not an infringement of copyright.”).

⁵⁵ The Supreme Court’s 1994 decision, *Campbell v. Acuff-Rose*, rejects any attempt to reduce the four statutory factors to a checklist. *Campbell* and subsequent cases remind us that the codification of fair use in the 1976 Copyright Act was not intended to change what is essentially a common law doctrine; and that the statutory factors are interrelated; and thus, that the factors must be “explored, and the results weighed together, in light of the purposes of copyright.” *Campbell v. Acuff-Rose*, 510 U.S. 569, 577 (1994); see also, Pierre Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1110 (1990) (“The factors do not represent a score card that promises victory to the winner of the majority.”)

⁵⁶ 17 USC 107.

⁵⁷ *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994). For an empirical analysis, see Clark D. Asay, Arielle Sloan, and Dean Sobczak, *Is transformative use eating the world*, 61 B.C. L. REV. 905 (2020).

⁵⁸ *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (2023), Slip Opinion at 12. (emphasis added).

⁵⁹ *L.e., Cariou v. Prince*, 714 F.3d 694 (2d Cir. 2013) which suggested that merely imposing a “new aesthetic” on an existing

prominently in the case law: the more transformative a use is, the less likely it is to substitute for the copyright owner's original expression. Consider classic fair uses such as parody, commentary, or criticism may include substantial portions of the author's original expression, but these uses are so intrinsically different that they do not usually pose any risk of expressive substitution. In contrast, merely adding an overlay of new expression provides no such comfort.

In terms of the first fair use factor, non-expressive uses are not just transformative, they are highly transformative.⁶⁰ By definition, a non-expressive use does not usurp the copyright owner's communication of her original expression to the public because the expression is not communicated.

(ii) If a use is non-expressive, then the third statutory factor which considers "the amount and substantiality of the portion used" will also favor finding of fair use.

The ultimate question under the third fair use factor is whether the amount of copying was reasonable in relation to a purpose favored by fair use.⁶¹ Although non-expressive uses typically involve making complete literal copies, that copying has been found to be reasonable because it is an intermediate technical step in an analytical process that does not lead to the communication of the underlying original expression to a new audience. Accordingly, courts in non-expressive use cases have found the third factor weighs in favor of the defendant.⁶²

work was enough to be transformative.

⁶⁰ A.V. v. iParadigms Liab. Co., 544 F. Supp. 2d 473, 482 (E.D. Va. 2008): "This Court finds the "purpose and character" of iParadigms' use of Plaintiffs' written works to be *highly transformative*. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students' written works from plagiarism. iParadigms achieves this by archiving the students' works as digital code and makes no use of any work's particular expressive or creative content beyond the limited use of comparison with other works." AV Ex Rel. Vanderhye v. iParadigms, LLC, 562 F. 3d 630, 640 (4th Cir, 2009): "The district court, in our view, correctly determined that the archiving of plaintiffs' papers was transformative and favored a finding of "fair use." *iParadigms' use of these works was completely unrelated to expressive content and was instead aimed at detecting and discouraging plagiarism.*" Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 97 (2nd Cir. 2014): "... we conclude that the creation of a full-text searchable database is a *quintessentially transformative* use."); Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 216-7 (2d Cir. 2015): "We have no difficulty concluding that Google's making of a digital copy of Plaintiffs' books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a *highly transformative* purpose, in the sense intended by Campbell." Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 217 (2d Cir. 2015): "... through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. *We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell as strongly favoring satisfaction of the first factor.*"

⁶¹ Campbell v. Acuff-Rose, 510 U.S. 569, 586-87 (1994). ("[T]he extent of permissible copying varies with the purpose and character of the use.") In *Campbell*, the Court characterized the relevant questions as whether "the amount and substantiality of the portion used ... are reasonable in relation to the purpose of the copying," and noting that the answer to that question depends on "the degree to which the [copying work] may serve as a market substitute for the original or potentially licensed derivatives[.]" *Id.* at 586-588.

⁶² AV Ex Rel. Vanderhye v. iParadigms, LLC, 562 F. 3d 630, 642 (4th Cir, 2009); Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87, 98 (2nd Cir. 2014) "In order to enable the full-text search function, the Libraries, as we have seen, created digital

(iii) *If a use is non-expressive, the fourth statutory factor which considers the effect on the “potential market for or value of the copyrighted work” will also favor a finding of fair use in many cases.*

The “market” and “value” referred to in the fourth fair use factor are not simply any benefit the copyright owner might choose to nominate; these terms mean something much more specific. A critical book review that quotes from a novel does not have an adverse market effect if it persuades people to buy different book instead;⁶³ a report from a plagiarism detection service might depress the market for helping students cheat on their homework, but that is hardly a cognizable injury under copyright law.⁶⁴ More generally, in *Campbell* and subsequent cases, the courts have recognized that the copyright owner has no protectable interest in preventing criticism, parody,⁶⁵ or simply locking up unprotectable ideas and expression.⁶⁶

By definition, if a use is non-expressive then it poses no direct threat of expressive substitution and thus should generally be preferred under the first factor (purpose and character) and considered harmless under the fourth factor (market effect).

The argument that copyright owners have an inherent right to charge for non-expressive uses and thus suffer an adverse market effect under the fourth factor is transparently circular. In theory, every defendant in every fair use case could pay the plaintiff for the right to engage in the challenged use, but if the use is fair there is no obligation to pay. To avoid such circular arguments, courts have limited market effect under the fourth factor to those

copies of all the books in their collections. Because it was reasonably necessary for the HDL to make use of the entirety of the works in order to enable the full-text search function, we do not believe the copying was excessive.” *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 221 (2d Cir. 2015) “Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier’s transformative purpose and was done in such a manner that it did not offer a competing substitute for the original.” *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 221-222 (2d Cir. 2015) “As with *HathiTrust*, not only is the copying of the totality of the original reasonably appropriate to Google’s transformative purpose, it is literally necessary to achieve that purpose. ... While Google makes an unauthorized digital copy of the entire book, it does not reveal that digital copy to the public. The copy is made to enable the search functions to reveal limited, important information about the books.”

⁶³ *Campbell v. Acuff-Rose*, 510 U.S. 569, 591-592 (1994) “We do not, of course, suggest that a parody may not harm the market at all, but when a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act. Because parody may quite legitimately aim at garroting the original, destroying it commercially as well as artistically, the role of the courts is to distinguish between biting criticism that merely suppresses demand and copyright infringement, which usurps it.” (cleaned up).

⁶⁴ *AV ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 464 (4th Cir. 2009) (“Clearly no market substitute was created by iParadigms, whose archived student works do not supplant the plaintiffs’ works in the ‘paper mill’ market so much as merely suppress demand for them, by keeping record of the fact that such works had been previously submitted In our view, then, any harm here is not of the kind protected against by copyright law.”)

⁶⁵ *Campbell v. Acuff-Rose*, 510 U.S. 569, 577-79 (1994); *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 482 (2d Cir. 2004) (“[C]riticisms of a seminar or organization cannot substitute for the seminar or organization itself or hijack its market.”); *Bill Graham Archives v. Dorling Kindersley, Ltd.*, 448 F.3d 605 (2d Cir. 2006) (“A copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody ... or other uses of its own creative work.”) (internal quotations omitted).

⁶⁶ *Sega Enters., Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992); *Sony Computer Entm’t, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000).

that represent a cognizable copyright interest.⁶⁷ Accordingly, in *HathiTrust*, the second Circuit rejected the plaintiff's argument that not being paid for text mining was a cognizable harm, noting that "[l]ost licensing revenue counts under Factor Four only when the use *serves as a substitute for the original* and the full-text-search use does not."⁶⁸ Likewise, in *Google Books*, the court insisted on focusing "on whether the copy brings to the marketplace *a competing substitute for the original*, or its derivative, so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."⁶⁹

The substitution the courts are referring to here is expressive substitution, not simply the threat of a more competitive marketplace. A non-expressive use can be harmless under the fourth factor even if it results in the creation of a competing product—as long as the competing product does not contain an infringing level of original expression taken from the plaintiff's work. In *Sega v. Accolade* and again in *Sony Computer Entertainment v. Connectix Corp.*,⁷⁰ the Ninth Circuit found that reverse engineering a gaming console in order to produce interoperable games (*Sega*), and a rival gaming platform (*Sony*), was fair use. In both cases the Ninth Circuit found that there was no cognizable market effect because the rival products did not contain any protectable expression derived from the plaintiffs' consoles. The defendants were entitled to use uncopyrightable elements from those consoles to make new independent creative expression possible.⁷¹

⁶⁷ *Campbell v. Acuff-Rose Music*, 510 U.S. 569, 591-92 (1994) (no cognizable market effect where parody or criticism depress demand for the original work); see also *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 607 (9th Cir. 2000) (noting that a videogame manufacturer's desire to foreclose competition in complementary products was understandable, but that "copyright law ... does not confer such a monopoly."); *Bill Graham Archives v. Dorling Kindersley, Ltd.*, 448 F.3d 605, 615 (2d Cir. 2006) ("[A] copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work.") (citations and quotations omitted).

⁶⁸ *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 100 (2d Cir. 2014) (emphasis added).

⁶⁹ *Authors Guild v. Google, Inc.*, 804 F.3d 202, 223 (2d Cir. 2015) (emphasis added).

⁷⁰ *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F. 2d 1510, 1523. *Sony Computer Entertainment v. Connectix Corp.*, 203 F. 3d 596, 608.

⁷¹ *Id.*

APPENDIX C: CLAIMS THAT GENERATIVE AI INFRINGES COPYRIGHT IN “STYLE” ARE MISCONCEIVED

The practice of allowing users to invoke the names of living artists in text-to-image generators may be problematic, but not because there is copyright in artistic style. The claim that copyright protects abstract artistic style is entirely at odds with the idea-expression distinction. However, although no one owns an artistic style, copyright law does recognize copyrightable characters.

(1) Confusing style with copyrightable characters

As I explain in a forthcoming article:

The most important implication of copyrightable characters in the context of generative AI is that, practically speaking, the level of similarity required to establish infringement is reduced in that context. This problem is compounded because the way LLMs learn to associate visual elements with text descriptions effectively primes them to memorize the very relationships that constitute a copyrightable character. When presented with a thousand different images associated with the word “Snoopy,” a model like Stable Diffusion learns which characteristics are consistently repeated across the entire set. In the words of the Ninth Circuit, the model focuses on the “consistent, identifiable character traits and attributes,” and gives more weight to those that are “especially distinctive.”⁷²

As noted above, memorization is more likely if a text description is closely associated with a particular image over and over again; it is also more likely if the image is relatively simple or relates to a single subject. Because the threshold of substantial similarity required to infringe on a copyrightable character is comparatively low, multiple variations of the same character in the training data will result in a latent concept for the character that is readily identifiable and easily extracted by invoking the name of that character. This explains why it is easy to provoke Midjourney to recreate copyrightable characters such as Snoopy and Mickey Mouse, but difficult to come close to infringing a Salvador Dali painting with a simple text description of the scene.⁷³

Some of the claims currently being made about copyright in style would be better recast in terms of copyrightable characters. For example, Sarah Andersen is a successful and award-winning cartoonist and illustrator; she is also the lead plaintiff in recent class action lawsuit filed in relation to Stable Diffusion, Midjourney and DreamUp.⁷⁴ Andersen is the author of a semi-autobiographical webcomic, Sarah’s Scribbles, a self-deprecating series focusing on a

⁷² DC Comics v. Towle, 802 F.3d 1012, 1021 (9th Cir. 2015).

⁷³ Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: <https://ssrn.com/abstract=4438593>).

⁷⁴ Andersen v. Stability AI Ltd. Case 3:23-cv-00201 (Document 1) (January 13, 2023 N.D. Cal.)

bug-eyed millennial's struggles with adulthood.⁷⁵ The figure below contrasts an original four panel comic strip by Anderson featuring her characteristic stressed millennial alongside two images created in Midjourney. The prompt for both images was “a Sarah Andersen style black and white comic where Sarah asks ‘why is copyright is so stressful?’”

Figure : A Sarah Andersen original and two Midjourney copies



Ignoring Midjourney's gibberish text for the moment, there is a certain similarity between the panels. The images are far from exactly the same, but there is a stylistic overlap in terms of minimalism, black and white, oversized heads, and exaggerated bug-eyes. The panels are also similar in that the characters are in a perpetual state of extreme emotion. The case for infringement would be stronger if we replaced the random text produced by Midjourney with something echoing one of Andersen's perpetual themes of social anxiety, body image, or laziness. If the Midjourney images above were deemed infringing, it would not be because they reproduce a particular style of drawing, it would be because all of those stylistic features are applied to depict a character that is uncomfortably close to Andersen's messy-haired protagonist.⁷⁶

(2) Artists can be injured by the use of their names as style prompts in ways that have little to do with copyright

For example, one of the most commonly invoked style prompts in early 2023 was Greg Rutkowski,⁷⁷ an artist who is well known for his richly detailed depictions of Dungeons &

⁷⁵ See e.g., Sarah Andersen, *Adulthood is a Myth: A Sarah's Scribbles Collection (Volume 1)* (2016).

⁷⁶ Note that the allocation of responsibility for infringement between Midjourney and the user requesting images in a Sarah Andersen style is unclear. Arguably, the user is the direct cause of the infringement and Midjourney would only be liable if it met the standards of vicarious, contributory, or inducement-based liability. But even so, if such uses are commonplace, it may undermine the fair use argument that justified ingesting Andersen's works into the training corpus without her consent.

⁷⁷ Melissa Heikkilä, "This artist is dominating AI-generated art. And he's not happy about it" MIT TECH. REV. Sep 16, 2022 (available at <https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/>) (noting that prompts in Midjourney and Stable Diffusion for the artist Greg Rutkowski were more popular than for Picasso and other more famous artists.)

Dragons and similar worlds in a style comparable to the romantic English painter, William Turner.⁷⁸

Figure: Comparison of Greg Rutkowski to William Turner



Rutkowski's name is primarily used as a shortcut to invoke high-quality digital art generally, or in relation to fantasy motifs, and not in attempt to recreate his style specifically.⁷⁹ But even if it were, the suggestion that Rutkowski has a copyright interest in painting fire-breathing dragons in this style is absurd. The idea that once an artist is well-known for depicting a certain subject matter or using a certain technique that these are off-limits to subsequent creators is an anathema to copyright.

Nonetheless the harm that Rutkowski suffers by having his genuine works crowded out in Internet searches by tens of thousands of images produced "in the style of Rutkowski" is very real. That harm could be avoided if the names of individual artists paired with images in the training data were replaced with more general descriptions, or if platforms like Midjourney and DALL·E-2 did not allow prompts in the names of individual living artists.

(3) Personality is not style

If generative AI re-created someone's distinctive appearance or voice, that person should have recourse under right of publicity. Congress should enact a national right of publicity

⁷⁸ I mean no disrespect to Mr. Rutkowski.

⁷⁹ This assessment is based on my review of prompts including "Greg Rutkowski" located using a Google image search on April 20, 2023.

law to ensure nationwide and uniform protection of individuals' inherently personal characteristics.

APPENDIX D: GENERATIVE AIs TYPICALLY LEARN ABSTRACTIONS AS OPPOSED TO COPYRIGHTABLE EXPRESSION

LLMs usually learn from the training data at a fairly abstract level. Moreover, the output of generative AI usually combines abstract latent features learned from the training data in a way that ensures model outputs look nothing like specific model inputs.

(1) Learning abstractions, not specific details

For example, when a text-to-image model such as Stable Diffusion or Midjourney is trained on hundreds of images with labels that include the words “coffee” and “cup,” it develops a model of what a coffee cup should look like. If the system is working properly, that model looks nothing like any individual coffee cup from the training data.

As part of the research for my forthcoming Houston Law Review article,⁸⁰ I compared a random set of coffee cup images from the Stable Diffusion training data⁸¹ with a newly rendered “cup of coffee that is also a portal to another dimension.” The figure below contrasts several of the coffee cups in the training data against the model output.

Figure: Coffee cups in the training data compared to model output



The comparison is instructive. The coffee cup image on the right has a vague similarity to some of the coffee cup images on the left—the cup is round, it appears to be made of white ceramic, it has a small single handle, the color of the liquid is essentially black, transitioning to brown. However, beyond these generic features, this cup is not substantially similar to any

⁸⁰ Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023) (Available at SSRN: <https://ssrn.com/abstract=4438593>)

⁸¹ Images based on a search of the Baio & Willson database on March 09, 2023 (https://laion-aesthetic.datasette.io/laion-aesthetic-6pls/images?_search=coffee+cup&_sort=rowid).

particular image from the training data.⁸² The process of training the Stable Diffusion model has not simply memorized images involving coffee cups, it has learnt something about the latent concept of a coffee cup distinct from cakes, sunsets, sunrises, newspapers, and men with facial hair—all of which can be seen in the training data examples on the left.

(2) Combining abstractions to produce something new

The output of generative AI usually combines abstract latent features learned from the training data. For example, when I instructed Midjourney to create a teddy bear in rich opulent clothing with ultra-realistic textures, with a hypnotic stare, reading a newspaper, it produced the following image.⁸³

Figure: Opulent Bear (Midjourney)



No doubt, the picture is influenced by thousands of images paired with each of the relevant keywords. All of the images of teddy bears in the training data inform a latent construct of a teddy bear nested within the model's latent space; likewise, all of the images of someone staring hypnotically inform a latent construct of a hypnotic stare.⁸⁴ Loosely speaking, when

⁸² I also reviewed images on Have I Been Trained, a website that purports to index "5.8 billion images used to train popular AI art models," i.e., the LAION-5B database (<https://haveibeen-trained.com/>).

⁸³ The prompt was: "32K RAW Photography, low angle shot, teddybear reading newspaper, rich opulent clothing, hypnotic stare, textures: ultra realistic." The form of the prompt was copied from another discord user and simplified to produce this example.

⁸⁴ Generative AI models don't form distinct "latent models" for separate concepts, rather they learn a comprehensive "latent space" that represents the diverse array of features present in the training data. In the context of machine learning, particularly with generative models, a "latent space" refers to the mathematical space where the AI model compresses and organizes the complex patterns it identifies in the training data. The term "latent" means hidden or not directly observable. In this case, the latent space embodies the underlying structure or patterns within the data that are not immediately

these elements are combined and unpacked, the result is something entirely new; it is not merely a copy or a derivative work based on something in the training data.⁸⁵ One of the fun things about this particular image is that although the bear's demeanor is consistent with him staring hypnotically, the bear is actually wearing sunglasses that leave the details of his gaze to our imagination. The output is heavily influenced by the training data in the sense that it can't generate concepts it has never encountered during training. Thus, while the specific combination (a teddy bear in rich opulent clothing with a hypnotic stare, reading a newspaper) is novel, all of its components exist in some form within the training data. But that is almost invariably true of human authored works as well.

apparent. This space is a high-dimensional continuum where similar features are located closer together, allowing the model to generate diverse outputs by navigating this space.

⁸⁵ The model does not merely "combine" and "unpack" these learned latent features. Instead, generates a novel instance that may share certain characteristics with the input prompt based on the latent space. So describing the output as a combination of learned features is a bit simplistic.

APPENDIX E: RECOMMENDED BEST PRACTICES FOR COPYRIGHT SAFETY FOR GENERATIVE AI

In a forthcoming article,⁸⁶ I make 10 recommendations for Copyright Safety for Generative AI that should be followed in order to reduce the risk of copyright infringement. The Key recommendations are summarized below:

Those who copy copyrighted material to train LLMs should:

1. Take reasonable measures to ensure that their models are not exposed to duplicates of the same copyrighted work.
2. Take reasonable measures in training, and in model deployment, to reduce the probability of infringement of copyright, trademark, and publicity rights. These measures may include (i) curating and pre-processing the training data (ii) using reinforcement learning through human feedback, and (iii) installing restrictions on model outputs.
3. Take reasonable measures to safeguard individual privacy interests.
4. Keep detailed records of the works used and their provenance. How to balance company's legitimate interests in preserving valuable commercial secrets and the rights of the public to know how their works have been used in training LLMs is a tricky question. Congress may wish to consider giving the Copyright Office or some other administrative agency the power to audit and inspect such records.

Implementing these recommendations in a fast-changing environment may be challenging—deduplication is an especially thorny technical problem. However, the absence of good faith efforts to do so could be relevant to a court's fair use determination.

I also recommend that, to assist companies developing LLMs to take appropriate steps to reduce potential copyright and related harms, the U.S. Copyright Office or some other administrative agency should maintain a registry of artists and copyright owners who do not want their names, or the names of their characters, used as style prompts in text-to-image platforms.

⁸⁶ Matthew Sag, Copyright Safety for Generative AI (May 4, 2023)(Available at SSRN: <https://ssrn.com/abstract=4438593>).

Questions from Senator Tillis
for Ben Brooks
Witness for the Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing
“Artificial Intelligence and Intellectual Property -
Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?
2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?
3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of

everything that is ingested by the AI publicly available?

- b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?
4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?
 - a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?
5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?
6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?
8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?
9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?
10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?
12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?
 - a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?
13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?
14. Your CEO has publicly suggested that, in the future, Stability AI and other AI models will be fully licensed. While hearing this is promising, what more can be done now – across the AI community – to address the current concerns amongst the copyright community?

15. Some AI developers have said that the ingestion of copyrighted works is transformative and qualifies as fair use. What impact does the Supreme Court's recent decision in *Andy Warhol Foundation v. Goldsmith* have on that position?

16. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data.

Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

17. Where does Stability AI get the training data that is used to train their AI systems?

18. What steps, if any, does Stability AI take to ensure that your training data does not include copyrighted material?

19. Has Stability AI sought out licenses from copyright owners for copyrighted content that is included in your training data? Please explain why or why not.
20. What considerations would impact your decision to use copyrighted material to train your model as opposed to non-copyrighted material?
21. Let's assume that under *Andy Warhol Foundation v. Goldsmith* the use of copyrighted works for training AI is not considered transformative. Do you believe the use of these works would still qualify as fair use looking at the four factors? Which particular factors support your position?
22. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model.

Could you please speak to how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

Questions from Senator Tillis
for Jeffrey Harleston
Witness for the Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing
“Artificial Intelligence and Intellectual Property -
Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?
2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?
3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of

everything that is ingested by the AI publicly available?

- b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?
4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?
 - a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?
5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?
6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?
8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?
9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?
10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author’s digital content.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?
12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?
 - a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?
13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?
14. What are artists saying to you about their voices and music being used without their consent? Is there something that we need to do to protect a person's voice so that others can't use generative AI to exploit their unique qualities?

15. Should someone be able to use an artist's voice to train an AI model so that an AI system can sing other people's songs or say something new? How is such different from a person imitating another person's sound?
16. Do you have any thoughts on some of the positive ways that AI can be used to support artists?
17. In your testimony, you outline a number of different ways that you believe AI-generated music violates copyright law. You also point out how often UMG and artists collaborate with other artists and partners to bring music to fans. Whether, it's the recording of a cover of a popular song, sampling of a song's lyrics or composition, or a collaboration on a remix, artists and the industry regularly produce new or reimagined music while following the law.

Could you please explain the legal and appropriate process of creating music involving samples, for instance, and compare that process to what is happening with generative AI music that we've seen in such volume over the last few months?

18. AI may be increasingly used instead of non-featured vocalists and studio musicians in the making of sound recordings. Assuming this trend continues, what impact will it have on the music industry?

Questions from Senator Tillis
for Karla Ortiz
Witness for the Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing
“Artificial Intelligence and Intellectual Property -
Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?
2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?
3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of

everything that is ingested by the AI publicly available?

- b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?
4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?
 - a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?
5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?
6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?
8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?
9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?
10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?
12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?
 - a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?
13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?
14. Can you describe some of the challenges that you face as a working artist and how you expect generative AI technology to impact those challenges?

15. Do you feel that the value of your work has been negatively impacted by the advancement of these new generative AI systems? If so, how?

16. While style is not protected by copyright, do you believe that allowing for style to be copied and for users of AI systems to include artists' names in prompts results in a greater chance that the output will be substantial similar to an ingested work and potentially infringing?

Questions from Senator Tillis
for Dana Rao
Witness for the Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing
“Artificial Intelligence and Intellectual Property -
Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?
2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?
3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of

everything that is ingested by the AI publicly available?

b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generative AI will take over jobs?

6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?
8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?
9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?
10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author’s digital content.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?
12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?
 - a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?
13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?
14. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data.

Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

15. Can you explain the thinking behind Adobe's decision to adopt an opt-in approach to the use of copyrighted works as training data?
 - a. Are you concerned that this approach will put Adobe at a competitive disadvantage with those who do not see the need to obtain consent?
16. Some of the most popular generative AI tools were built by indiscriminately scraping material – including copyright protected material – from the Internet. Can you explain Adobe's approach to building its AI tools and how it's different than other companies?
17. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model.

Could you please explain how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

18. Some have suggested different licensing structures for compensating copyright owners for the use of their works in AI training. What licensing structures have you seen or used that have worked to the mutual benefit of both AI companies and copyright owners?

Questions from Senator Tillis
for Matthew Sag
Witness for the Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing
“Artificial Intelligence and Intellectual Property -
Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?
2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?
3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of

everything that is ingested by the AI publicly available?

b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generative AI will take over jobs?

6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?
8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?
9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?
10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?
12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?
 - a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?
13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?
14. Language Learning Machines are increasingly being used to generate source code and help software developers write software. Such models can require a vast amount of source code and thus can turn to open-source software (OSS) for scraping publicly available source code.

- a. If AI models are trained on OSS, does that infringe on the copyright of the respective authors?
 - b. If an AI model is used to generate code, does that generated code constitute a derivative work? And if that AI model was trained on copyleft-licensed OSS, must it also be licensed under copyleft?
15. Some AI developers have said that the ingestion of copyrighted works is transformative and qualifies as fair use. What impact does the Supreme Court's recent decision in *Andy Warhol Foundation v. Goldsmith* have on that position?
16. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data.

Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

17. To what extent should the use of AI impact whether a human creator receives a copyright? In other words, if AI is being used as a “tool,” should a human still be able to receive a copyright if they have independently contributed creative content?
18. Let’s assume that under *Andy Warhol Foundation v. Goldsmith* the use of copyrighted works for training AI is not considered transformative. Do you believe the use of these works would still qualify as fair use looking at the four factors? Which particular factors support your position?
19. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model.

Could you please explain how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

20. Some have suggested different licensing structures for compensating copyright owners for the use of their works in AI training. What licensing structures

have you seen or used that have worked to the mutual benefit of both AI companies and copyright owners?

Response to questions from Senator Tillis for Mr. Ben Brooks, Head of Public Policy

Artificial Intelligence and Intellectual Property – Copyright

Judiciary Subcommittee on Intellectual Property

August 2023

- 1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?**

The United States can demonstrate global leadership by developing a measured response to AI that realizes the full potential of these technologies while addressing emerging risks. As Congress considers the future of AI oversight, we encourage policymakers to vigorously promote open innovation in AI. These technologies will be the backbone of our digital economy, and it is essential that the public can scrutinize their development. Specifically, we urge policymakers to:

- **Foster innovation in open models.** Open models promote transparency, competition, and security in AI. By inspecting open models, researchers and authorities can 'look under the hood' to verify performance, identify risks, and develop new mitigations. By customizing open models, creators and developers can build new AI tools or launch new AI ventures without spending millions on research and computing. By building on open models, organizations can establish their own secure and independent AI capabilities without relying on a handful of firms for critical AI infrastructure. Future policy should account for the role of open models within the AI ecosystem, and support the development and deployment of open models as a public resource.
- **Adopt a risk-based approach to oversight.** Performance requirements for AI systems should be proportional to risk. Requirements may be more stringent where the likelihood or consequence of harm is greater, without imposing those requirements on lower risk systems or lower risk applications. For example, an AI system deployed in a higher-stakes domain such as healthcare, finance, education, or public administration may be subject to more rigorous requirements governing reliability, interpretability, and assurance than an AI system deployed in a lower-stakes domain. In addition, oversight frameworks should be adaptable. They should have the flexibility to respond to new technology and new research. Oversight frameworks should avoid prescribing a single means of compliance that may be ineffective or outdated: safety will depend on layers of mitigations across the AI supply chain.
- **Support diversity in the AI ecosystem.** The AI ecosystem is diverse, from large firms building proprietary technology through to everyday developers freely sharing their technology. Future policy should nurture this diversity. In particular, policymakers should

carefully consider the impact of regulation on grassroots innovation. For example, 'one size fits all' requirements that assume every AI developer is a well-funded corporation may have a chilling effect on collaborative research by independent developers, researchers, or entrepreneurs. Developers may be reluctant to contribute to AI research or share AI resources, leaving the community dependent on a small number of firms for access to AI technology.

These principles can help to ensure that AI is deployed in a way that meets public expectations for safety while promoting transparency and competition in AI. Other jurisdictions recognize the importance of these principles. For example, the United Kingdom has committed to a sector-specific regulatory approach that focuses on the risks presented by AI in specific domains.¹ The European Union has recognized the importance of open innovation, and the diversity of AI supply chains.² However, it is yet to be seen whether the final version of the draft AI Act will adopt a risk-based approach to AI models, and we continue to advocate for amendments to the final Act to protect grassroots developers.³ In addition, a range of jurisdictions including Singapore, Japan, and the European Union have reformed their copyright laws to create safe harbors for AI training that achieve similar effects to fair use.⁴ These reforms aim to establish a culture of 'open learning' that promotes access to large and diverse datasets, helping to make AI safer, more effective, and less biased.

2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?

Users should know when they are interacting with AI applications that generate content or make significant decisions.⁵ To that end, policymakers may consider disclosure obligations for application developers who provide certain AI services to a user. Additionally, policymakers may consider privacy obligations that require users to affirmatively consent prior to the collection of their data for AI training, since this data may be used to target the user with personalized content.

3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available? Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available? Under what circumstances, if

¹ British Government, *A Pro-innovation Approach to AI Regulation*, March 2023 at [3.2].

² European Parliament, draft AI Act, May 2023, Recitals 12a, 12b, 12c. See also draft Commission and Council text.

³ European Parliament, draft AI Act, May 2023, Articles 2, 3, and 28b.

⁴ See the respective text and data mining exceptions. See also the *Regulation of Technologies Review* conducted by the UK Government Chief Scientific Adviser in March 2023 at 9: "If the government's aim is to promote an innovative AI industry in the UK, it should enable mining of available data, text, and images (the input) and utilise [sic] existing protections of copyright and IP law on the output of AI."

⁵ Stability AI, 'Advocating for Open Models in AI Oversight', May 2023. See our submissions to the National Telecommunications and Information Administration and the Office of Science and Technology Policy.

any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

Datasets should be assessed for safety, bias, and suitability before they are used to train an AI model. Independent auditing is one way to help promote scrutiny of these datasets. By inspecting datasets, researchers, developers, and users can anticipate potential risks or limitations in a model. For example, by inspecting image datasets, researchers can determine whether an image model fairly represents a particular community or demographic. By inspecting language datasets, researchers can determine whether a language model accounts for cultural, political, or language diversity. Developers can work to correct these limitations and adjust undesirable behavior, either by training a new model or by ‘fine-tuning’ an existing model with additional data.

Open datasets may not be feasible in all cases. For example, models may be customized or ‘fine-tuned’ with proprietary data, and a developer may not be able or willing to share that data publicly. The content and structure of a dataset affects the performance of an AI model, and we acknowledge that developers may choose to keep their datasets closed. Nonetheless, model developers should ensure that datasets are screened for unsafe content, bias, and opt-out requests from creators. In sensitive contexts – such as healthcare, finance, or public administration – additional dataset requirements may be determined by regulators.

4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable? Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

Whether a prompt is copyrightable is a fact-sensitive inquiry. Certain prompts may consist of unique and highly detailed instructions, with original expressive content, created by a user after significant experimentation. These prompts may satisfy the criteria for copyrightability in certain circumstances. In our view, the copyrightability and registrability of an AI output is a separate inquiry (see response to Question 12 below).

5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?

We believe AI is a tool that can help creators express themselves, but it is not a substitute for creators. Instead, AI can help to accelerate the creative process. AI tools can help existing creators boost their productivity, experiment with new concepts, and perform complex tasks as part of a wider workflow. In addition, AI can lower barriers to entry for people who do not have the resources or training to realize their creative potential. In this way, AI can help open up

economic opportunities for new and existing creators, much like earlier tools such as computers, software, smartphones, and social media.

However, there is a risk that the economic benefits of AI accrue disproportionately to AI firms. That's why Stability AI is committed to putting AI in the hands of everyday creators and developers through open models. Open models enable everyday creators to participate in this new industrial revolution as builders – not just consumers – of technology. Creators can adapt, integrate, and experiment with open models to develop AI applications that best support their work (see examples from our written testimony). They can develop these AI tools without spending millions on research or computing power. In this way, we expect the economic benefits of open models will be shared by a broader community of creators, developers, and entrepreneurs, not just Silicon Valley.

In the long term, we are sensitive to the implications of AI for routine tasks and the effects of AI on certain kinds of work. To that end, we encourage policymakers to carefully scrutinize the impact of AI across the national economy. Over time, it may be necessary to reflect on how we learn, train, and work in light of these tools.

- 6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?**

Stability AI is unable to comment on matters subject to litigation.

- 7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?**

AI is a tool, not a replacement for creators. AI can help to accelerate the creative process, but AI depends on a human in the loop to provide creative direction. AI may deliver comparable performance to humans in the technical execution of specific creative tasks (e.g. extending, filling in, or editing a photorealistic image). However, audiences value a range of attributes in art beyond technical execution. They take into account the composition as a whole, the context of the work, the creator, and subjective interpretations of meaning. By itself, AI cannot impart these attributes to content. We expect the community will continue to value human works – indeed, we may value them at a premium. Smartphones didn't destroy photography, and word processors didn't diminish literature, despite radically transforming the economics of creation. Instead, they created new fields of artistic expression, giving rise to new demand for services, new markets for content, and new creators.

- 8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do**

you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?

We believe that training AI models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law, including the Constitutional goal of 'promot[ing] the progress of science and useful arts'.⁶ These models learn the unprotectable and non-expressive ideas, facts, and structures within a visual or textual system. 'Open learning' of these facts about our world is essential to recent developments in AI, and it is doubtful that these groundbreaking technologies would be possible without it. The US has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation. We encourage policymakers to preserve this culture of 'open learning' in any future reform.

However, there are specific harms that may warrant future intervention. For example, the improper use of likeness can be problematic if it wrongfully implies a person's endorsement of, affiliation with, or promotion of a work or idea. The improper use of personal likeness should be governed by clear rules that specify impermissible use. These may include right of publicity laws (for performance or voice likeness) and deepfake laws that prohibit specific kinds of misleading conduct in political, commercial, or personal settings (including non-consensual intimate imagery). Policymakers may consider strengthening these laws. Further, policymakers can help to fortify the wider information ecosystem by encouraging the adoption of content provenance standards by AI applications and advanced content moderation systems by social media or streaming platforms.

9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?

AI models review existing content to learn the relationship between words, ideas, and fundamental visual or textual features. They can apply this knowledge to help a user generate new and unseen content as part of a creative workflow. In this sense, AI is comparable to other assistive technologies, from cameras to software. Like other tools, AI may accelerate the creative process; increase the volume of creative outputs; and change the economics of creation. Like other tools, AI may be used properly or improperly by a user. However, these characteristics do not relevantly distinguish AI from other technologies that alter the ease, speed, volume, or cost of performing a creative task.

10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

⁶ U.S. Constitution, Article I, Section 8, Clause 8.

Machine-readable metadata can help to communicate essential information about a particular work. For example, Stability AI is exploring machine-readable metadata that indicates whether a creator would like to be removed from datasets used for AI training. This metadata would be additional to existing protocols, such as robots.txt, that indicate whether a web page can be ‘crawled’ for purposes such as search engine indexing. Today, machine-readable opt-outs are required in order for a rightsholder to exercise certain opt-out rights under the European Union text and data mining exception.⁷

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?

Existing frameworks are sufficient to address emerging challenges in AI, including statutory authorities that apply across sectors (e.g. the *Federal Trade Commission Act*), domain-specific regulatory authorities (e.g. FDA, DOT, SEC, and DOL), and existing product liability principles. The risk of an AI system is determined by how that AI system is deployed for a specific application in a specific environment. Transparency, auditability, and documentation requirements will vary accordingly, and domain-specific rules will best account for these variations.

Existing regulatory agencies are equipped with the legal tools to respond to these AI deployments, from AI decision making (e.g. automated vehicles) to AI content (e.g. healthcare interactions with a patient or business interactions with a consumer). However, regulatory agencies may not be adequately resourced to respond to these challenges. Prior to further legislative intervention, we encourage Congress to conduct a rapid gap analysis to determine (i) where regulatory competence falls short in any sector or domain and (ii) where agency funding and personnel may need to be supplemented.

12. Have you reviewed the U.S. Copyright Office’s Registration Guidance for “Works Containing Material Generated by Artificial Intelligence” and, if so, what are your views on the guidance? Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?

In principle, we acknowledge a threshold of authorship below which a work with negligible human input may not qualify for registration. However, we are concerned that the U.S. Copyright Office’s Registration Guidance and recent decisions (such as *Kashtanova*) do not account for the many ways in which human input may rise above the threshold. For example, a creator (i) may fine-tune their own model on their own content to better evoke their established style, (ii) provide detailed creative direction via prompts, inputs, and settings that narrowly define the range of possible outcomes, or (iii) use the model’s suggestion as a starting point, and refine the initial image via editing.

⁷ European Union, Directive 2019/790 (Copyright Directive), 2019, Article 4.

Sufficient creative control demonstrated in one or more of these ways should qualify a work for registration. A user with clear expressive intent, and who has demonstrated that they directed the model, should be able to register their work. We welcome further clarification on this issue. Uncertain or discretionary guidance means that professional creators may be unfairly disadvantaged by their use of AI tools within a wider creative workflow.

- 13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?**

Stability AI participated in the U.S. Copyright Office listening session on AI and visual arts in May 2023. These sessions were comprehensive and professional, featuring a diverse range of viewpoints. Notably, these sessions were structured around different modalities to reflect important nuances in the copyright treatment of different media. While we object to aspects of the recent guidance on AI (see response to Question 12 above), we welcome further engagement with the U.S. Copyright Office in relation to AI, including registrability.

- 14. Your CEO has publicly suggested that, in the future, Stability AI and other AI models will be fully licensed. While hearing this is promising, what more can be done now – across the AI community – to address the current concerns amongst the copyright community?**

There are practical steps that AI developers can take today to address emerging concerns. These include respecting opt-outs in training data; watermarking and tagging AI-assisted content to help users and platforms distinguish it from human-generated content; and mitigating defective behavior such as overfitting. These are detailed in our written testimony. In addition, we are actively exploring new ways to incentivize creators in the digital economy. These are early days, but we look forward to sharing more information with the Subcommittee as these initiatives unfold.

- 15. Some AI developers have said that the ingestion of copyrighted works is transformative and qualifies as fair use. What impact does the Supreme Court's recent decision in *Andy Warhol Foundation v. Goldsmith* have on that position?**

For the reasons given in our testimony, we believe training is an acceptable, transformative, and socially-beneficial use of content that is supported by the fair use doctrine.

- 16. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data. Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?**

A 'base' AI model is typically trained on broad datasets, including content aggregated from the Internet. A base model will demonstrate some level of competence in basic tasks. However, to make AI models useful, they will need to be customized or 'fine-tuned' for specific tasks using high-quality data. That may include proprietary or confidential data. This data should be accessed and handled appropriately, and we urge greater transparency in how private data is used in the development of AI models. Further, user data should be treated with care, since it may be used to train AI models that target a particular individual with personalized content. We have previously advocated for robust privacy obligations that require affirmative user consent before an AI application is able to collect or retrieve personal data for AI training.

One way to ensure that data is properly accessed and handled is by customizing AI models 'in house'. For example, a regulated financial institution may need to customize an AI model to assist in analysis, decision making, or customer support. The financial institution will need to audit the performance of the model for reliability; train or fine-tune the model without exposing sensitive customer data to third-parties; and retain full control over the AI model without relying on a third-party provider.

Open models can help organizations to customize these AI models safely and securely. By building on open models, organizations can 'look under the hood' to identify potential risks or limitations. They can use the open model to train a custom model for specialized applications without exposing their proprietary data to a third-party AI firm. After customizing the model, they retain full control over their AI capabilities, including the distinctive settings or 'parameters' that define the model's performance, without relying on a third-party AI firm. In this way, organizations can help to ensure the security of their proprietary or sensitive data.

17. Where does Stability AI get the training data that is used to train their AI systems?

Stable Diffusion is trained on a subset of two billion images from the LAION-5B dataset, a public dataset of five billion image URLs and text captions assembled by the nonprofit Large AI Open Network. We filter this dataset to remove unsafe images, among other criteria. By filtering training data before it ever reaches the AI model, we can help to prevent users from generating unsafe content. In addition, we filter this data for opt-out requests from creators, and we will incorporate these opt-out requests in future training.

18. What steps, if any, does Stability AI take to ensure that your training data does not include copyrighted material?

Stability AI has proactively solicited opt-out requests from creators, and will honor these over 160 million opt-out requests in upcoming training. Going forward, we are exploring new technical standards for machine-readable opt-outs, so that opt-out metadata follows the content wherever it goes. In addition, the datasets used for training our flagship model, Stable Diffusion, respected

industry-standard digital protocols like robots.txt, which indicate whether a website consents to automated data collection for ancillary purposes such as indexing or analysis.

19. Has Stability AI sought out licenses from copyright owners for copyrighted content that is included in your training data? Please explain why or why not.

We believe that training AI models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law, including the Constitutional goal of “promot[ing] the progress of science and useful arts”.⁸ Through training, these models develop an understanding of the relationship between words, concepts, and fundamental visual or textual features. The model doesn’t rely on any single work in the training data, but instead learns by observing recurring patterns over vast datasets – billions of image and caption pairs, and hundreds of billions or trillions of words. The model does not store the material in this training data. They do not “collage” or “stitch” together original works, nor do they operate as a “search engine” for existing content. Instead, AI models apply this generalizable knowledge to help a user create new and unseen content.

However, we are sensitive to emerging concerns. As detailed in our testimony, we are working to address these concerns through technology, standards, and good practices. In addition, we are actively exploring new ways to fairly incentivize creators in the digital economy, and we would be pleased to share more details with the Subcommittee as these initiatives unfold.

20. What considerations would impact your decision to use copyrighted material to train your model as opposed to non-copyrighted material?

We believe that training AI models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine. Access to large and diverse datasets is essential to make AI safer, more effective, and less biased. However, as detailed in our testimony, we acknowledge emerging concerns among creators. We are working to address these concerns through technology, standards, and good practices.

21. Let’s assume that under *Andy Warhol Foundation v. Goldsmith* the use of copyrighted works for training AI is not considered transformative. Do you believe the use of these works would still qualify as fair use looking at the four factors? Which particular factors support your position?

We are unable to comment on matters subject to litigation. As detailed in our testimony, we believe training is an acceptable, transformative, and socially-beneficial use of content that is supported by the fair use doctrine.

⁸ U.S. Constitution, Article I, Section 8, Clause 8.

- 22. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model. Could you please speak to how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?**

Image models are trained on vast datasets consisting of digital images and their associated text captions. During training, these images are reviewed by the AI model to understand the relationship between words, ideas, and fundamental visual features. These images may be reviewed a number of times as the model is optimized to improve performance and mitigate risks, but they are not retained in the trained model. The model can subsequently apply this knowledge to help a user generate new works.

Questions from Senator Tillis
for Jeffrey Harleston
Witness for the Senate Committee on the Judiciary Subcommittee on
Intellectual Property Hearing “Artificial Intelligence and Intellectual
Property - Part II: Copyright”

1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?

The development of public policy surrounding AI is in its infancy, presenting the U.S. with an important opportunity to lead the world. As the Senate considers guidelines and rules for a responsible and safe AI ecosystem, we note the helpful commitments made by the G7, including the [Hiroshima Leaders’ Communiqué](#) (paragraph 38) and the [Ministerial Declaration of the Digital and Tech Ministers’ Meeting](#), (see paragraphs 42, 45 and 47).

The [EU’s AI Act](#), currently under consideration, includes helpful proposals on government review of generative AI models before release, continued assessment of those models, recordkeeping provisions, transparency and labeling obligations, and more. As one of the bill’s key recitals reads:

“...it is appropriate for the Commission and the AI Office to monitor and periodically assess the legislative and governance framework of such models and in particular of generative AI systems based on such models, which raise significant questions related to the generation of content in breach of Union law, copyright rules, and potential misuse.”

On the other hand, there are some policies, including ones that were adopted years ago, before the rise of generative AI, that the U.S. should avoid. For example, the Japanese polices regarding Text and Data Mining exceptions to copyright law as enacted in 2009 and amended in 2018. Another example includes the policies in Singapore, which were enacted in 2021. We would note that the United Kingdom explicitly rejected such policies last year in recognition of the irrevocable harm it would inflict upon their creative industries.

2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?

Yes. Content generated by an AI should be identified as such. Consumers have a right to know that what they are consuming is real. There is exceptionally high risk that content created by generative AI could mislead or deceive consumers, and that risk necessitates disclosure.

However, merely *disclosing* purely generative AI does not go far enough. If an artist's name, image, likeness, or voice is used without permission, or if an AI engine was trained on unauthorized copyrighted content, mere disclosure does not solve the underlying problem or mitigate the harm. To ensure consumers and fans aren't duped, and artists' livelihood and rights are protected, federal right of publicity legislation and enforcement of copyright law should also be enacted.

In short, the best way to mitigate the potential risks posed by generative AI is to address and regulate how AI models are trained in the first instance.

3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?

Detailed recordkeeping of data, materials, and information used to train a generative AI engine is essential for the following reasons:

- (1) **Accurate, truthful, and detailed recordkeeping would allow copyright holders to know when their content was used without permission.**

This is a fundamental tenet of 233 years of American copyright law and jurisprudence. AI developers need permission from copyright owners before their works can be used to train an AI algorithm.

Recordkeeping is the lynchpin of a copyright owner's ability to enforce their rights. Without the ability to see how an AI was trained, an untenable "Catch-22" exists: copyright law requires rightsholders to prove a registered copyright was violated before they can enforce their rights. Since most AI training sets are not transparent and available to

rightsholders, they must initiate expensive, lengthy litigation in order to reach discovery and prove that their content exists in the training sets. However, without *proof* that a copyright has been violated, rightsholders cannot even initiate the litigation it would take to reach discovery. In other words: rightsholders must sue in order to prove their registered works were used to train an AI, but they *can't* sue unless they can prove their registered works were used. The inability of rightsholders to protect themselves is an unreasonable consequence of the AI era, and Congress should remedy the problem.

- (2) **It would require those who create AI engines to be more thoughtful and careful about the content they use to train their AI engines.** Clarity on recordkeeping requirements would ideally inspire those creating AI engines to seek authorization in the first instance – reducing intellectual property theft and establishing a robust marketplace free of litigation and uncertainty.
- (3) Importantly, **transparency of training information would allow consumers and users of an AI to know what went into the AI that they are using.** Confidence in accuracy, veracity, and trustworthiness of training sets is essential to establish the foundation of a legitimate generative AI marketplace, maximizing potential benefits while mitigating the potential harms.
 - a. Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available?

There is no justifiable circumstance where an AI creator should not have to make their training sets available. While it has been argued that making training set information transparent would threaten “trade secrets” or proprietary information, those arguments do not hold water. The potential for bias, misinformation, and very real harm to American intellectual property owners is too great to justify that position. AI developers have a responsibility to disclose the material used to train their systems to the societies in which they operate.

- b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

An AI developer should make everything that is ingested for training of their model available, especially to rightsholders.

- 4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

A prompt provided by an AI user is, essentially, a text-based work. Therefore, it should be copyrightable to the same extent any text-based work is copyrightable. As the Supreme Court has held:

“The *sine qua non* of copyright is originality. To qualify for copyright protection, a work must be original to the author. Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity.” (*Feist Publications, Inc. v. Rural Telephone Service Company, Inc.*, 1991)

- a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

No. As a general matter, the copyrightability of a prompt and the copyrightability of an AI output are two separate and distinct inquiries, but which apply the same standards for copyrightability. If an output is generated purely by AI, it fails the copyrightability test as there is no human author.

That being said, many prompts (e.g., “create a picture of a cat on a surfboard”) will not contain sufficient originality or creativity needed to be copyrightable. And if the output is generated purely by AI, it fails the copyrightability test as there is no human author.

- 5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?

When copyright is violated, it has a very real financial impact on an artist's livelihood, with ripple effects felt throughout the entire music ecosystem.

Speaking as a music publisher and record label, the existence of content that violates our rights harms our ability to license a creator's music. That means lost U.S. GDP, tax revenue, and an inability to support as many employees, artists, and small businesses. It results in an inability to invest in and support as many new artists, leading to fewer opportunities for songwriters, fewer jobs for producers, studios, instrument and equipment manufacturers, musicians, managers, lawyers, stylists, directors, graphic designers, choreographers, video production crews, caterers, and so on. It means fewer tours supporting fewer venues across America.

In 2021, during the height of the COVID-19 pandemic, the core copyright industries added \$1.8 trillion dollars of value to U.S. GDP (accounting for 7.76% of the entire U.S. economy) and employed 9.6 million American workers (accounting for 4.88% of the nation's workforce). The copyright industries also provided the sustenance that fed our societal mental health during that unprecedented and difficult time of isolation, anxiety, and loss.

And it's particularly egregious because it effectively strips the artist/songwriter of their rights of integrity and control: they don't get to choose whether or not their work is used to create new works by others. And since those new works could be political, defamatory, violent, racist, or offensive it's not only troubling, but it can also be tremendously damaging to the artists' brand and commercial potential.

Imagine if someone stole everything of value from a business and used it to start a business to compete against them. That's exactly what's happening with a lot of the larger machine-learning AI models in existence today. This is textbook unfair competition: hundreds of digital services worldwide (as well as individual creators, advertising agencies, small businesses, content creation companies, and others) have followed the law and licensed the work; AI developers should operate under the same rules.

Unless we protect creators in this new era, the long-term consequences on human creativity and our culture could be dire. Some 120,000 songs are uploaded to streaming services *every day*. AI-generated songs will only get easier to make and that number will only rise. Will that make it harder or easier to discover good music created by humans? Will that improve the experience

for music fans? Will this flood of content be better or worse for artists attempting to make a living from their music? We think the answers are clear.

It will take some time for a legitimate, legal marketplace to develop, and you have the power to determine the speed at which it materializes. There is a hunger for marketplace actors who respect copyright. Pending litigation results in marketplace uncertainty stymies forward progress that benefits everyone. There is an urgent need for appropriate “rules of the road” for generative AI and we encourage you to act decisively and without delay.

6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

As always, this would be a fact-based analysis, but generally speaking:

TRAINING: If the AI system has been found to be infringing because it trained on copyrighted works without permission, the actors who engaged in that training violation would be liable for infringement. That could be a company or organization, a group of individuals, or a single individual.

PROMPTING: A “prompter” would not necessarily be liable if the content generated in response to their prompt included copyrighted material. A copyright violation occurs if the prompter makes use of the infringing work after its creation. If a prompter *sought* to produce a work through an AI that was infringing, that might render them liable under either direct or contributory infringement principles.

INFRINGING OUTPUTS: In terms of an output, the infringer(s) would be the person or company that is violating the rights guaranteed under U.S. Code Title 17 Sec. 106. If an AI produces something that infringes an underlying work, the question becomes who is copying, performing, transmitting, displaying (etc.) that infringing work.

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?

No – nothing can replace human creativity, experience, and expression. But in many cases, AI-generated imagery is already indistinguishable from human-created art – especially by the layman – and other forms of media are not far behind. People will always relate best to other people and generative AI will

never be able to make art informed by the human-lived, and often cultural, experience, but the output quality of AI tools will likely equal that of human-created works in the near future, which is why it is critical that Congress take steps to protect human creativity.

Because generative AI is capable of producing remarkably high-quality outputs, it can mislead or deceive consumers into believing something purely generated is real, factual, or has occurred. In the context of music, increasingly high-quality deep fakes or voice clones lead to cultural appropriation, consumer confusion, unfair competition against the actual artist, market dilution, and potentially irreparable damage to the artist's reputation, brand, and livelihood.

8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?

U.S. copyright law has enabled countless technological innovations since it was first established in 1790. Copyright law protects creators' innovations and rather than stymie technological innovation, it has *encouraged and fueled* it. Ethical technological innovation does not tread on the rights of other creators and innovators. Quality content drives a hunger for devices and technology that make the content shine, easier to access, and amazing to experience; and technology companies need quality content to spur a market for their services, software, and devices – just as content owners need technology to help their content reach fans.

When the law is clear and guardrails against appropriation and abuse are firmly established, technology and content *both* thrive. The proper legal guidelines and guardrails that protect our artists' innovation and expression are essential, and to trample the rights of creators in order to create unprecedented shortcuts for AI developers is unjust.

The ideal “win-win” scenario is one in which AI companies build their systems using content licensed from creators and content owners in an open market, and all participants benefit from the associated commerce.

9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras,

electronic music, and the like, all of which allow the public to develop and advance knowledge?

Creators use those machines as tools and instruments to produce original works that spring from their original thoughts and imagination. Generative AI is unique in that the creative elements are derivative by nature and born of technology that has been trained on a body of human-created works that have come before. In the case of generative AI, the creative choices are made by the technology, not the human. When using a camera, the photographer chooses what to shoot, how to frame it, where to focus and add light, how to edit it, etc. When a “photograph” is created by an AI, those choices are largely made by machine learning and guided by the materials and information it used to train the algorithm.

Allowing “the public to develop and advance knowledge” inherently suggests human knowledge. While AI can do many amazing things to advance scientific achievement, create efficiencies, and problem solve, we should be careful as a society not to equate the knowledge and learning generated by a machine learning AI with human knowledge and achievement. As computers assume the tasks so many of us struggled ourselves to learn and absorb, our global society should take care to ensure humanity maintains its own genuine intelligence, creativity, and knowledge.

10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author’s digital content.

Content created by the music industry is already incredibly rich in detailed metadata and digital information. The global digital music marketplace is reliant on that information for far more than ownership, licensing, and sales information. Metadata and the standardized digital messages sent between business partners may be invisible to the consumer, but together they form the “nervous system” of the vibrant digital music marketplace.

While each company surrounds its content in proprietary metadata, Digital Data Exchange, LLC (DDEX) enables the world’s music business partners to communicate that metadata seamlessly. DDEX is a not-for-profit, membership organization comprised by a consortium of media companies, music licensing

organizations, rights owners, digital service providers, and technical intermediaries.

Media-specific industry standards, such as DDEX for music, should continue to be used to transport data about creative works. Information regarding AI is already being implemented into the DDEX system and can help facilitate a legitimate generative AI marketplace for music. Creators and distributors also use industry standard identifiers for contributors and media (such as ISNI, ISRC, ISWC, and UPC for music). These same standards can be used to transmit metadata and content obtained from record labels for AI training purposes. New and developing labeling standards such as C2PA may be used to indicate the provenance of partially AI-created output and identify contributing works in the future, but the space is still developing.

Information regarding AI is already being implemented into the DDEX system and can help facilitate a legitimate generative AI marketplace for music.

11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?

No. There is a need to future-proof the law to fit a technology evolving and advancing at breakneck speed. There's also a clear problem today where rightsholders have great difficulty enforcing their rights because copyright law is not perfectly fit to allow rightsholders to protect themselves. Moreover, in order to cultivate a lawful, legitimate, trustworthy, and ethical AI marketplace, Congress needs to establish the fundamental "rules of the road" that enable that development.

It will take more than ensuring transparency and detailed recordkeeping to accelerate ethical AI development. Specifically, the U.S. should pursue the following changes to law:

- (1) Enact a federal Right of Publicity statute to ensure intellectual property protection for a name, image, likeness, or voice.
- (2) Ensure the transparency of AI training materials and allow a rightsholder to view detailed records of training inputs, without having to initiate litigation. Solving the rightsholder litigation "Catch-22" is paramount.
- (3) Require labeling of works substantially generated by AI.

(4) Future-proof copyright law to ensure it holds up to an ever-evolving AI technology.

12. Have you reviewed the U.S. Copyright Office’s Registration Guidance for “Works Containing Material Generated by Artificial Intelligence” and, if so, what are your views on the guidance?

a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?

We think the Copyright Office established the right principle: that human creativity is necessary for copyright protection and that material that is not generated through the human creative process should not be subject to copyright protection. When registering our works, we believe this means that AI-generated material should be disclosed to the Office when it takes the place of human creative expression.

There is a breadth of AI tools used in the creation of music, and art has always pushed the boundaries of technology. We assume the Copyright Office guidance will develop and evolve over time as the technology advances and the AI market more fully develops.

13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?

Yes. The music community was represented in both proceedings through our trade associations. In addition, UMG participated in the recent Copyright Office roundtable on Music and AI this past May.

These proceedings have been extremely helpful. While their efforts are ongoing, it is our hope they will serve as needed, necessary voices within Congress and the Administration – creators must be represented when decisions regarding AI policy are made, and we’re deeply hopeful they’ll actively advocate on our behalf. We’re also hopeful they act decisively and quickly, as the speed of AI advancement necessitates it.

14. What are artists saying to you about their voices and music being used without their consent? Is there something that we need to do to protect a person's voice so that others can't use generative AI to exploit their unique qualities?

Artists are understandably upset that their name, image, likeness, or voice is stolen and used to suggest that they're said, sang, or done something they never did. It's a horrifying violation of privacy, and since it could damage the artist's reputation, it could irreparably damage their career. An artist's voice and persona *is* their livelihood and to steal it – no matter the means – is wrong.

Adding to the harm, deep-fake and/or unauthorized recordings or visuals of artists generated by AI not only lead to consumer confusion, but they also unfairly compete against the actual artists, diluting the markets for their music and devaluing their brands.

AI-generated, mimicked vocals trained on vocal recordings extracted from our copyrighted recordings go beyond Right of Publicity violations and concerns about consumer deception, unfair trade practices, and privacy – copyright law has clearly been violated.

The Senate should enact a federal Right of Publicity statute, ensuring an artist's persona is protected as intellectual property.

15. Should someone be able to use an artist's voice to train an AI model so that an AI system can sing other people's songs or say something new? How is such different from a person imitating another person's sound?

Yes, but one should only use an artist's voice to train AI with the proper consent.

An AI voice clone is very different from a human imitation of another human's sound. In terms of economic impact, a generative AI model has no limits and can create an infinite number of recordings that sound exactly like an artist whose voice was stolen. This avalanche of recordings competes against the artist in the marketplace, and every song streamed is money stolen from the artist.

16. Do you have any thoughts on some of the positive ways that AI can be used to support artists?

When used in the service of artists and artistry, generative AI can be powerful and help artists reach new fans in exciting ways. In my testimony, I described how an artist we distribute was able to use a generative AI engine to train on his voice and the intonation of native speakers in six different languages. This allowed him to release a song in seven different languages at the same time, all in his own voice. It's thrilling to think of what such a tool could mean for music and artistry.

Aside from the potential artistic impact, generative AI is also poised to make the life of an artist easier. Touring with greater efficiency, finding new fans, streamlining licensing, monetizing music in new ways never imagined – generative AI could do great things for artists.

17. In your testimony, you outline a number of different ways that you believe AI-generated music violates copyright law. You also point out how often UMG and artists collaborate with other artists and partners to bring music to fans. Whether, it's the recording of a cover of a popular song, sampling of a song's lyrics or composition, or a collaboration on a remix, artists and the industry regularly produce new or reimagined music while following the law.

Could you please explain the legal and appropriate process of creating music involving samples, for instance, and compare that process to what is happening with generative AI music that we've seen in such volume over the last few months?

When an artist wishes to use a sample of a pre-existing recording in a new recording, clearance is generally sought from both the owner of the original (sampled) sound recording and the original (sampled/interpolated) musical work. That clearance is typically handled by the artist's record label and involves obtaining licenses from those original copyright owners in exchange for consideration. There is a well-developed market in clearing and licensing these samples, but these uses are subject to exclusive copyrights. Accordingly, the original creators/owners are entitled to deny these requests.

To date, in the generative AI market there has been simply a wholesale taking of content – entire catalogs – without *anyone seeking permission*. This is a process that cannot continue and is likely to lead to extensive litigation unless Congress asserts the basic tenets of copyright law, making clear that existing law applies to AI developers, just as it applies to everyone else.

18. AI may be increasingly used instead of non-featured vocalists and studio musicians in the making of sound recordings. Assuming this trend continues, what impact will it have on the music industry?

I am aware of a few instances where generative AI has been used instead of background vocalists or studio musicians. It wouldn't be in our interest to use generative AI in that way as it would result in an inability to obtain a copyright for those elements of the completed recording.



**Questions from Senator Tillis
for Karla Ortiz (08/08/23)
Witness for the Senate Committee on the Judiciary Subcommittee on Intellectual
Property Hearing “Artificial Intelligence and Intellectual Property — Part II:
Copyright”**

1. *Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?*

The U.S. has the world’s most vibrant economy for creators, including writers, artists, designers, and software programmers. I am among the millions of Americans who’ve built a career in that creative industry. If generative AI companies plan to participate in that economy, they need to comply with the same laws that the rest of us do. So far, they have not.

First and foremost, I believe that dataset transparency in generative AI (currently being considered in the EU) is a vital policy for keeping AI companies accountable. If AI companies are required to disclose their datasets, they will have an obvious incentive to make sure their data is acquired legally and ethically.

Further, I do not believe that the U.S. should enact any statutory exemption for AI training, akin to the Text and Data Mining exemption in the EU. Experience has already shown that generative AI companies will find ways around it. For example, certain AI companies are already abusing the exemption through a practice known as “data laundering”, where for-profit companies exploit a training exemption supposedly restricted to research. The fair-use doctrine under U.S. copyright law suffices for any generative AI company devoted to research.

I also wish to list again some of the policies that I think would benefit the U.S. greatly:

- a. Congress should update the Copyright Act to reaffirm that copyright requires human authorship.
- b. Congress should empower existing agencies to regulate the use of data to train generative AI models. This can take the form of requiring disclosure of training data, limiting the types of data that can be used to train AI models, closing “research to commercial” loopholes and ensuring compliance with these regulations.
- c. Congress should pass laws expressly authorizing those who have had their data used to train AI models without their consent the



right to vindicate those rights in federal court and to seek statutory damages. This can take the form of an amendment of the Copyright Act to authorize an express civil cause of action for those who have their data used to train AI models without permission. This can also take the form of passing a law authorizing a federal civil right of publicity cause of action.

- d. Immediate taxing of all companies that replace jobs with any form of generative AI.
 - e. Immediate protection of citizens' data that powers generative AI, including likenesses, voice, biometric data, private data, copyrighted works and so on.
 - f. A copyright registration should be deemed sufficient for copyright-infringement litigation as long as the registration is gained before trial.
2. *A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?*

With the caveat that I have no plans to use AI for my own work, I think an AI-disclosure rule may still be insufficient. The problem with an AI-disclosure rule is that it is possibly too blunt a tool. It may make sense today, when works can be readily classified into either human- or AI-generated. Over time, however, this will become difficult, as more work could be a combination of human and AI contribution. What amount of AI contribution needs to be disclosed? How can that even be measured or verified? And what accountability will there be for those who violate the disclosure rules to try to pass off their AI work as human-made? I believe the better policy is to make generative AI systems legally accountable at the source.

3. *What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?*

I strongly believe that dataset transparency is a critical policy for ensuring legal and ethical accountability of generative AI companies and the systems they create.

First, it creates a strong incentive for these companies to adopt legal and ethical practices for data collection and licensing.



Second, it's better for the generative AI industry at large, because customers of these systems will want to know that the outputs of generative AI systems are safe, legal, and ethical to use in other settings.

Third, it allows creators and artists to discover whether their data has been scraped for use in generative AI models. This third point is particularly important because in the case of large language models, the scraped data might be private or personal, and thus potentially affects every citizen with even the smallest digital footprint.

The U.S. has long imposed analogous labeling requirements for food and medicine for similar reasons. Disclosure requirements enhance safety and create confidence.

- a. *Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available?*

AI companies should always be required to make these records available.

- b. *Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?*

These records should always be made available. In a practical sense, what every generative AI system is offering is a certain kind of access to the underlying training data. All the “intelligence” in the system is derived from that training data. So it is conceptually impossible to frame the outputs of such a system as being separate from the inputs.

4. *Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?*

That is primarily a question for the U.S. Copyright Office. If a generative AI prompt is, say, a sequence of words, then the copyrightability of that sequence should be judged by the same standards as other textual works. I would note that the current AI image generators, such as Stable Diffusion and Midjourney, are introducing features to make prompting easier, in response to complaints from users that even the current text-prompting systems are too difficult and complicated. It seems inevitable that these systems will move in the direction of supporting simpler prompts, not more expressive ones.

- a. *Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?*



This is also primarily a question for the U.S. Copyright Office. In my view, copyrightability of a prompt and its output are two separate issues. As a human artist, when I make an oil painting, all the creative decisions originate in my brain and my hand. Whereas in a generative AI system, the person providing the prompt is supplying only a tiny fraction of what we might call creative agency. Mostly, the algorithmic decisions are being made by the generative AI system itself. So we have to consider the prompt separately from the output. Otherwise, we'll be led to absurd results. For instance, if the prompt were a copyrightable text like "provide an exact copy of an existing oil painting by Karla Ortiz / if you please" and the system produced such an image, then the copyrightability of the prompt obviously should not make the resulting infringing image copyrightable. This would also set a deeply unfair standard of essentially allowing AI users and AI companies to "launder" copyrighted works: anyone could type a prompt describing a previously copyrighted work and then claim copyright on both the prompt and the AI output. This act would single handedly destroy the need for anyone to respect copyrights.

5. *What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?*

As someone who works as a concept artist in the entertainment industry, it's apparent that the loss of jobs is not merely a "concern." A key reason generative AI systems exist is to replace human labor. IBM, for instance, [has announced](#) that it already plans to replace approximately 7,800 human jobs with AI. Emad Mostaque, CEO of Stability AI, [has openly predicted](#) that AI will have a "bigger economic impact than the pandemic." OpenAI has [published a paper](#) concluding that language models will have a significant impact on jobs, especially higher-income positions. Two major entertainment-industry creative unions, the WGA and SAG-AFTRA, are currently on strike in part because of disagreements with studios about the appropriate role for generative AI.

In my particular industry, every day brings more accounts of students and professionals seeing their projects reduced or losing opportunities altogether. I personally have been a part of three productions where generative AI has materially lessened my duties. As a freelancer who gets paid hourly, it has already impacted my bottom line. I expect that freelance concept artists and artists in the commercial-entertainment industries have and will continue to experience declining job assignments. I expect that there will also be downward pressure on pricing, and many of us will just have to find some other way to support ourselves and our families as the work dries up.

6. *If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?*



Both, depending on where in the process the infringement is happening. Infringement can occur a) during the training of the AI system, b) as part of the distribution and operation of the AI system, c) when the AI system is generating outputs, and d) possibly other points.

7. *In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?*

No, it cannot. And what’s more, AI companies agree. Human-generated content is indispensable as training data because it contains the widest variety of expression. Training an AI system on its own output leads to a [condition known as](#) “model collapse” where the system starts to “forget” things and thereby becomes less valuable. Because of this entanglement, it’s not a stretch to predict that if the AI companies bankrupt human creators, they’ll just end up bankrupting themselves. Thus, ensuring the health of the human creative industry should not be seen as a cost imposed on the AI industry, but as an investment in its long-term health and prosperity.

8. *A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?*

During the last 50 years, U.S. copyright law has moved almost uniformly in one direction: toward greater protection for authors and creators. Why? Because as our national economy has become more reliant on industries rooted in copyright, affirming these rights has become vital. It’s no coincidence that the last 50 years have also been the most prosperous in U.S. history. But now, the nascent AI industry is nudging Congress to reverse direction and erode protections for copyright. This makes no sense. If the generative AI opportunity is as big as these companies claim, then they should have no objection to complying with the established copyright regime that has been so successful for so long. Copyright law has **never** been a hobble to U.S. global competitiveness—on the contrary, it has been crucial to creating jobs and wealth that simply don’t exist in other countries. Let’s build on what works, rather than tearing it down.

9. *In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?*

There is a critical, fundamental difference between generative AI and previous technologies: generative AI systems are trained on vast quantities of existing copyrighted work, and representations of these works are retained within the AI model. The value of a generative AI system lies specifically in its ability to recreate expressions found in that



work. This is not true of, say, a video camera, or an image-editing program like Photoshop, which by default contain no representations of copyrighted works.

10. *What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.*

I don't think new burdens should be imposed on creators to defend their works from AI training. Burdens like these would be contrary to the long-term policy of U.S. copyright law, which over time has eliminated most procedural formalities for achieving copyright protection. Under current U.S. copyright law, those who want to use copyrighted works have the burden of finding the author and asking permission. That is fair. Some creators may not want to make that easy! That should be their right. I, for instance, would prefer that my work is never used for AI training, and copyright law grants me that right.

11. *Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?*

Probably not. As late as 2022, AI companies tended to see their work as being rooted in both commerce and research. They routinely disclosed details of their training methodology, including the training dataset, to help advance the AI industry as a whole. Starting in 2023, these companies have put up new walls, becoming increasingly cagey and opaque about these facts. Creators and artists are not asking AI companies to be subject to different principles of fairness than, say, food or pharmaceutical companies in terms of disclosing the ingredients that go into their products. But the techniques for implementing these principles may necessarily be different, given some of the unique features of generative AI technology. I do believe, however, that regulations and laws need to urgently clarify that these companies must be transparent, they must be thorough with their recordkeeping, and they must publicly share records of their training data.

12. *Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?*

Yes, I have reviewed this document. In general, it seems the U.S. Copyright Office is applying the human-authorship requirement to AI-generated works in a way that is conceptually and legally consistent with historical practices and U.S. law. (Bearing in mind, as the Copyright Office notes, that "AI-generated works implicate other copyright issues not addressed in this statement," such as the possibility of copyright infringement during the process of training, deploying, and operating these systems.) I believe the Copyright Office has made a correct decision that will protect human creators and the jobs they rely upon.



- a. *Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?*

The U.S. Copyright Office’s registration guidance says that “[w]hen an AI technology determines the expressive elements of its output, the generated material is not the product of human authorship.” I thoroughly agree with this conclusion, and hope this basic idea, that authorship belongs to humans, becomes the foundation of any further action or policy.

13. *Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?*

Yes, I have participated in discussions with the U.S. Copyright Office as a speaker for their listening sessions in the Visual Arts category. I personally found the U.S. Copyright Office to be wonderful and professional. I do have a few suggestions for the U.S. Copyright Office that would be helpful to the discussions and also helpful to creative professionals like myself.

For starters, I believe that every governmental conversation and/or exploration concerning generative AI should represent all viewpoints fairly, and not be tilted in favor of industries poised to materially benefit from generative AI. For example, at my recent panel appearance at the Copyright Office, three artists (including myself) provided one perspective, while two generative AI industry representatives and three lawyers all spoke in favor of generative AI. While I think the listening sessions still went well, I believe better care in who gets to speak in these events should be taken. For instance, inviting more technical and academic experts—who have not been employed or funded by AI companies—to have a voice in these discussions and help determine policy. Further, those whose data is the subject of training (for example, artists, coders, filmmakers) and those who may be the subject of training (for example, educators, healthcare workers, etc.) should be represented in these discussions and have a role in determining regulatory policy.

I would also love to see the U.S. Copyright Office streamline copyright registration by making batch registration possible for artists. Currently artists like myself do not enjoy the same privileges as photographers, who are able to register mass quantities of their work all at once. Making this small shift would immediately benefit my artist community, and allow us to gain affordable access to the additional legal protections that copyright registrations provide.



14. *Can you describe some of the challenges that you face as a working artist and how you expect generative AI technology to impact those challenges?*

Generative AI would fundamentally change the lives of working artists. Consider generally the process and steps it takes to deliver finalized personal visuals to a client: A client describes what they wish to see. I, as the artist, send a variety of sketches to showcase potential visual avenues. The client shares their thoughts on the sketches, and we make adjustments until we feel it's ready to serve as the foundation for the final version. Once time is spent polishing, the final version is sent to the client. Generative AI replaces the artist in every one of those steps. It essentially eradicates the need for an artist. It does that work faster, cheaper, and delivers more quantity than any human artist could ever do. And all of this is based on exploiting the artists' work in the first place as training data.

It's important to note that artists have different industries they make a living in, each with their own challenges. I speak as an artist who works specifically in the commercial entertainment industries. Working as an artist in any industry, let alone the entertainment industry, is already an immensely challenging endeavor. While the entertainment industries are the only ones where an artist can find stable and consistent employment, these industries are notoriously difficult to get into and with good reason. For starters, the number of jobs available are already very small. Furthermore, the high level of technical ability required to gain these jobs, the time needed to make the necessary connections to enter these industries, and the high demands of these jobs, all contribute to the industry's difficult reputation. Due to the specialized nature of some of these jobs, wages do tend to be high (though they have gone down throughout the years). Because of the nature of our jobs, as described above, I see generative AI to be in a unique position where it can take the few jobs that are already available.

I expect generative AI to massively impact our industries. I could see a future where there is only one visual director editing the errors out of generative AI in a production, where normally there would have been a team of dozens of artists.

If left to proliferate, generative AI will make an already difficult living impossible. I genuinely do not see how my industry will survive.

15. *Do you feel that the value of your work has been negatively impacted by the advancement of these new generative AI systems? If so, how?*

I am proud of the work I do. I have trained my entire life to be able to have the technical ability and understanding to create art the way I do. I have enjoyed a successful career with many accolades. I never once doubted my future as an artist—until now.



Generative AI has affected the value of every single artist I know, in foreseen and unforeseen ways. The skills we worked so hard to gain have now been automated. In a cruel twist of fate, that automation was only possible because our own works were taken—without consent, credit or compensation—to train generative AI systems. With generative AI, at the click of a button you can make digital replicas of artists, you can also generate countless images at quantities never before seen, and you can do so via a low monthly subscription. No singular human artist can ever hope to compete.

Because of this exploitative novelty, artists, myself included, have experienced a sudden shift in our industry. We have a harder time negotiating livable wages when there’s cheap software out there that can do our jobs. We have quietly seen an almost immediate devaluation of the skills we offer. Although no official studies have been made, reports of generative AI’s impact amongst peers at all levels in our industry continue to surface. For instance, I’ve heard accounts of legendary veterans in my industry asked to paint over AI generated imagery, essentially asking some of the best artists of our time to no longer paint the evocative imagery countless have seen and simply fix a visual error here and there for less pay. I’ve also heard accounts of students losing out on internships because the position was now utilizing AIs, entry-level jobs being canceled, regular clients suddenly disappearing, usually busy times now being dry.

While the value of an artist is immeasurable, while we are all proud and confident in what we do, market forces simply do not agree. Especially with such a formidable cheaper and faster alternative to the costs of employing an artist. Again, if left to proliferate as is, being an artist may be something only a select few will ever be able to do.

16. *While style is not protected by copyright, do you believe that allowing for style to be copied and for users of AI systems to include artists’ names in prompts results in a greater chance that the output will be substantially similar to an ingested work and potentially infringing?*

First—we should be clear that “substantial similarity” is not an element of copyright infringement. (See *Range Rd. Music, Inc. v. East Coast Foods, Inc.*, 668 F.3d 1148, 1154 (9th Cir. 2012).) Rather, it is a special evidentiary rule that helps courts determine if copyright infringement occurred in cases where a work *was not* directly and entirely copied. But when a work is directly and entirely copied—for instance, during the training phase of a generative AI system, where millions or billions of works are directly and entirely copied—then a finding of copyright infringement is automatic. Because a generative AI system can only “remix” expressive elements from the training data, there is a plausible argument that every output of that system infringes on the copyrights of the owners of the training data. Though some reject this idea as being impossibly broad, it is consistent with a quantitative account of how these systems work.



Second—style is indeed not copyrightable. But we should be careful to distinguish the two meanings of the word “style”. In one sense, the word “style” denotes general moods, art movements and approaches—such as “realistic”, “surreal”, “fantasy art” and so on. As an artist, I have learned from the style of other artists, and I hope others learn from mine. But the other meaning of the word “style” refers to how I, as an artist and individual, uniquely approach and create *my* work. The only reason the name “Karla Ortiz” within a prompt means anything to a generative AI system is because it has been trained in part on *my* work—without consent, credit, or compensation. Furthermore, certain AI companies have promoted the ability of their systems to generate works in the style of certain artists, such as myself. Even if my style isn’t copyrightable (though my work certainly is), I ought to have a right to prevent my name from being commercialized in this manner—which has also happened without consent, credit, or compensation. Thus, as a complement to copyright law, I hope Congress will consider a federal right of publicity law that will help artists and creators ensure that their names and reputations are not misused by AI companies.

**Questions from Senator Tillis
for Dana Rao**

**Witness for the Senate Committee on the Judiciary Subcommittee on
Intellectual Property Hearing “Artificial Intelligence and Intellectual
Property - Part II: Copyright”**

1. **Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?**

The United States has led the world through technological transformations in the past and we have all learned that it is important to be proactively responsible about the impact these technologies can have on us. Pairing innovation with responsible innovation will ensure that AI ultimately becomes a transformative and true benefit for society.

At Adobe, we believe a comprehensive analytical framework for encouraging responsible AI development should consider:

- the impact of unfair AI bias on high-risk uses of the technology;
- the importance of ensuring access to data to spur the growth of AI;
- the impact of AI on creators and IP;
- the economic impact of AI on productivity and job creation; and
- mitigations against the spread of deepfakes.

We believe AI regulation should appropriately categorize AI systems as high risk and low risk and focus review on those high-risk systems. Examples of high-risk areas include AI systems that make decisions about health, employment, housing or finance. Any regulation or review should be conducted on a use case basis (versus categorizing AI as general purpose and applying the same standards regardless of use case). This will ensure that we are focusing on mitigating risk for those high-risk use cases, without unnecessarily slowing down innovation, so companies can continue to develop cutting-edge technologies and maintain U.S. leadership in the global technology industry.

Additionally, transparency in digital content is critical, especially in the age of AI. Congress should support provenance solutions and standards for providing consumers with more transparency about the content they are consuming – such as the Content Credentials technology and its underlying [C2PA](#) standard. Congress should also require any product or platform that receives content with provenance metadata attached to preserve and display that metadata (and not strip it away).

And as discussed in our written testimony, AI is trained on data, and training on larger datasets helps ensure the AI is more accurate and less biased. This is why we have seen other jurisdictions such as Japan, China, the United Kingdom, and the European Union recently taking steps to liberalize copyright laws and creating text and data mining exceptions specifically for AI/ML training. However, one of the important implications of AI's need for data is the impact on copyright and creators' rights.

We believe there is a way both to protect creators' rights and also to ensure the access to data that AI requires. The U.S. must balance the two to maintain a long-term leadership position in this space.

To protect creators, we encourage government and industry to work together to:

- provide artists a "Do Not Train" tag to allow them to opt out of training
- provide artists a way to secure copyright and attribution in a world with AI-assisted works
- establish a new federal right to allow artists to protect themselves against people misusing AI to impersonate their style
- and help advance new economic opportunities for creators enabled uniquely by AI.

Pairing these creator protections with support for access to data for AI research will ensure AI innovation continues to develop in the right way, for everyone.

2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?

Yes. At Adobe, we believe transparency in digital content is critical, especially in the age of AI. This is the mission behind the Adobe-led [Content Authenticity Initiative](#) (CAI). The CAI is a group of more than 1500 members from across industries working to help enable creator attribution and fight misinformation through open standards for provenance technology and supporting tools like Content Credentials. Content Credentials allow creators to attach important information to a piece of content like their name, date, and what tools were used to create it. That information travels with the content wherever it goes. For generative AI, Content Credentials can show you whether a piece of content was human-created, AI-edited or AI-generated. In Adobe's own AI image-generating model, [Firefly](#), we automatically attach Content Credentials that indicate that content was AI-generated. This level of transparency in digital content will help consumers make more informed decisions about whether to trust the content they see online and help create a more trustworthy digital space.

3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?

- a. Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available?

b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

At Adobe, all of our AI products and features are developed in accordance with our [AI Ethics principles](#) of accountability, responsibility, and transparency. As part of our commitment to transparency, we share the types of datasets used to train the generative AI models we build and release in a commercial product. For example, the first version of Adobe Firefly was trained on licensed images from our own Adobe Stock photography collection, openly licensed content, and public domain images where copyright has expired.

We understand that practically speaking it can become very difficult to understand what datasets are used in commercially licensed models or open-source models, or datasets that are an amalgamation of other datasets. In addition, AI models may be trained to dynamically update with user interaction data, in which case the complete source of training data will be difficult to quantify or characterize. Accordingly, we suggest the requirements of disclosure may be best limited to references to named datasets (if any), or accurate characterization of the types of data on which the model is trained (e.g., user data, data licensed under a particular license agreement, copyrighted data, data in the public domain, etc.).

4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

Under US Copyright law, you cannot copyright an idea. What you can copyright is the expression of an idea. For example, you cannot copyright the word “book” – that is an idea. But you can copyright a painting of a book because that’s an expression of the idea. When you apply this principle to image-generating AI, it means that a prompt may not be copyrightable because the prompt represents the idea, and the output is based on the AI’s interpretation of that prompt. When you type in “cat driving a 1950s car through the desert”, the AI decides whether it’s a Siamese cat or a Tabby cat; it decides whether to make the car a convertible or a pickup truck; it chooses the color of the sky, the number of cactuses in the desert and so on. The AI is expressing the idea, not the human. Copyright law is designed to protect the rights of human creators – therefore an AI output (the AI’s expression of the idea) may not be copyrightable.

However, most creators will not use the raw, unmodified AI output as their final creation. Many creators are ideating and brainstorming in a generative AI tool and then adding their own style and expression to an image. Some artists are using AI just to quickly change the sky color of their artwork, rather than laboring through a tedious pixel-by-pixel manual task. Take our cat in a car in the desert example. Maybe the artist starts with the AI-generated output but then uses other

non-AI tools to change the color of the car, add some mountains in the desert, add another cat in the passenger seat, and so on. Or, maybe they already created the image of the cat in the desert using non-AI tools and they used AI to simply swap out one of the cactuses for a desert tree. In these instances, the human creativity and expression they are adding to their work should be copyrightable, but they will need to prove what was AI-generated and what was human-created. Fortunately, there are tools (such as any that implement Content Credentials) that will allow a creator to distinguish the portions of a work created by AI versus the portions of a work expressed by a human.

Currently, we do not believe that the text prompts typically used are of sufficient specificity to approximate the expression of an idea – they are more akin to an idea itself – and therefore may not be copyrightable. However, as the space evolves, it is possible that prompts may take different forms, and could become of specific precision and detail such that parts of the output could be considered the creator’s expression of an idea.

5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?

We believe AI holds vast potential to unlock new opportunities for creators. First, we believe that Generative AI will increase productivity by automating tedious or repetitive tasks, allowing creators to do in seconds what used to take hours and increasing the creative output of every creative professional. Second, by making creating easy and fun, billions of latent creators can now become creative, because the tools are easy to use. As more people create, we believe the demand for professional content will only continue to grow leading to more opportunities for creative professionals. Third, we believe that creativity will be one of the more difficult aspects of human thinking for AI to imitate. In the creative world, we believe consumers of art prefer their art to have a soul and a story along with it in order to create the real connection and meaning that viewers, readers, and listeners want from it. This should mean that the creative professions are in a better place to be safeguarded from AI replacement. All of these points should encourage the United States to invest in creative skills training, to create a workforce equipped with the higher-level abstraction skills needed to differentiate against AI tools.

We also believe the AI business model itself can unlock new economic opportunities for creative professionals, and Adobe is exploring ways to help creators monetize their work in the age of AI. One approach would be to enable creators to license their style directly to consumers. In this approach, people could subscribe to a particular artist, who has worked with an AI model to have it replicate their style. The consumer could then create their own work in that artist’s style, but for a fee. This would be an entirely new revenue stream for artists, and we are currently testing an AI model with creators to do just this.

6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

Questions about copyright infringement should center around the output of a generative AI system and whether a particular output infringes on the copyright of someone whose image was used to train an AI system that generated that output and would depend on very specific facts of the case.

We believe generative AI companies have a responsibility to take steps to mitigate potential infringement, as do users of the tool.

For example, we believe that in cases where someone is intentionally using AI to try to impersonate an artist's likeness or style, the person misusing the AI tool in that way should be held accountable. This type of misuse can cause economic harm to artists. In terms of whether a tool itself should be held accountable, this should typically be decided on a contractual basis based on the user and the provider. In Adobe's case, we are indemnifying enterprise users for the works generated directly by Adobe Firefly. Other AI providers may choose different approaches, and customers should be free to choose the model they like.

To help enable artists to enforce their rights against misuse, we believe Congress should establish a federal anti-impersonation right (FAIR) that would give artists a right to enforce against someone intentionally and commercially impersonating their work through AI tools, without having to rely solely on copyright. This new law should include statutory damages to alleviate the burden on artists to prove actual damages, directly addressing the unfairness of an artist's work being used in a manner that could cause a creator direct economic harm.

7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?

As noted above, in the creative world, we believe consumers of art will want a soul and a story along with the art they are experiencing, to create the real connection and meaning that viewers, readers, and listeners want from their art and the artist that created it. In addition, we believe the evolution of style takes human creativity.

8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?

Placing thoughtful safeguards around AI development and use will help us harness the full potential of AI to benefit society. We believe there are important steps that industry and

government can take to protect creators' rights in the age of AI while ensuring the U.S. continues to lead the world in AI development.

We believe government and industry should leverage the Content Authenticity Initiative's Content Credentials technology. Content Credentials are built on an open standard and have several important capabilities when it comes to protecting creators:

- **Do Not Train.** First, we believe that government and industry should support enabling creators to attach a "Do Not Train" credential in the metadata of their work. This gives them the option to keep their data out of AI training datasets. The Adobe Content Credential enables Do Not Train tags as part of its open standard.
- **Enabling creators to obtain IP for AI-assisted works.** Second, creators using AI tools want to ensure they can obtain copyright protection over their work in this new era of AI assisted digital creation. We believe that AI output alone will not receive copyright protection, but we believe the combination of human expression with AI expression will and should. With Content Credentials, creators can capture the tools they used throughout their creative process and distinguish their own non-AI efforts from the AI expressions in their work. This will give them the proof they need to obtain a copyright.
- **Enabling creator attribution.** Third, it's very easy to reproduce digital content in different forms and in different outlets, and the original author attribution can often get lost or stripped away. Content Credentials allow creators to cryptographically associate their identity with their work so they can get credit for it, wherever it goes.

We also believe it is important for the law to protect artists against the economic harm that could be caused by AI-generated works. We propose that Congress establish a new federal anti-impersonation right (FAIR) that would protect artists from someone using AI to impersonate their work or style. This new law would allow artists to protect their livelihood from people misusing this new technology, without having to rely solely on copyright.

We also know that to create good and unbiased AI, you need a lot of data. Like the human brain, AI learns from the information you give it. If AI has more facts to learn from and a wider breadth of experiences to build its experience of the world around it on, it will be more likely to generate accurate responses that avoid perpetuating harmful biases.

It is important to pair creator protections, like we proposed above, with support for access to data for AI research, whether it is through fair use or other means, so we can ensure AI innovation continues to develop in the right way, for everyone.

9. **In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?**

Like other technological advances in the creative space, such as advances in photography, video cameras, electronic music, etc., consumers are excited about generative AI as a new technology that has the potential to unlock vast new opportunities for creators. Generative AI differs from traditional machine-aided creativity because it allows direct creation of a work typically through a text interface. In the past, such as in Photoshop's own AI tools, the AI features worked on existing content. However, we believe in essence both types of AI are still machine-aided creativity. The creative professional will use the output of generative AI as the first step in their creative process, and will refine the work to match their vision. The Generative AI approach does, however, significantly advance the professional's ability to reach their end state much faster than before. In addition, we expect Generative AI will be used by creative professionals in all kinds of new and innovative ways, pushing forward the fields of innovation and the arts in unforeseeable ways.

- 10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.**

In the digital world, it is very easy to reproduce digital content in different forms and in different outlets, and it is often hard to maintain the source attribution for the original work. The Adobe-led Content Authenticity Initiative is a global coalition working to fight misinformation and enable creator attribution through open standards for provenance technology and supporting tools like Content Credentials. With Content Credentials, creators can cryptographically associate their identity with their work (for example, an image or a piece of digital art), and then if the work is reused or repurposed, their identity will travel with their work across all platforms that support Content Credentials.

To support this approach, Congress should require that all platforms that receive such attributions maintain them, so the attributions are not stripped away and artists can receive credit for their work.

- 11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?**

As governments look to establish standards and safeguards to help guide the responsible development of AI, they should begin by leveraging existing frameworks, such as the NIST AI Risk Management Framework. Adobe collaborated with other industry leaders to help develop this framework. We believe AI governance should be global in nature and build upon industry best practices.

This framework outlines the characteristics of trustworthy AI, which includes transparency "about an AI system and its outputs." Adobe strongly agrees with the need for transparency

when it comes to AI. As stated in our AI Ethics Principles, and as discussed above, transparency means we are open about how we use AI and the types of datasets we use to train our AI. We are committed to working together with our community to design and implement AI that respects our customers.

12. Have you reviewed the U.S. Copyright Office’s Registration Guidance for “Works Containing Material Generated by Artificial Intelligence” and, if so, what are your views on the guidance?

a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?

Adobe has reviewed the U.S. Copyright Office’s Registration Guidance for “Works Containing Material Generated by Artificial Intelligence” and we are aligned with the approach the Copyright Office has taken. Adobe Senior Director, IP & Advertising Law, J. Scott. Evans, recently participated in a [U.S. Copyright Office AI listening session on visual arts](#), where he shared Adobe’s perspective on copyright in the age of AI and subsequently shared our perspective in this [blog post](#) for the Copyright Alliance.

13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?

Yes. Adobe Senior Director, IP & Advertising Law, J. Scott. Evans, recently participated in a [U.S. Copyright Office AI listening session on visual arts](#), where he shared Adobe’s perspective on copyright in the age of AI. We are encouraged to see these offices seeking feedback and taking a collaborative approach and Adobe is happy to continue to share our perspective and provide feedback in these ongoing discussions.

14. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data.

Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

We do not feel qualified to speculate about what other companies may or may not be doing.

15. Can you explain the thinking behind Adobe’s decision to adopt an opt-in approach to the use of copyrighted works as training data?

a. Are you concerned that this approach will put Adobe at a competitive disadvantage with those who do not see the need to obtain consent?

In designing the first model of our own family of generative image AI models, Adobe Firefly, we chose a path that supports creators and customers by training our model only on licensed images from our own Adobe Stock photography collection. This provides us with a dataset that is designed to minimize exposure from legal concerns.

However, it's important to remember AI is only as good as the data on which it is trained, which is why the question of data access is important to companies and organizations building foundation models. Like the human brain, AI learns from the experiences or information you give it. And like the human brain, the more information you give it, the better it will perform. An AI system trained on a small dataset is more at risk of producing wrong or unsatisfactory results, or reproducing harmful biases that exist within the dataset.

If you had never been taught what a car is, it would be hard to accurately depict one or answer a question about what it is or what it does. To produce accurate results, AI needs a large dataset representing the universe of possible answers to learn from. Additionally, a narrow dataset can lead to unfair bias. If you have only ever been taught that lawyers are men, you are likely to conjure up an image in your head of a man when someone is talking about a lawyer, even though more than half of the graduates of law schools are women. AI works the same way. Training on a larger dataset can help ensure you capture a broader set of perspectives in the data itself, so that when you type in "lawyer," you will see a result set that reflects the society in which you live.

We believe that designing the model this way mitigates the risk of infringing on someone's intellectual property, which allows enterprises and individual creators to create confidently with it. However, because we trained on a narrower dataset rather than scraping the web, our approach required extra engineering effort, testing, and mitigation practices to ensure accuracy and reduce bias in Firefly outputs. We encourage the government to support access to data to ensure that AI innovation can flourish both accurately and responsibly especially in areas where a large dataset of licensed data is not readily available.

16. Some of the most popular generative AI tools were built by indiscriminately scraping material – including copyright protected material – from the Internet. Can you explain Adobe's approach to building its AI tools and how it's different than other companies?

Adobe recognized the various unanswered legal questions around access to data in designing our own family of generative AI image models, [Adobe Firefly](#), which we launched in March 2023. We chose a path that supports creators and customers by training on a dataset that is designed to be commercially safe, where users have protections against potential legal liability that could be caused by other types of training, as the law on what is permitted is not currently decided.

We trained our first Firefly model on licensed images from our own Adobe Stock photography collection, openly licensed content, and public domain images where copyright has expired. To help ensure copyrighted or branded materials are not created as part of Firefly's output, we have a content moderation team that performs extra filtering on the images before they become part of the Firefly dataset. By designing the model with licensed content, the resulting model is more insulated against legal exposure from potential copyright claims.

- 17. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model.**

Could you please explain how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

Adobe used licensed data, and not any unauthorized data, to train the first model of Firefly, our text-to-image generative AI tool.

- 18. Some have suggested different licensing structures for compensating copyright owners for the use of their works in AI training. What licensing structures have you seen or used that have worked to the mutual benefit of both AI companies and copyright owners?**

Adobe trained the first model of Firefly, our text-to-image generative AI tool, only on licensed images from our own Adobe Stock photography collection. However, it is still in early days of addressing the question of creator compensation and we have not yet seen a license structure announced that appears to address this. However, as we discussed in our testimony, we believe that we do need to enable access to data while giving creators a way to address the economic dispossession caused by a person misusing AI to impersonate their work. We believe this model will most directly address the harm caused by AI impersonation.

Answers to Questions

Arising In Relation to the U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property Hearing on “Artificial
Intelligence and Intellectual Property – Part II: Copyright and
Artificial Intelligence” held on July 12, 2023

Matthew Sag
Professor of Law in Artificial Intelligence, Machine Learning,
and Data Science
Emory University School of Law

Chair Coons, Ranking Member Tillis, Members of the Subcommittee:

Thank you for the opportunity to answer the following Questions for the Record.

Question 1. Given generative AI is developing all over the world and countries are responding to it in different ways, are there policies or regulations being adopted elsewhere that you recommend that the U.S. consider or avoid?

(1) In relation to authorship, there is a strong international consensus that authorship requires some kind of subjective intention to manifest or communicate a belief or a state of mind that is entirely lacking in current and foreseeable computer technology. No changes are required in U.S. law at this time.

A notable exception to this international consensus is Section 9(3) of the UK Copyright, Designs and Patents Act 1988. This provides that for “computer-generated” works, (meaning a work generated by a computer “in circumstances such that there is no human author”) “the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken.” Section 9(3) was not drafted in contemplation of generative AI and the provision is generally regarded as not particularly useful in that context, or any other.

In fact, the section shows some of the risks in trying to make computer generated works copyrightable. The phrase “the person by whom the arrangements necessary for the creation of the work are undertaken” is deeply ambiguous in the context of generative AI: it could refer to the authors of the works in the training data, the person or persons who selected the training data, the person or persons who trained the model, the person or persons who provided fine-tuning feedback in a process of reinforcement learning through human feedback, the person or persons who provided the necessary computing resources to train the model or to run the model in response to a user query, or to the person or persons who

wrote the prompt. This is only the beginning of the uncertainty, but I believe I have made my point.

(2) In relation to the use of copyrighted works and training data, the United States' world leading position in AI research is partly attributable to the flexibility afforded by the fair use doctrine. Given the pace of technological change, there may be some wisdom in continuing to allow the courts to apply the fair use doctrine to particular situations, rather than prescriptively legislating in more detail. The track record of other jurisdictions trying to legislate for issues we deal with through fair use is mixed, at best. Attempt to write technology specific rules often fail to predict the past, let alone the future.

On the other hand, business and research communities may benefit from an express copyright limitation that allows for text data mining, similar to Article 3 and Article 4 of the EU Digital Single Market Directive. I have reservations about this approach if it displaces fair use and I anticipate that any "clarifying" legislation will nonetheless require significant judicial clarification.

However, the U.S. could benefit from a clear safe harbor for the use of copyrighted works as training data for machine learning. I address this in my answer to question 8, below.

Question 2. A recent survey on how consumers view AI found that most consumers – nearly 80% – believe the use of AI should be explicitly disclosed. Do you agree? Why or why not?

In principle, it is hard to disagree with calls for disclosure and transparency. However, the devil is in the details.

It seems simple enough to propose that images generated by an AI without any significant human creative import should be labeled either through water marketing, metadata, or some other description as "AI generated."

However, we need to be clear on why such labeling is important and how critical terms will be defined.

- The line between AI-generated and human-generated may be difficult to draw when a person uses and then edits AI-generated text, or when a person uses AI-powered editing tools to manipulate a work that was initially human-authored.
- In some contexts, disclosure and labeling may be important because we want to assess whether the work should be considered as creative or expressive. Accordingly, the threshold for labeling something as AI-generated might be quite high.
- In other contexts, labeling may be important because the public needs to know whether the content they are being presented with has been manipulated, or even entirely manufactured. If a news report features an image of the Pope in a white puffy

jacket, the tools used to create the image are far less important than the fact that the image is fake. Accordingly, any manipulation of the image or text should be disclosed.

My point is that the reasons we want disclosure need to align with the details of any proposed requirement, and that there may be tradeoffs between different objectives.

3. What are the benefits and disadvantages of requiring an AI company to keep records of everything that is ingested and to make those records publicly available?

a. Under what circumstances, if any, should an AI company NOT be required to make its records of everything that is ingested by the AI publicly available?

b. Under what circumstances, if any, should an AI company be required to make its records of everything that is ingested by the AI publicly available?

The term “ingested” is imprecise and creates unnecessary ambiguity.

Machine learning models “learn” by exposure to the training data. The training data influences the model, but it does not become part of the model, except in rare cases of overfitting (usually due to a failure to effectively deduplicate the training data). If Congress legislates in relation to “ingestion,” technology companies will plausibly respond that the models do not ingest anything.

Record-keeping:

Any entity that relies on the fair use doctrine to assemble a significant quantity of copyrighted works for use as training data for machine learning should be required to keep accurate records that identify those copyrighted works and their provenance.

For each model, those records should also keep track of whether the work was ultimately used in training. Some works initially copied for the purpose of training may never actually be used because they are duplicates of other works, or because they have undesirable properties (for example, the inclusion of pornography or hate speech).

Any record keeping provision should apply to all works copied as part of the potential set of training data, not just those works actually used to train the model. Depending on the work, maintaining accurate records may require archiving a permanent copy of the work. Identifying a work by title and URL, may not provide sufficient information.

Disclosure:

(a) Any entity subject to the recordkeeping requirement suggested above (“a covered entity”) should be required to publicly disclose summary information about the composition of copyrighted works in the training data associated with each new publicly released or commercially significant model in a timely fashion. The recordkeeping requirement should not apply to works included with the express permission of the relevant copyright owner, works in the public domain, etc.

(b) Covered entities should be required to disclose examples of specific works in each relevant category that were and were not included in the training data.

(c) Technology companies and researchers may have a legitimate interest in not publicly disclosing the entire contents of the training data. The composition of the training data could be a valuable trade secret in some circumstances. Moreover, information about the training data combined with other information could disclose personal or sensitive information about individuals in some contexts. Covered entities that do not disclose the entire contents of their training data to the public should be required to provide a mechanism to allow individuals to easily determine whether their works were part of the training data for a given model.

(d) Covered entities should be required to make the training data available for auditing by the Copyright Office, or some other government agency. The purpose of the audit would be to determine the adequacy of the summary in (a) and the disclosure mechanism in (c).

Regulatory burden:

The recordkeeping provision proposed above would impose a minimal burden on technology companies and researchers, because the steps outlined above should already be part of any data workflow. The disclosure provisions proposed above need to be carefully considered to ensure that they are reasonable and proportionate. To the extent that such provisions apply to university and not for profit researchers, they should be calibrated to that environment. The Copyright Office could be delegated authority to define exemptions to recordkeeping and/or disclosure requirements.

Question 4. Do you think that generative AI prompts provided by users are copyrightable? And if so, under what circumstances could they be copyrightable?

A user prompt *could* be copyrightable in some specific cases, however, in practice most user prompts will not be copyrightable.

Much like computer software, a prompt could be copyrightable as a literary work if it is original, if it contains a non-trivial amount of creativity, and if it is not simply an uncopyrightable idea or instruction. The mere fact that a literary expression is used as a prompt does not make it ineligible for copyright. However, most prompts will lack sufficient

originality and creativity beyond their functional attributes and will in practice be ineligible for copyright protection.

a. Do you think that whether the prompt used is copyrightable or not should impact the copyrightability of the resulting AI output generated as a result of the provided prompt?

The copyrightability of a prompt does not itself determine the copyrightability of the content generated by an AI when the prompt is invoked. The copyrightability of an AI generated image depends on whether the image reflects a person's original conception of that image in sufficient detail. A prompt may be independently copyrightable as a literary work but entirely fail to produce an image that reflects the prompt-author's conception. For example, a copyrightable haiku could be used as a prompt, but the resulting image would not be copyrightable.

It is important to understand that the relationship between words typed into the context window of a generative AI and the output of that AI is quite different to the relationship between instruction and output in other contexts. When a composer writes a set of instructions in the form of sheet music, she does so with a very specific output in mind. Even though performers add a layer of creativity, subjectivity, personal interpretation in implementing sheet music instructions and producing musical sounds, there is a very tight correlation between the work envisaged by the composer and the work ultimately performed by a performer following the sheet music. In contrast, when a user inputs a prompt into a text-to-image generator, such as Stable Diffusion and Midjourney, she often has no specific expectation of the appearance of the resulting image. The process is deterministic in that the same prompt will result in the same image if the initial state of the program is the same, but from the user perspective, (1) specific outputs are generally so unpredictable as to appear random, and (2) platforms appear to be designed such that the initial state of the program is never the same. The same is true with ChatGPT, you never really know what the answer to a prompt will be, and it is never exactly the same because the system varies the initial state every time.

The copyrightability of AI generated content does not depend on the copyrightability of the instructions used to generate that content. As I explained in my written testimony:

Generative AI is often used as a tool in the creative process. A person who instructs a Generative AI with enough detail, such that model output reflects that person's original conception of the work, should be regarded as the author of the resulting work. However, simple text prompting is unlikely to meet this standard.

... there is no reason in principle why prompts couldn't be detailed enough to meet the traditional threshold of authorship in some cases. Sophisticated prompts that specify details of an image should be sufficient to meet the requirement that the work that results from and reflects a person's original conception of the expression.

Furthermore, refining text prompts and choosing between different outputs should also be recognized as way in which a human using Generative AI could meet the authorship standard.

Question 5. What does the impact of generative AI have on the creative industry? Specifically, what are your thoughts regarding the concern that the proliferation of generation AI will take over jobs?

Generative AI has enormous potential to make creative people more productive and to allow more people to express their creativity. By reducing the cost of creation, generative AI will enable individuals and companies to do more with less — whether that implies an increase in creative production or any decrease in employment in creative industries is a difficult question to answer in the abstract. I am not an economist or an industry specialist, but I have studied economic history and how the law responds to and shapes new technology for the past 20 years.

In the early days of the World Wide Web (in the mid-1990s) it was easy to foresee how digitization and distributed networking could disrupt existing business models. However, most of the new opportunities and new business models the web made possible were unforeseen, and I suspect, unforeseeable. This is a recurrent story with disruptive technology. The jobs that might be lost loom large because they are foreseeable, the new jobs that the technology makes possible seem like optimistic speculation because we don't know with specificity what they will be. I don't wish to trivialize the legitimate anxieties professionals in the creative industries have in relation to AI, just to place them in a broader context.

The people whose jobs are most at risk from generative AI are those that produce content that is easily fungible with other content. Three obvious examples would be: stock photography, cover art, and narrative content for search engine optimization. In contrast, work that is customized, work that is part of a long-term relationship, or work that is valued because of its connection to the personality of a particular author does not appear to be at risk.

In answering this question, I have assumed that Congress will ensure that individuals are given adequate protection from the use of AI to generate synthetic content that recreates their voice, image, or likeness (i.e., deepfakes). One of the things that emerged most clearly from the July 12 hearing was that some of the most pressing concerns people have in relation to generative AI have nothing to do with copyright, and everything to do with the increased capacity of deepfake technology.

Question 6. If a generative AI system is found to infringe a copyrighted work, who should be liable for the infringement – the AI company, the user providing the prompts to the AI tool, or both?

Before I answer this question, it is important to note that AI produced content that infringes copyright is extremely rare. If we apply our traditional tests of infringement and seek to identify “substantial similarity” between the outputs of generative AI models and the copyrighted works used to training those models, we will almost inevitably come up short. The reason for this, as I explained in my written testimony, is that generative AI models ‘usually learn from the training data at a fairly abstract level. Moreover, the output of generative AI usually combines abstract latent features learned from the training data in a way that ensures model outputs look nothing like specific model inputs.’”

The output of generative AI is most likely to infringe on copyrightable characters, where, practically speaking, the level of similarity required to establish infringement is more abstract.¹ Setting that issue aside, infringing output may result from (1) overfitting, usually a product of failing to effectively deduplicate the training data, or (2) from user instructions.

Who is liable for isolated instances of infringement that result from using generative AI will depend on whether courts apply the volitional act requirement in this context. Although copyright infringement does not require a particular mental state—you can infringe copyright by mistake, or even by subconsciously copying—it is widely assumed that a defendant may be held directly liable only if it has engaged in volitional conduct.² Technology companies will plausibly argue that although they provide the infrastructure to create images/music/text, if the user who actually prompts the system to create the output is the one who “makes the copy.” Understandably, courts have not addressed the application of the volitional act requirement in the context of generative AI, and it is arguable that this context should be distinguished from the automated copying and transmission systems where the volitional act requirement has been applied in the past.

If the user is treated as the one who “makes” the offending copy, then under current law the technology provider will only be liable if the requirements of the vicarious, contributory, or inducement-based liability are made out.

If the technology provider is deemed to “make” the offending copy, it is essentially subject to strict liability. Congress may wish to consider whether some intermediate standard is desirable, such that technology providers have an obligation to take reasonable measures to

¹ For a more detailed explanation, see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: <https://ssrn.com/abstract=4438593>)

² In *Religious Technology Center v. Netcom On-Line Communication Services, Inc.*, 907 F. Supp. 1361, 1370 (N.D. Cal. 1995), the district court held that the defendant Internet service provider was not liable for the automatic reproduction of a copyrighted work by its computer system. The court refused to impose direct liability on the service provider, reasoning that: “Although copyright is a strict liability statute, there should still be some element of volition or causation which is lacking where a defendant’s system is merely used to create a copy by a third party.”

prevent infringement, but are not held liable for the independent choices of the platform users.

Regardless of who is liable for individual instances of infringing output, the existence of infringing output may have implications for whether the fair use defense applies to the assembly of the training corpus in the first place. As I explain in a forthcoming Law Review article:

If ordinary and foreseeable uses of generative AI result in model outputs that would infringe on the inputs no matter what intervening technological steps were involved, then the non-expressive use rationale no longer applies. If training LLMs on copyrighted works is not justified in terms of non-expressive use, then there is no obvious fair use rationale to replace it, except perhaps in the non-commercial research sector. If LLMs just took expressive works and reconveyed that same expression to a new audience with no additional commentary or criticism, or no distinct informational purpose, that would be a very poor candidate for fair use.³

Question 7. In your opinion – currently or in the foreseeable future – can AI generated material ever replace the quality of human created work?

The literal answer to this question is, yes. We have already seen examples of AI content that has been adjudged to be as good as human authored work.

However, I believe the spirit of the question is really about whether there is something special or significant that should make us regard human authored works more highly than AI generated content. In some cases, the answer is clearly yes, in much the same fashion as we often regard an original work of art as more desirable than a copy. But in many cases, there is nothing intrinsically special about human authored content.

Question 8. A balance needs to be struck in terms of how to encourage innovation, how to be responsible, and how to ensure that there is clarity for all using this technology. How do you propose we do this in the copyright space in a way that allows the U.S. to stay competitive and remain the global leader?

The fair use doctrine already gives U.S. technology companies and researchers a substantial advantage over their peers in many other developed nations. Courts in the United States have a strong track record of applying the fair use doctrine in a way that balances innovation with respect for the interests of copyright owners. Properly applied, the fair use doctrine allows for technical acts of reproduction that do not interfere with the copyright owner's interest in controlling the communication of their original expression to the public. The

³ Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: <https://ssrn.com/abstract=4438593>).

courts were correct to rule that peer-to-peer file sharing was not fair use; they were also right to find that Google Books and HathiTrust were fair use.

The U.S. could supplement the fair use doctrine by establishing a safe harbor regime for non-expressive uses, without prejudice to the general application of Section 107 of the Copyright Act. The application of the safe harbor could be conditioned on taking certain affirmative steps to protect both copyright and non-copyright interests of authors and copyright owners of works used in training data. Any such safe harbor should be optional, not compulsory, to avoid First Amendment entanglement. Key requirements for the safe harbor should be modeled on the best practices for deduplication and avoiding overfitting that have already been identified in the computer science and legal literature;⁴ but they must also provide room for the development of new best practices that will doubtless emerge. The safe harbors could be designed to protect interests that relate to right of publicity and trademark related concerns.

Question 9. In the copyright context, what differentiates the technology of generative AI from other machine-aided creativity, such as photography, video cameras, electronic music, and the like, all of which allow the public to develop and advance knowledge?

There is a long history of technologies that have enabled new forms of creativity or reduced the cost or skills required to engage in creativity, and each one has disrupted existing market structures. Generative AI as part of this tradition, but there are some important distinctions to be drawn.

Authorship, originality and ownership:

In traditional machine-aided creativity, such as photography or electronic music, the authorship of the tool-user is rarely questioned. Although operating a camera take less skill and training than painting with oil on a canvas, courts have long recognized that the combination of minute anesthetic decisions involving framing, timing, lighting, positioning, etc. are enough to make the photographer the author. In contrast, as discussed above and in my written testimony, much of the content produced by generative AI does not meet the authorship standard and is thus uncopyrightable.

Reproducibility and Volume:

AI can generate vast amounts of new, seemingly creative content at a speed and volume far beyond human capacity. In contrast, traditional machine-aided creative processes are constrained by human capabilities and time.

⁴ For an introduction to this literature, see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023)(Available at SSRN: <https://ssrn.com/abstract=4438593>).

Copyright law is premised on the fact that information goods are expensive to create but cheap to copy. Generative AI may undermine that premise by making novel information goods cheap to create. I don't think this makes copyright obsolete, but it does explain why the uncopyrightability of generative AI content is no cause for concern.

Question 10. What steps can and should the creative community take today to ensure that their work is more easily attributed to them, regardless of whether their work is used for training an AI model? For example, indicating authorship and contact information via the metadata of the author's digital content.

Creative communities may need to revisit the contractual terms under which their work is distributed. Many open-source and creative commons licenses implicitly allow for works to be used in machine learning training, but these contracts were not drafted with this scenario mind. The users of social media platforms and cloud hosting services may also be surprised by the extent to which they have already agreed to allow their photos, videos, music, and social media posts to be used to train machine learning algorithms. The same goes for any creative professional who contracts with an aggregator, such as a stock photo agency.

Persistent metadata about signals appropriate and inappropriate uses of a work would be advantageous, but I believe other witnesses are better placed to address this issue.

Question 11. Are existing laws and regulations sufficient to deal with the issues relating to transparency and record keeping by AI companies?

No.

I am not aware of any laws or regulations that require companies developing generative AI tools to disclose the details of which copyrighted works were used in training, or even to keep a complete and accurate record of those works.

For a proposal for such a requirement, see my answer to question 3.

Question 12. Have you reviewed the U.S. Copyright Office's Registration Guidance for "Works Containing Material Generated by Artificial Intelligence" and, if so, what are your views on the guidance?

a. Do you think that the Copyright Office got it right? Are there aspects of the guidance that could stand to be clarified or revised?

Please refer to Appendix A of my written testimony, "When Should A Human Be Credited With Authorship Of Something Created Using Generative AI?"

Question 13. Both the U.S. Patent and Trademark Office and the U.S. Copyright Office have engaged in extensive outreach regarding AI. Have you participated in this outreach and, if so, how did you find it? What more can and should these offices do?

The USPTO and the Copyright Office deserve recognition for their outreach in relation to the intersection of AI with intellectual property.

I believe it would be beneficial for one of these agencies to convene a working group to suggest best practices for generative AI.

Although copyright infringement, trademark infringement, and interference with privacy and personality rights, do not pose the same existential risk as Skynet or an out-of-control paperclip factory, these copyright and copyright-adjacent risks are foreseeable in relation to generative AI. These risks are also a lot more likely.

Like other issues in AI safety, addressing the potential for copyright infringement and other related harms will require technical solutions informed by legal, ethical, and policy frameworks. The USPTO or the Copyright Office could make a significant contribution without the need for additional legislation by exploring options for continuing the development of generative AI while reducing potential harms and adverse impacts.⁵

Question 14. Language Learning Machines are increasingly being used to generate source code and help software developers write software. Such models can require a vast amount of source code and thus can turn to open-source software (OSS) for scraping publicly available source code.

a. If AI models are trained on OSS, does that infringe on the copyright of the respective authors?

b. If an AI model is used to generate code, does that generated code constitute a derivative work? And if that AI model was trained on copyleft-licensed OSS, must it also be licensed under copyleft?

There is a very close relationship between the two parts to this question.

(1) An AI model trained on open-source software is likely to qualify as fair use as long as the outputs of the model are not substantially similar to the protectable original expression of

⁵ For an initial proposal for best practices for “Copyright Safety for Generative AI,” see Matthew Sag, *Copyright Safety for Generative AI* (May 4, 2023) (Available at SSRN: <https://ssrn.com/abstract=4438593>).

the copyrighted inputs. However, making this assessment in the context of computer software is complicated by the anomalous nature of software.

(2) An AI model trained on open-source software may not even need to qualify as fair use if the open-source license permits copying. Whether copying is permitted depends on the exact terms of the license and the specific details of how the training data is used. Some open-source licenses allow for unrestricted reuse, some allow for unrestricted non-commercial reuse, some prohibit the creation of derivative works, some are contingent on the license terms being carried forward to derivative works based on the open-source software.

Let's consider an open-source license that authorizes reproduction and the creation of derivative works, but under the condition that any derivative works must also be licensed under the same terms. I.e., a viral license. In that scenario, an AI developer could reproduce the works as part of the training data and fall within the terms of the license by (a) releasing the trained model under the same open-source license,⁶ or (b) ensuring that the model did not amount to a derivative work. This is quite plausible because the learned weights and biases of an AI model trained on open-source software represent an abstraction and generalization of the input data, rather than a copy of the training data. There are some clear examples of generative AI coding tools memorizing and repeating examples from the training data, but if this can be avoided the trained model will not amount to a derivative work.

I should say more about derivative works because this is an aspect of copyright law that people find very confusing. The scope of the Copyright Act's right "to prepare derivative works based upon the copyrighted work" (Section 106(2)) is often misunderstood. A poem inspired by a painting is not a derivative work. An index to a textbook is not a derivative work. A frequency table showing how often words are used in a novel is not a derivative work. A piece of software that is not substantially similar to the software in the training data is not a derivative work.

Making a derivative work necessitates recasting a qualitatively and quantitatively significant amount of the primary work's original expression into a new form or a new version. Assessing whether this threshold has been met requires some understanding of what made the primary work copyrightable in the first place. Suppose I reduced a novel such as *Fifty Shades of Gray* down to a table of individual words and the frequency with which they appeared in the text. I could program a computer to randomly construct an alternative novel, *Gray Fifty Shades Of*, which followed traditional rules of English grammar and used the same individual words in the same proportions (plus or minus 5%, to give it some flexibility). A few things should be obvious about, *Gray Fifty Shades Of*: (i) it would be terrible; (ii) it would not exist, but for *Fifty Shades Of Gray*, (iii) but it would not convey any of the original expression of the primary work. Without some nontrivial overlap in original expression *Gray*

⁶ Complying with attribution requirements in some of the creative commons licenses could be tricky.

Fifty Shades Of would not be a derivative work. On the other hand, a sequel to the primary work that uses the same characters and settings would be very likely to be a derivative work.

Question 15. Some AI developers have said that the ingestion of copyrighted works is transformative and qualifies as fair use. What impact does the Supreme Court’s recent decision in *Andy Warhol Foundation v. Goldsmith* have on that position?

The Supreme Court’s 2023 decision in *Andy Warhol Foundation v. Goldsmith* (“*AWF*”) emphasizes that the question of “whether an allegedly infringing use has a further purpose or different character ... is a matter of degree, and the degree of difference must be weighed against other considerations, like commercialism.”⁷

AWF reaffirms the importance of transformative use and implicitly rejects lower court rulings that had found uses to be transformative where there was no significant difference in purpose. Simply adding a layer of new expression or a new aesthetic over-the-top of someone else’s expressive work and communicating both the old and new expression to the public in a commercial context, without further justification, is not fair use. The Second Circuit was wrong to suggest in *Carion v. Prince*, 714 F.3d 694 (2d Cir. 2013) merely imposing a “new aesthetic” on an existing work was enough to be transformative. It was correct to retreat from that position in *Andy Warhol Foundation v. Goldsmith* 11 F.4th 26 (2021). The Supreme Court’s decision in *AWF* simply reinforces the position that the Second Circuit had already taken. It is not a major change in the law of fair use, even if it did puncture some wishful thinking about fair use.

AWF helpfully clarifies the reason why a transformative use has featured so prominently in the case law: the more transformative a use is, the less likely it is to substitute for the copyright owner’s original expression. Consider classic fair uses such as parody, commentary, or criticism may include substantial portions of the author’s original expression, but these uses are so intrinsically different that they do not usually pose any risk of expressive substitution. In contrast, merely adding an overlay of new expression provides no such comfort. Deriving uncopyrightable abstractions from training data and using those obstructions to generate novel images/music/text is highly transformative.⁸ Nothing in the *AWF* indicates to the contrary.

⁷ *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (2023), Slip Opinion at 12. (emphasis added).

⁸ *A.V. v. iParadigms Liab. Co.*, 544 F. Supp. 2d 473, 482 (E.D. Va. 2008): “This Court finds the “purpose and character” of iParadigms’ use of Plaintiffs’ written works to be *highly transformative*. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students’ written works from plagiarism. iParadigms achieves this by archiving the students’ works as digital code and makes no use of any work’s particular expressive or creative content beyond the limited use of comparison with other works.” *AV Ex Rel. Vanderhye v. iParadigms, LLC*, 562 F. 3d 630, 640 (4th Cir, 2009): “The district court, in our view, correctly determined that the archiving of plaintiffs’ papers was transformative and favored a finding of “fair use.” *iParadigms’ use of these works was completely unrelated to expressive content and was instead aimed at detecting and discouraging plagiarism.*” *Authors Guild, Inc. v. HathiTrust*, 755 F. 3d 87, 97 (2d Cir. 2014): “... we conclude that the creation of a full-text searchable database is a *quintessentially transformative use*.”;

Question 16. Scraping the Internet for data – text, images, audio, video, etc. – for use in training AI models has all the current focus. However, once this has been done the focus may shift to sources of data that are not as readily accessible, such as private user data. Do you foresee companies using cloud-based file storage systems – such as Microsoft OneDrive, Google Drive, Dropbox etc. – as a potential source of data to be scraped? What are your thoughts on this?

(1) Scraping training data from closed sources, such as private cloud storage, without express permission would almost certainly violate federal and state laws targeted at computer hacking.

Most obviously, it would amount to access without authorization in contravention of the Computer Fraud and Abuse Act. Furthermore, web scraping that involved defeating or circumventing technological protection measures would also likely violate the DMCA’s anti-circumvention provisions, see Section 1201. Web scraping can also implicate people’s privacy rights. As of 2023, there is no broadly applicable federal privacy law and many of the relevant state statutes are quite recent.

Furthermore, in the right circumstances, web scraping could also give rise to a claim sounding in unfair competition, unfair and deceptive trade practices, trespass to chattels, conversion, trade secret claims, tortious interference with a contract, tortious interference with a prospective economic advantage, unjust enrichment, and misappropriation.

(2) Of course, none of this would prevent a company like Microsoft, Amazon, or Google from using their customers private cloud storage data as training data if that activity is permitted under their terms of service. Cloud storage companies tend to give themselves an incredibly broad latitude in this regard and Congress may wish to consider whether additional consumer and business protections are required in this context.

(3) Web scraping is often a violation of the terms of service of the relevant website. Whether terms of service are enforceable contracts is a question of state law. The Solicitor General recently expressed doubts as to the enforceability of terms that are merely posted on a website without requiring some express affirmation or agreement in an amicus brief in *ML Genius Holdings LLC v. Google LLC*. In that case, the Second Circuit held that the browse-wrap terms of service that placed limits on the reproduction of music lyrics posted to the ML Genius website were preempted under the Copyright Act. The Supreme Court denied

Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 216-7 (2d Cir. 2015): “We have no difficulty concluding that Google’s making of a digital copy of Plaintiffs’ books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a *highly transformative* purpose, in the sense intended by Campbell.” *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 217 (2d Cir. 2015): “... through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. *We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell* as strongly favoring satisfaction of the first factor.”

ML Genius’ petition for certiorari. Congress should consider whether legislation is needed to clarify the scope of Copyright Act preemption in relation to contracts.

(4) Even in cases where training and AI model on copyrighted works amounts to a non-expressive use, the particular facts of a given case could tip the balance against fair use. I will say more on this in my answer to question 18, but for the moment it is worth noting that a court could consider that obtaining training data by violating the CFAA, Section 1201 of the Copyright Act, privacy laws, or binding contractual restrictions, is a significant factor militating against fair use.

Question 17. To what extent should the use of AI impact whether a human creator receives a copyright? In other words, if AI is being used as a “tool,” should a human still be able to receive a copyright if they have independently contributed creative content?

Please refer to Appendix A of my written testimony, “When Should A Human Be Credited With Authorship Of Something Created Using Generative AI?”

Question 18. Let’s assume that under *Andy Warhol Foundation v. Goldsmith* the use of copyrighted works for training AI is not considered transformative. Do you believe the use of these works would still qualify as fair use looking at the four factors? Which particular factors support your position?

The assumption is far-fetched. For the reasons explained in my answer to question 15, there is no good reason to think that the recent Supreme Court decision undermines the fair use status of non-expressive uses.

However, even without making the assumption, I can offer some thoughts on how the individual fair use factors should be applied in relation to generative AI.

(i) Assuming that machine learning training amounts to a non-expressive use, its “purpose and character” will favor a finding of fair use under the first statutory factor.

The first fair use factor calls for an evaluation of “the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes.”⁹ Since the Supreme Court’s 1994 decision in *Campbell v. Acuff-Rose*, whether the defendant’s use was “transformative”—meaning that the use added “something new, with a further purpose or different character”—has been the central question under the first factor.¹⁰ The Court’s 2023 decision in *Andy Warhol Foundation v. Goldsmith* (“AWF”) emphasizes that the question of “whether an allegedly infringing use has a further purpose or

⁹ 17 USC 107.

¹⁰ *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994). For an empirical analysis, see Clark D. Asay, Arielle Sloan, and Dean Sobczak. *Is transformative use eating the world*, 61 B.C. L. REV. 905 (2020).

different character ... *is a matter of degree*, and the degree of difference must be weighed against other considerations, like commercialism.”¹¹

If a machine learning model was trained in such a way that its outputs were substantially similar to its inputs, the copying that produced the training data would not amount to a non-expressive use, it would not be considered transformative, and the remaining fair use factors would go against fair use. On the other hand, assuming that the outputs are not substantially similar to the inputs, the copying that produced the training data would be a non-expressive use, it would be highly transformative, and it would thus be of a “purpose and character” that was consistent with fair use — regardless of whether it was undertaken by a commercial or nonprofit entity. In this scenario, which should be the more common one, the remaining fair use factors would be addressed as follows...

(ii) The second factor, the nature of the copyrighted work, has no independent relevance, it is not a factor that goes either for or against fair use, it is the context in which the other three factors must be evaluated.

Factor two, the nature of the copyrighted work, simply reminds courts to take context into account when addressing the substantive considerations of purpose and character (factor one), amount and substantiality (factor three), and effect (factor four).

Some authorities suggest that the nature of the work, whether it is creative/informational, or published/unpublished is a stand-alone consideration such that some works merit greater copyright protection than others. This approach is ill-conceived. The nature of the work is not an independent factor that weighs in favor or against a finding of fair use, it is simply the context in which courts must apply the substantive considerations of purpose, proportion, and effect set out in factors one, three, and four respectively.

To elaborate, images are not less worthy of copyright protection than text, but it is much harder to selectively comment on an image or use just part of an image as evidence than it is with purely textual works. Accordingly, full quotations of an image might be reasonable and proportional in circumstances where partial quotation of the text would be.

The statute is not wrong to direct courts to think about the nature of the work; indeed, it would be quite impossible to analyze the purpose, proportion, and effect of the defendant’s use without taking into account the nature of the work. Moreover, works like computer

¹¹ Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith (2023), Slip Opinion at 12. (emphasis added).

software may have special characteristics that inform a fair use analysis.¹² Note that the majority's fair use analysis in *Google v Oracle* begins with the second factor.¹³

(iii) If a use is non-expressive, then the third statutory factor which considers "the amount and substantiality of the portion used" will also favor finding of fair use.

The ultimate question under the third fair use factor is whether the amount of copying was reasonable in relation to a purpose favored by fair use.¹⁴ Although non-expressive uses typically involve making complete literal copies, that copying has been found to be reasonable because it is an intermediate technical step in an analytical process that does not lead to the communication of the underlying original expression to a new audience. Accordingly, courts in non-expressive use cases have found the third factor weighs in favor of the defendant.¹⁵

(iv) If a use is non-expressive, the fourth statutory factor which considers the effect on the "potential market for or value of the copyrighted work" will also favor a finding of fair use in many cases.

The "market" and "value" referred to in the fourth fair use factor are not simply any benefit the copyright owner might choose to nominate; these terms mean something much more specific. A critical book review that quotes from a novel does not have an adverse market effect if it persuades people to buy different book instead;¹⁶ a report from a plagiarism

¹² *Google LLC v. Oracle America, Inc.*, 141 S. Ct. 1183, 1198 (2021) ("... fair use can play an important role in determining the lawful scope of a computer program copyright ... It can help to distinguish among technologies. It can distinguish between expressive and functional features of computer code where those features are mixed. It can focus on the legitimate need to provide incentives to produce copyrighted material while examining the extent to which yet further protection creates unrelated or illegitimate harms in other markets or to the development of other products.")

¹³ *Id.* at 1201.

¹⁴ *Campbell v. Acuff-Rose*, 510 U.S. 569, 586-87 (1994). ("[T]he extent of permissible copying varies with the purpose and character of the use.") In *Campbell*, the Court characterized the relevant questions as whether "the amount and substantiality of the portion used ... are reasonable in relation to the purpose of the copying," and noting that the answer to that question depends on "the degree to which the [copying work] may serve as a market substitute for the original or potentially licensed derivatives[.]" *Id.* at 586-588.

¹⁵ *AV Ex Rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 642 (4th Cir. 2009); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 98 (2d Cir. 2014) "In order to enable the full-text search function, the Libraries, as we have seen, created digital copies of all the books in their collections. Because it was reasonably necessary for the HDL to make use of the entirety of the works in order to enable the full-text search function, we do not believe the copying was excessive." *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 221 (2d Cir. 2015) "Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original." *Authors Guild, Inc. v. Google, Inc.*, 804 F.3d 202, 221-222 (2d Cir. 2015) "As with *HathiTrust*, not only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose. ... While Google makes an unauthorized digital copy of the entire book, it does not reveal that digital copy to the public. The copy is made to enable the search functions to reveal limited, important information about the books."

¹⁶ *Campbell v. Acuff-Rose*, 510 U.S. 569, 591-592 (1994) "We do not, of course, suggest that a parody may not harm the market at all, but when a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act. Because parody may quite legitimately aim at garroting the original, destroying it commercially as well as artistically, the role of the courts is to distinguish between biting criticism that merely suppresses demand and copyright infringement, which usurps it."

detection service might depress the market for helping students cheat on their homework, but that is hardly a cognizable injury under copyright law.¹⁷ More generally, in *Campbell* and subsequent cases, the courts have recognized that the copyright owner has no protectable interest in preventing criticism, parody,¹⁸ or simply locking up unprotectable ideas and expression.¹⁹

By definition, if a use is non-expressive then it poses no direct threat of expressive substitution and thus should generally be preferred under the first factor (purpose and character) and considered harmless under the fourth factor (market effect).

The argument that copyright owners have an inherent right to charge for non-expressive uses, and thus suffer an adverse market effect under the fourth factor, is transparently circular. In theory, every defendant in every fair use case could pay the plaintiff for the right to engage in the challenged use, but if the use is fair there is no obligation to pay. To avoid such circular arguments, courts have limited market effect under the fourth factor to those that represent a cognizable copyright interest.²⁰ Accordingly, in *HathiTrust*, the second Circuit rejected the plaintiff's argument that not being paid for text mining was a cognizable harm, noting that "[l]ost licensing revenue counts under Factor Four only when the use *serves as a substitute for the original* and the full-text-search use does not."²¹ Likewise, in *Google Books*, the court insisted on focusing "on whether the copy brings to the marketplace a *competing substitute for the original*, or its derivative, so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."²²

The substitution the courts are referring to here is expressive substitution, not simply the threat of a more competitive marketplace. A non-expressive use can be harmless under the

¹⁷ *AV ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 464 (4th Cir. 2009) ("Clearly no market substitute was created by iParadigms, whose archived student works do not supplant the plaintiffs' works in the 'paper mill' market so much as merely suppress demand for them, by keeping record of the fact that such works had been previously submitted In our view, then, any harm here is not of the kind protected against by copyright law.")

¹⁸ *Campbell v. Acuff-Rose*, 510 U.S. 569, 577-79 (1994); *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 482 (2d Cir. 2004) ("[C]riticisms of a seminar or organization cannot substitute for the seminar or organization itself or hijack its market."); *Bill Graham Archives v. Dorling Kindersley, Ltd.*, 448 F.3d 605 (2d Cir. 2006) ("A copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody ... or other uses of its own creative work.") (internal quotations omitted).

¹⁹ *Sega Enters., Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992); *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000).

²⁰ *Campbell v. Acuff-Rose Music*, 510 U.S. 569, 591-92 (1994) (no cognizable market effect where parody or criticism depress demand for the original work); see also *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 607 (9th Cir. 2000) (noting that a videogame manufacturer's desire to foreclose competition in complementary products was understandable, but that "copyright law ... does not confer such a monopoly."); *Bill Graham Archives v. Dorling Kindersley, Ltd.*, 448 F.3d 605, 615 (2d Cir. 2006) ("[A] copyright holder cannot prevent others from entering fair use markets merely by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work.") (citations and quotations omitted).

²¹ *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 100 (2d Cir. 2014) (emphasis added).

²² *Authors Guild v. Google, Inc.*, 804 F.3d 202, 223 (2d Cir. 2015) (emphasis added).

fourth factor even if it results in the creation of a competing product—as long as the competing product does not contain an infringing level of original expression taken from the plaintiff's work. In *Sega v. Accolade* and again in *Sony Computer Entertainment v. Connectix Corp.*,²³ the Ninth Circuit found that reverse engineering a gaming console in order to produce interoperable games (*Sega*), and a rival gaming platform (*Sony*), was fair use. In both cases the Ninth Circuit found that there was no cognizable market effect because the rival products did not contain any protectable expression derived from the plaintiffs' consoles. The defendants were entitled to use uncopyrightable elements from those consoles to make new independent creative expression possible.²⁴

(v) However, other considerations may nonetheless tilt the fourth factor against fair use.

Non-expressive uses that substantially undermine copyright incentives could be considered unfair. To recap on my written testimony,

- (1) A court in some future case may well consider whether a defendant had lawful access to the works used as training data under the fourth factor.
- (2) Likewise, a future court might extend the fourth factor to consider whether, in scraping material from the Internet, the defendant ignored robot.txt files indicating a desire to opt out of search engine indexing and similar activities. Likewise, a court might conclude that scraping material from a website in violation of its terms of use was relevant to the fourth factor, if the inability to rely on such exclusions substantially undermined copyright incentives.
- (3) A plaintiff might argue that it is unfair to systematically extract valuable uncopyrightable material from a website or other information source and then use that material as a substitute for the functionality of the website. This argument would be strongest where the systematic extraction was likely to significantly undermine the website's incentives for original content production.

This argument is hard to reconcile with the view that the idea-expression distinction is meant to encourage competition where the competing product does not include too much of the plaintiff's original expression. But it is not foreclosed by existing precedent.

These may be valid considerations under the fourth factor, but I would not elevate them to independent factors or prerequisites. They may have different salience in different cases and will generally be more relevant in commercial fair use cases than non-commercial ones.

²³ *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F. 2d 1510, 1523. *Sony Computer Entertainment v. Connectix Corp.*, 203 F. 3d 596, 608.

²⁴ *Id.*

Question 19. One concern about generative AI that has been raised by creators is that unauthorized copies of their works are being made during the process of collecting data and training a respective model. Could you please explain how copies and how many copies of such data are made and when within the lifecycle of creating and executing an AI system – from start to end?

The answer to this question may vary significantly depending on the model being trained. However, I can answer the question with reference to a generic example:

Model development begins by identifying and obtaining access to the relevant training data. It is hard to imagine that any large model could be trained without at least one locally stored copy of the training data. To avoid overfitting (and thus hopefully minimize the risk of copyright infringement and other analogous harms), it is important to deduplicate the training data. Practically speaking, this is hard to do without creating a semi-permanent local copy. To address questions of bias and filter out toxic materials, the potential training data needs to be analyzed carefully before training begins. Again, this is much more practical with access to a semi-permanent local copy. Storing a semi-permanent local copy also makes sense if the developer anticipates the need to retrain the model from time to time. Continued access to the training data in its original form may also be necessary to evaluate the performance of the model, and to take additional steps to mitigate the potential for copyright infringement, or other undesirable outcomes.

The training process itself does not involve copying or storing documents in their original format. However, segmenting the training data into tokens and converting those tokens into a numerical representation is, technically, another form of copying.

To elaborate, the data used to train models like GPT-3 and other text-based large language models do not consist of words or symbols that are meaningful to, or intelligible by, humans. At the beginning of the training process the raw text data is broken down into smaller pieces, known as tokens. These tokens could be as short as one character or as long as one word (in English and similar languages). For example, the text “U.S. Senate Hearing” is probably broken down into the tokens “U” “.” “S” “Senate” “Hear” and “ing”. Each token is then mapped to a unique numerical value.

This is a one-to-one mapping, so you could reverse engineer the original human-readable text from these numerical representations, and thus this still qualifies as a “copy” for the purposes of copyright law’s reproduction right.

The actual training process for a model like GPT-3 involves feeding the numerical representations of tokens into the model and having it make predictions about the next token. Through this process, the model learns the structures, rules, and patterns in the language. The model doesn’t (or shouldn’t) retain any specific copyrighted works from the training data.

The model itself is an absurdly large statistical model that can be used to predict the next token given a set of input tokens. This model is not a copy of the training data.

When the model is deployed it generates entirely new content based on the statistical patterns it learned during training.

In summary, using copyrighted works as training data for generative AI necessarily involves at least two steps that would qualify as creating a copy under the reproduction right in Section 106(1). In practice, there may be some additional technical copies created, but none with any independent economic significance.

Question 20. Some have suggested different licensing structures for compensating copyright owners for the use of their works in AI training. What licensing structures have you seen or used that have worked to the mutual benefit of both AI companies and copyright owners?

As long as commercial AI developers respect machine-readable opt-outs, refrain from sourcing training data from sites of known infringement, respect paywalls, and other technological exclusions, we can expect a vibrant voluntary licensing system to emerge. Already, AI developers are negotiating access deals with media companies and stock photography agencies.

A compulsory license in relation to AI training would be difficult to administer and would interfere with voluntary licensing. Existing statutory licenses compensate right holders on a per-play or per-use basis and thus avoid the need to assess the merit or contribution of a given work. There is no easy way to assess how much a single work contributes to a machine learning model. If every work used to train a model is valued equally, then the remuneration an author or artist received would not be calibrated to the importance or value of her work; it would also tend to approach zero as the number of works in the training data increased.

Thank you again for the opportunity to assist the Senate in this hearing.



Matthew Sag

July 12, 2023

Chair Chris Coons
U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property
Washington, DC 20510

Ranking Member Thom Tillis
U.S. Senate Committee on the Judiciary
Subcommittee on Intellectual Property
Washington, DC 20510

Re: July 12, 2023 Senate Judiciary Subcommittee on Intellectual Property Hearing:
Artificial Intelligence and Intellectual Property – Part II: Copyright

Dear Chair Coons and Ranking Member Tillis:

On behalf of the Computer & Communications Industry Association (“CCIA”),¹ we write to explain how existing copyright law already covers emerging technology like artificial intelligence (“AI”). CCIA requests that this statement be included in the record of the hearing scheduled for July 12, 2023.

CCIA’s members are leaders in AI innovation. Most modern AI technology utilizes hardware manufactured by CCIA member companies, and much of it relies on AI technology released as open-source contributions by CCIA members. Because of this, CCIA members have a significant interest in ensuring that the various types of AI-related inventions are promoted, rather than suppressed, by the U.S. copyright system. CCIA believes that existing law is generally capable of addressing issues related to artificial intelligence and serves to promote creative activity in AI technology. While unique issues might arise in the future that require additional legislation or regulation, the technology-neutral nature of the present Copyright Act is sufficient to address present issues regarding AI and copyright.

I. Authorship

One primary issue, AI authorship, has already been addressed by the Copyright Office. As properly interpreted by the Copyright Office, a work produced by an AI algorithm or process, without the involvement of a natural person contributing to the resulting work does not qualify as a work of authorship protectable under U.S. copyright law. This interpretation follows in a long line of cases and guidance finding that only a natural person can create a work of authorship protectable via copyright.

¹ CCIA is an international, not-for-profit trade association representing a broad cross-section of technology and communications firms. For more than 50 years, CCIA has promoted open markets, open systems, and open networks. CCIA members employ more than 1.6 million workers, invest more than \$100 billion in research and development, and contribute trillions of dollars in productivity to the global economy. For more, visit www.ccianet.org.

The Copyright Office currently refuses to register a work that was not created by a human being. It explains that the copyright law only protects “the fruits of intellectual labor” that “are founded in the creative powers of the mind.” *Trade-Mark Cases*, 100 U.S. 82, 94 (1879). Because copyright law is limited to “original intellectual conceptions of the author,” the Office will refuse to register a claim if it determines that a human being did not create the work. *Burrow-Giles Lithographic Co. v. Sarony*, 111 U.S. 53, 58 (1884). The Copyright Office adds that it “will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author.” There is no need for this provision to change. Artists who incorporate technology into their artistic process can still copyright the overall artistic product, so long as the human artist has contributed a sufficient amount of original material to the combined work.

II. Training

Existing law also adequately addresses the issue of the use of copyrighted material for training AI models. The existing statutory framework and related case law concerning the fair use right, 17 U.S.C. § 107, clearly permit the ingestion of large amounts of copyrightable material for the purpose of an AI algorithm or process learning its function.

The flexible and balanced copyright law regime in the U.S. has been key to American success in innovation in emerging technologies like AI. The U.S. leads the way in AI development in large part due to the fair use right. AI algorithms and other processes often require the ingestion of large amounts of material. Assembling that material may entail converting it into a more usable format, e.g., translating image files into mathematical image representations. In addition, backup copies of the materials may be necessary to protect against loss of data in the event of system failure. Temporary reproductions of portions of the material in a computer’s random access memory are a normal part of any computer program, including the process of training an AI algorithm. These copies are not viewable or consumable by the outside world. These non-expressive copies are not consumable by the public. They do not function as market substitutes for copies of the ingested works.

Numerous appellate courts have correctly found the mass copying of raw material to build datasets for automated computational analysis to be fair use under 17 U.S.C. § 107. See, e.g., *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015); *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 640 (4th Cir. 2009); *Perfect 10 v. Amazon.com, Inc.*, 508 F.3d 1146, 1165 (9th Cir. 2007); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003).² Training AI is a form of this computational analysis.

² See also Jonathan Band, *Israel Ministry of Justice Issues Opinion Supporting the Use of Copyrighted Works for Machine Learning*, Disruptive Competition Project (Jan. 19, 2023), <https://www.project-disco.org/intellectual-property/011823-israel-ministry-of-justice-issues-opinion-supporting-the-use-of-copyrighted-works-for-machine-learning/>.

Judge Leval's opinion in *Google* provides the clearest analysis of why the creation of datasets for computational analysis, and their subsequent uses, are fair uses.

While this activity is already fair use, if more clarification is needed in the future, Congress could consider enacting a specific provision permitting the creation and use of datasets for computational analysis. Such an unambiguous provision could provide useful guidance to AI developers and courts.

III. Infringement

Finally, existing copyright law adequately protects against infringement. If the output of an AI system resembles existing copyrighted material, then the ordinary analysis of whether copyright infringement has occurred would apply. In short, the question would be whether the AI system had access to the allegedly infringed work, and whether the AI system's output is substantially similar to the allegedly infringed work. The first question is trivial to answer by examining whether the work in question was part of the training data used by the AI system. If it was not used in training, the AI system did not have access to it. The second question is answered as it would be in any other copyright case.

The sole remaining question in output infringement is who would be held liable for the infringement. Here, as in other areas of law, principles of agency and causation will guide the courts. Generally, however, liability should lie on the end-user who requests and publishes a copyright-infringing work. Much like many other areas of technology, including photography, AI systems are strong examples of a "staple article or commodity of commerce suitable for substantial noninfringing use." *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417, 440-42 (1984). Misuse of AI systems to infringe copyright, much like misuse of a VCR or computer to impermissibly replicate copyrighted content, is attributable to the user, not the manufacturer of the system being abused.

* * * * *

CCIA hopes this letter helps illuminate the flexible nature of current copyright law and how it stands ready to handle the new issues raised by the creation of generative AI. We would be pleased to provide further information to your offices and those of other members of the Subcommittee on any topic which might aid you in your work.

Sincerely,

Ali Sternburg
Vice President, Information Policy, CCIA
Joshua Landau
Senior Counsel, Innovation Policy, CCIA



July 17, 2023

The Honorable Chris Coons
Chairman
Senate Judiciary Committee
Subcommittee on Intellectual Property
218 Russell Senate Office Building
Washington, D.C. 20510

The Honorable Thom Tillis
Ranking Member
Senate Judiciary Committee
Subcommittee on Intellectual Property
113 Dirksen Senate Office Building
Washington, D.C. 20510

Dear Chairman Coons and Ranking Member Tillis:

We write today to thank you for holding the recent hearing on *Artificial Intelligence and Intellectual Property – Part II: Copyright* and to submit for the record a short paper outlining our positions on copyright and AI. We also thank the Subcommittee for its attention to the significant copyright implications surrounding the development and use of generative artificial intelligence. As the only organization in the United States representing the entire creative community on copyright law issues, we stand ready to assist your efforts to ensure the concerns of America's creators and copyright owners are effectively addressed.

The Copyright Alliance is a non-profit, non-partisan public interest and educational organization representing the copyright interests of over 2 million individual creators and over 15,000 organizations in the United States, across the spectrum of copyright disciplines. The Copyright Alliance is dedicated to advocating policies that promote and preserve the value of copyright, and to protecting the rights of creators and innovators. The individual creators and organizations that we represent rely on copyright law to protect their creativity, efforts, and investments in the creation and distribution of new copyrighted works for the public to enjoy.

There are five fundamental principles that must form the basis of a common understanding amongst stakeholders, courts, policymakers, and the public when it comes to the relationship between copyright and generative AI.

1. ***When formulating new AI laws and policies, it is essential that the rights of creators and copyright owners be respected.*** When making determinations about AI policies, it is vital for policymakers and stakeholders to understand that any new laws and policies relating to AI must be based on a foundation that preserves the integrity of the rights of copyright owners and their licensing markets. The interests of those using copyrighted materials for ingestion by AI systems must not be prioritized over the rights and interests of creators and copyright owners.

2. ***Long standing copyright laws and policies must not be cast aside in favor of new laws or policies obligating creators to essentially subsidize AI technologies.*** Established copyright laws must not be weakened based on a mistaken belief that doing so is necessary to incentivize AI technologies. This is especially true when there is no evidence of market failure or problems warranting changes to the law. AI-specific statutory exceptions to copyright law that would effectively strip rightsholders of their ability to control and be compensated for the use of their copyrighted works for ingestion purposes are not necessary and should be rejected.
3. ***The ingestion of copyrighted material by AI systems implicates the right to reproduce copyrighted works.*** Section 106(1) of the Copyright Act vests copyright owners with the right to prevent the reproduction of their copyrighted works. When an unauthorized copy is made of a work protected by copyright, there is a violation of the copyright owner’s right to reproduce the work, absent a valid defense. It is important to understand that copyright infringement at the input stage is distinguishable from output stage infringement because the reproduction right is a “stand-alone” right—it is violated by copying a work (without authority) *regardless of whether a specific output of AI system is infringing.*
4. ***The ingestion of copyrighted material by AI systems is not categorically fair use.*** Some AI developers argue that ingestion of copyrighted works by AI tools always constitutes fair use. This is incorrect. Determining whether a particular use qualifies for the fair use defense to infringement requires a fact-specific inquiry that is considered on a case-by-case basis. Courts will need to evaluate fair use defenses involving AI systems the same way they evaluate fair use in all contexts: by applying the four factors set forth in section 107 of the Copyright Act to the specific uses at issue. Importantly, while some AI companies argue that ingestion of copyrighted materials qualifies as fair use because it is a “transformative” purpose, the Supreme Court recently made clear that whether a use is transformative is not dispositive of fair use and is merely one of many considerations under the first fair use factor. Finally, under the fourth factor, courts will need to determine the extent of the “effect of the use upon the potential market for or value of” the works ingested by that system. 17 U.S.C. §107(4).
5. ***AI systems must implement safeguards to prevent infringing AI-generated outputs.*** Overfitting and allowing prompts that call for copyright protected-material and “in the style of” are more likely to result in AI-generated outputs that infringe one or more copyrighted works. While merely imitating the style of an existing artist does not constitute infringement, it is essential that AI companies implement effective safeguards to prevent the likelihood of output-related infringements. This is yet another reason why the AI companies should license ingested works because when they do so, the parties can negotiate these safeguards.

Another important consideration at the intersection of AI and copyright is the role that licensing is already playing (and will continue to play). Many creators and rightsholders already license their copyrighted works—which AI developers recognize as immensely

valuable—for commercial AI uses, and many more are on the cusp of doing so. Where a copyright owner offers licenses for uses relating to the training of AI systems, it is essential that the licenses be respected by any copyright or AI legal regime. The marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the right of copyright owners to license, or choose not to license, their works for AI purposes.

Lastly, as Subcommittee members recognized during the hearing, transparency regarding ingestion of copyrighted works is essential to helping ensure that the rights of copyright owners are respected and that AI development is being implemented in a way that is responsible and ethical. Adequate and appropriate transparency and record-keeping benefit both copyright owners and AI developers in resolving questions regarding infringement, fair use, and compliance with licensing terms. Transparency is a crucial component of any AI policy.

The Copyright Alliance and our members support responsible, respectful, and ethical development and use of AI technologies and a thriving and robust AI economy. As AI technology continues to evolve, it is critical that the underlying goals and purposes of our copyright system are upheld and that the rights of creators and copyright owners are respected. We once again thank you for your leadership in AI policy, and we look forward to working together on these important issues with members of Congress, the U.S. Copyright Office and other stakeholders.

Sincerely,



Keith Kupferschmid
CEO
Copyright Alliance

cc:

Senator Marsha Blackburn
Senator John Cornyn
Senator Tom Cotton
Senator Mazie Hirono
Senator Jon Ossoff
Senator Alex Padilla
Senator Peter Welch



POSITION PAPER

Artificial Intelligence

The Copyright Alliance supports the responsible development of AI technologies and a thriving and robust artificial intelligence (AI) economy. The continuing development of AI systems represents a profound achievement of the digital age that brings with it tremendous opportunities. In fact, many in the creative industry are already using or plan to use AI for the creation of a wide range of works that benefit society. But as with many advances in technology, these new opportunities come with challenges.¹

Advancements in AI have led to a new frontier in generative technologies, and thus they are often accompanied by difficult legal questions surrounding both the ingestion of copyrighted works into AI systems and 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 the underlying goals and purposes of our copyright system are upheld and that the rights of creators and copyright owners are respected.

When examining the intersection of AI and copyright, the following general principles must form the basis of a common understanding amongst stakeholders, courts, policymakers, and the public.

- *When formulating new AI laws and policies, it is essential that the rights of creators and copyright owners be respected.* When making determinations about AI policies, it is vital for policymakers and stakeholders to understand that any new laws and policies relating to AI must be based on a foundation that preserves the integrity of the rights of copyright owners and their licensing markets. The interests of those using copyrighted materials for ingestion by AI systems must not be prioritized over the rights and interests of creators and copyright owners.
- *Long standing copyright laws and policies must not be cast aside in favor of new laws or policies obligating creators to essentially subsidize AI technologies.* Established copyright laws must not be weakened based on a mistaken belief that doing so is necessary to incentivize AI technologies. This is especially true when there is no evidence of market failure or problems warranting changes to the law. AI-specific statutory exceptions to copyright law that would effectively strip rightsholders of their ability to control and be compensated for the use of their copyrighted works for ingestion purposes are not necessary and should be rejected.

¹ This paper addresses topics specific to the use of copyrighted works for ingestion by AI systems. There are several other questions that will arise as to who, if anyone, is the "author" of a work generated by an AI system, who, if anyone, is responsible for any copyright infringement committed via such system, and the copyrightability of AI-generated works in general. Those subjects will be the focus of future position papers.

- *Education is paramount in the AI space.* There must be efforts to educate participants and users in the AI industries to respect third-party rights such as copyright and otherwise act in an ethical and lawful manner.

Some of the most relevant areas of interest for the copyright community include:

Benefits of Licensing

Independent-to-large-scale creators and copyright owners produce high-quality works that are often ideal for ingestion by AI machines, and copyright law incentivizes those creators and rightsholders to lawfully enhance and aggregate their copyrighted works for that purpose—such as through semantic enrichment, metadata tagging, content normalization and data cleanup.

Where a copyright owner offers licenses for uses for ingestion by AI systems, it is essential that these licenses be respected by any copyright or AI legal regime, especially in the case of ingestion of copyrighted material used for text and data mining (TDM). There is already high demand for corpuses of copyrighted works for ingestion by AI systems, and copyright owners already enter into licensing agreements for TDM use. This licensing activity is evidence of existing markets for TDM. It is important that the conditions of those licenses are respected and that they are not undermined by new exceptions that excuse unauthorized uses.

Copyrighted works are also being licensed and used for AI projects that in turn generate works that serve as market substitutes for the ingested works. In some cases, the output could qualify as derivatives of the ingested, copyrighted works. In either scenario, copyright owners and creators would be harmed from the unauthorized use of their works, and it is essential that those using the copyrighted works license such uses. In short, the marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the ability of copyright owners to license their works for AI uses. Finally, copyright owners may sometimes choose not to license their works for use in generative systems that may produce competing output, and those choices must be respected.

Fair Use

There are some who believe that use of copyrighted works for AI ingestion will always qualify as a fair use under section 107 of the Copyright Act. That view is inaccurate. While there may be instances where ingestion qualify as a fair use under section 107, that likely would not be the case if a TDM license is available, the use is commercial, or the resulting AI generated work harms the actual or potential market for the ingested work. The answer will depend on the facts in each particular case.

Some AI developers have, without authorization, used copyrighted works incorporated into data sets or pre-trained AI created by non-commercial third parties in their commercial products—a practice known as data laundering. Neither this kind of unauthorized use nor the work of the non-commercial entity necessarily qualify as fair use. Ultimately, AI systems should only ingest works that they have the authority to use.

Transparency

Best practices from corporations, research institutions, governments, and other organizations that encourage transparency around development and use of AI already exist. Transparency includes such things as recording what works are ingested by AI systems and for what purpose, which helps to ensure that copyright owners' rights are respected. Infringement analyses, fair use defenses, and licensing terms disputes can all benefit from transparency best practices, and they can also be crucial in promoting safe, ethical, and unbiased AI systems.

Education & Awareness

As technologies rapidly advance, we caution against forging ahead in a way that would disregard the fundamental legal considerations at the heart of our copyright system. It is crucial that those leading AI projects are aware of the legal implications of using copyrighted works as input material, and those that arise from AI-generated output. Policymakers, in conjunction with stakeholders, must work together on educational initiatives that aim to establish common understandings and educational guidelines that ensure the rights of all are understood and respected as AI technologies evolve.



July 12, 2023

The Honorable Chris Coons
Chair, Subcommittee on Intellectual Property
Senate Judiciary Committee

The Honorable Thom Tillis
Ranking Member, Subcommittee on Intellectual Property
Senate Judiciary Committee

Dear Chairman Coons and Ranking Member Tillis,

Thank you for holding a hearing on artificial intelligence and copyright. I appreciate your interest in this topic and the opportunity to share DiMA's perspective.

DiMA represents the world's leading audio streaming companies, whose innovations are the economic engine that have revitalized the music industry. DiMA and its members – Amazon, Apple Music, Feed.fm, Pandora, Spotify, and YouTube – advocate for policies that ensure that music fans have legal access to music anytime, anywhere they want it, and that artists and songwriters can connect with old fans and make new ones around the world.

Questions about the use and impact of AI technology, its applications and how they intersect with existing law are an area of focus for music industry stakeholders, including DiMA and our member companies. AI has been used as a tool in the music industry for many years, and as the technology continues to rapidly evolve, it has the ability to assist creators (including professional human musicians and songwriters) and improve the way music is created, distributed, and consumed.

Music has long been at the forefront of potentially disruptive and new technology, and the development of new technology has often been met with initial concern by many in the music industry. However, the success of today's streaming-driven music industry is definitive proof that music and technology can, should, and most often do, learn to work together and enrich our musical traditions.

Importantly, AI is not one size fits all. We believe this hearing, and future Congressional engagement, are key to helping to highlight the different types of AI and its uses, and providing greater context around conversations, which we believe will ultimately lead to more fruitful and productive conversations.

Current discussions around AI lack grounding definitions, including the lines between generative and assistive AI and foundation models versus applications. The entire music industry will benefit from establishing a common set of facts in the discussion, and/or focusing questions around specific AI applications. Similarly, policymakers benefit from shared substantive expertise about AI technologies, evolving trends, and the potential effects on artistic expression, innovation, and commercial markets before proposing changes. DiMA members are willing to assist the government in pursuing an evidence-based path. For example, I recently participated in a listening session held by the U.S. Copyright Office regarding AI and Music and Sound Recordings, which I hope will be the start of this kind of analysis.

At a high-level, DiMA members believe that existing U.S. copyright laws, including (1) copyrightability (including originality, *de minimis* contributions, scenes a faire, and the idea/expression dichotomy); (2) infringement (including questions of unlawful appropriation, substantial similarity, and causation); and (3) the Digital Millennium Copyright Act, as well as laws outside of copyright to protect the right of publicity, including one's name, voice, and likeness, are likely to be sufficient to address creations made with or by AI technology. Different legal doctrines can and should be employed to consider questions arising from AI-generated music, but copyright law should not be stretched or changed to address questions that more properly arise under laws relating to trademark, right of publicity, or unfair competition.

Finally, these discussions around AI – particularly as they relate to music – must take into account data, a topic on which DiMA members have extensive experience. There have long been data challenges in the music space, including with ensuring that accurate and complete metadata identifiers are included in recordings at the time of distribution. These challenges existed before streaming, continue to exist today and are highly relevant to discussions of the treatment of AI-generated music. Music streaming services should not be – and cannot be – the arbiters as to what is or is not AI-generated. It is not possible with the existing data, and any new data must come from the copyright owners and creators. Data accountability must exist throughout the chain from creation to distribution.

Thank you again for holding this important hearing. Please do not hesitate to contact DiMA if we can be of assistance as you continue consideration of this topic. We appreciate the opportunity to have our perspective considered and look forward to continued engagement on this topic.

Sincerely,

Garrett Levin

Garrett Levin
President and CEO
Digital Media Association (DiMA)



MOTION PICTURE ASSOCIATION

STATEMENT OF THE MOTION PICTURE ASSOCIATION, INC.

BEFORE THE

SENATE COMMITTEE ON THE JUDICIARY
SUBCOMMITTEE ON INTELLECTUAL PROPERTY

REGARDING THE HEARING

ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY – PART II: COPYRIGHT
JULY 12, 2023

I. INTRODUCTION

The Motion Picture Association, Inc. (“MPA”) appreciates the opportunity to provide this statement for the record following the subcommittee’s July 12, 2023 hearing titled “Artificial Intelligence and Intellectual Property – Part II: Copyright.” The MPA is a not-for-profit trade association founded in 1922. The MPA serves as the voice and advocate of the film and television industry, advancing the business and art of storytelling, protecting the creative and artistic freedoms of storytellers, and supporting the creative ecosystem that brings entertainment and inspiration to audiences worldwide. The MPA’s member companies are Netflix Studios, LLC, Paramount Pictures Corp., Sony Pictures Entertainment Inc., Universal City Studios LLC, Walt Disney Studios Motion Pictures, and Warner Bros. Entertainment Inc. These companies and their affiliates are the leading producers and distributors of filmed entertainment in the theatrical, television, and home-entertainment markets.

While the title of the July 12 hearing indicated a focus on copyright, much of the discussion centered on a separate, non-copyright issue: the potential enactment of a new federal statute regulating the use of individuals’ names, images, and likenesses (“NIL”), as a means to address some of the concerns raised by recent developments in artificial intelligence. The MPA has been deeply engaged on NIL issues in state legislatures for many decades and stands willing to work with Congress and other stakeholders to address the concerns of recording artists, actors, and others about the use of digital replicas to replace their performances without their consent, as well as uses that deceive the public. Many of these concerns, and the proposed legislative responses to them, have been framed under the rubric of the “right of publicity.” As described in this statement, however, many of the concerns animating the discussion and proposed solutions have little in common with the traditional right of publicity—an area in which there exists a robust and established body of legal jurisprudence, the development of which the MPA and its

members have been actively engaged in to ensure the protection of important speech-related interests.

The concerns raised at the hearing regarding uses of digital replicas deserve attention. But it is vitally important that any legislation in this area respect First Amendment rights and creative freedoms, including those of filmmakers, broadcasters, photographers, journalists, and others who employ new technologies to entertain and educate audiences in the U.S. and around the world. In this statement, MPA summarizes the vital First Amendment issues implicated by a potential federal NIL statute, emphasizing that creation of such a right would constitute a content-based regulation of speech, subjecting it to strict scrutiny, which requires both the existence of a compelling state interest and narrow tailoring to serve that interest. MPA also offers the subcommittee guidelines should it pursue legislation in this area and stands ready to serve as a resource to the Committee as it examines these issues.

II. REGULATION OF USES OF NAME, IMAGE, AND LIKENESS

A. Existing State Right of Publicity Laws

Regulation of uses of an individual's name, image, and likeness has traditionally been governed by the body of state law known as "right of publicity." Approximately half the states have enacted right-of-publicity statutes, while almost all of the rest protect this right through the common law.¹ Importantly, the right of publicity applies only to commercial uses of an individual's NIL—for example, in an advertisement or on merchandise. But right of publicity does not—and, consistent with the First Amendment, may not—regulate uses of or references to individuals' NIL in "expressive works"—works such as books, plays, news articles and broadcasts, songs, and movies and television programs.² Such expressive works are non-commercial speech fully protected by the First Amendment.³

Recognizing the need to prevent right-of-publicity statutes from encroaching on First Amendment rights, states that have enacted or amended such statutes in the past several decades have routinely included explicit statutory exceptions, known as "expressive-works exemptions," which make clear that this body of law has no application in the context of creative and

¹ See generally <https://rightofpublicityroadmap.com/>.

² See Restatement (Third) of Unfair Competition, § 47 (right of publicity does not include "the use of a person's identity in news reporting, commentary, entertainment, works of fiction or nonfiction, or in advertising that is incidental to such uses").

³ The Supreme Court has defined "commercial speech" as "speech which does no more than propose a commercial transaction." *Bolger v. Youngs Drug Prod. Corp.*, 463 U.S. 60, 66 (1983) (internal quotation marks omitted). "If speech is not 'purely commercial'—that is, if it does more than propose a commercial transaction—then it is entitled to full First Amendment protection." *Mattel, Inc. v. MCA Records, Inc.*, 296 F.3d 894, 906 (9th Cir. 2002).

journalistic works.⁴ States have codified these exemptions precisely because they know that, if not properly cabined to commercial uses, right-of-publicity statutes risk chilling vast swaths of speech, including art, humor, political commentary, journalism, and criticism, all of which are the lifeblood of a free and well-functioning democracy, allowing us to debate, scrutinize, and laugh at the world around us. And absent such exemptions, right-of-publicity defendants sued over references to and depictions of real people would bear the burden of asserting their First Amendment rights as affirmative defenses in individual cases, an expensive and arduous task that itself chills speech. Despite the inapplicability of right-of-publicity laws to uses in expressive works, individuals unhappy with their portrayals in such works nonetheless routinely assert such claims. The courts just as routinely hold that the First Amendment bars these attempts at censorship.⁵

The U.S. Supreme Court has addressed state regulation of NIL just once, in *Zacchini v. Scripps-Howard Broadcasting Co.*, 433 U.S. 562 (1977), which involved an entertainer who performed a “human cannonball” act at an Ohio county fair. Zacchini sued a broadcaster that aired his entire act, alleging infringement of his “right to the publicity value of his performance.” *Id.* at 565. The Supreme Court held that the First Amendment did not bar Zacchini’s claim. But crucial to the Court’s reasoning was that Zacchini was not merely alleging misappropriation of his identity, but instead misappropriation of his *entire act*. And scholars have recognized that *Zacchini* was not about the “right of publicity” as that phrase is commonly used, but, rather, about a right akin to a common-law copyright, which protects event producers from unauthorized broadcasts of their events.⁶

⁴ See, e.g., Cal. Civ. Code § 3344.1 (1999); Ohio Rev. Code Ann. § 2741.09 (1999); Ark. Code Ann. § 4-75-1110 (2016); N.Y. Civ. Rights Law § 50-f(2)(d) (2020); La. Stat. Ann. § 51:470.5 (2022).

⁵ See, e.g., *Porco v. Lifetime Ent. Servs., LLC*, 195 A.D.3d 1351 (N.Y. App. Div. 2021) (First Amendment barred claim by convicted murderer over portrayal in docudrama); *De Havilland v. FX Networks, LLC*, 21 Cal. App. 5th 845 (Cal. Ct. App. 2018) (First Amendment barred claim by actress over portrayal in docudrama); *Sarver v. Chartier*, 813 F.3d 891, 896 (9th Cir. 2016) (First Amendment barred claim by individual allegedly portrayed in movie *The Hurt Locker*); *Tyne v. Time Warner Entm’t Co., L.P.*, 901 So.2d 802 (Fla. 2005) (First Amendment barred claims involving movie *The Perfect Storm*); *Matthews v. Wozencraft*, 15 F.3d 432 (5th Cir. 1994) (“Courts long ago recognized that a celebrity’s right of publicity does not preclude others from incorporating a person’s name, features, or biography in a literary work, motion picture, news or entertainment story. Only the use of an individual’s identity in advertising infringes on the persona.”).

⁶ See, e.g., Lee Levine & Stephen Wermiel, *The Court and the Cannonball: An Inside Look*, 65 Am. U. L. Rev. 607, 637–38 (2016) (“[A]lthough the Court used the shorthand phrase ‘right of publicity’ to describe the cause of action before it ... the Justices’ focus ... strongly suggests that the right they believed they were confronting was in the nature of a common law copyright (as the Ohio Court of Appeals had indeed characterized it) and had little to do with the right to control the use of one’s image in an otherwise distinct creative work....”); *Wisconsin Interscholastic Athletic Ass’n v. Gannett Co.*, 658 F.3d 614, 628 (7th Cir. 2011) (observing that a “producer of the entertainment,” such as “the NFL, FIFA, or the NCAA,” “normally signs a lucrative contract for exclusive, or semi-exclusive, broadcast rights for the performance,” and that “*Zacchini* makes clear that the producer of entertainment is entitled to charge a fee in exchange for consent to broadcast”).

B. A Potential Federal Name, Image, and Likeness Right

While several witnesses at the July 12 hearing expressed support for enactment of a federal “right of publicity,” the substance of what they were suggesting appears to be something quite different: a new right that—unlike state right-of-publicity statutes—would apply *in expressive works*. This new right, more accurately termed a “digital-replica right” than a “right of publicity,” would aim to prevent the unauthorized digital creation of what appears to be a performance by a recording artist or an actor—for example, the “Heart on my Sleeve” song that falsely appeared to be sung by Drake and The Weeknd.⁷ MPA acknowledges the concerns raised by such developments and looks forward to working with Congress and stakeholders to address them. But in doing so, policymakers must tread carefully, as creation of a new right that would apply in expressive works raises serious First Amendment concerns and risks interfering with core creative freedoms.

III. ANY DIGITAL REPLICA RIGHT WOULD BE A CONTENT-BASED RESTRICTION ON FREE SPEECH, AND WOULD BE SUBJECT TO STRICT SCRUTINY UNDER THE FIRST AMENDMENT.

The First Amendment’s free-speech guarantee provides extraordinary freedom for creators of expressive works, including by protecting the ability of creators to use the names and likenesses of real people. Indeed, the creative freedom guaranteed by the First Amendment, coupled with the robust protections of the Copyright Act—itsself an “engine of free expression”⁸—undergird this nation’s position as the unchallenged world leader in motion pictures, music, and other creative endeavors. Over 70 years ago, the Supreme Court recognized that “motion pictures are a significant medium for the communication of ideas,” and are thus fully protected by the First Amendment.⁹ And the fact that movies or other works are distributed for profit does not lessen their First Amendment protection.¹⁰

⁷ See Joe Coscarelli, *An A.I. Hit of Fake ‘Drake’ and ‘The Weeknd’ Rattles the Music World*, N.Y. Times, April 19, 2023, <https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html>.

⁸ *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 558 (1985).

⁹ *Joseph Burstyn, Inc. v. Wilson*, U.S. 495, 501 (1952); see also *Schad v. Borough of Mount Ephraim*, 452 U.S. 61, 65 (1981) (“Entertainment, as well as political and ideological speech, is protected; motion pictures, programs broadcast by radio and television, and live entertainment, such as musical and dramatic works fall within the First Amendment guarantee.”).⁹ Thus the “First Amendment... safeguards the storytellers and artists who take the raw materials of life—including the stories of real individuals, ordinary or extraordinary—and transform them into art, be it articles, books, movies, or plays.” *Sarver v. Chartier*, 813 F.3d 891, 905 (9th Cir. 2016).

¹⁰ See *Joseph Burstyn, Inc.*, 343 U.S. at 501 (“It is urged that motion pictures do not fall within the First Amendment’s aegis because their production, distribution, and exhibition is a large-scale business conducted for private profit. We cannot agree.”); *303 Creative LLC v. Elenis*, 143 S. Ct. 2298, 2316 (2023) (rejecting argument that speech receives lesser First Amendment protection when sold for a profit or by a corporation: “none of that makes a difference.”).

A. Strict Scrutiny Would Apply to a Federal Digital-Replica Right.

A digital-replica right would constitute a content-based restriction of speech. In *Reed v. Town of Gilbert, Arizona*, the Supreme Court made clear that a content-based law is “presumptively unconstitutional” and subject to the most demanding level of constitutional review: strict scrutiny.¹¹ *Reed* explained that “[g]overnment regulation of speech is content based if a law applies to particular speech because of the topic discussed or the idea or message expressed.”¹² And a law that regulates speech based on “particular subject matter” is “obvious[ly]” a content-based law subject to strict scrutiny.¹³ This is true “regardless of the government’s benign motive, content-neutral justification, or lack of animus toward the ideas contained in the regulated speech.”¹⁴ *Reed*’s “stringent standard” applies to all content-based restrictions of speech.¹⁵ Thus, applying *Reed*, the Ninth Circuit held that “California’s right of publicity law clearly restricts speech based upon its content,” and therefore must survive strict scrutiny to be constitutional.¹⁶

1. A Federal Digital-Replica Right Must Serve a Compelling Government Interest.

As with any other content-based restriction on speech, a law establishing a digital-replica right could clear the strict-scrutiny hurdle only if it serves a compelling governmental interest.¹⁷ Replacement of living professional performers with digital replicas without their consent might be found by courts to provide a compelling state interest sufficient to satisfy constitutional requirements, as such uses could interfere with those performers’ ability to earn a living.

¹¹ 576 U.S. 155, 163 (2015).

¹² *Id.*

¹³ *Id.*

¹⁴ *Id.* at 165 (internal quotation marks omitted).

¹⁵ *Nat’l Inst. of Fam. & Life Advoc. v. Becerra*, 138 S. Ct. 2361, 2371 (2018).

¹⁶ *Sarver*, 813 F.3d at 903; see also Eugene Volokh, *Freedom of Speech and the Right of Publicity*, 40 Hous. L. Rev. 903, 912 n.35 (“The right of publicity is clearly content-based: It prohibits the unlicensed use of particular content (people’s name or likenesses).... But even if it’s seen as content-neutral, strict scrutiny is still the proper test, because the right of publicity doesn’t leave open ample alternative channels for the speaker to convey the content that he wishes to convey.”).

¹⁷ *Reed*, 576 U.S. at 163.

2. A Federal Digital-Replica Right Must be Narrowly Tailored.

But a digital-replica right must also clear a second constitutional hurdle: the requirement that any law establishing such a right be “narrowly tailored to serve” the asserted compelling government interest.¹⁸ This is not merely a formal legal requirement, but a highly practical one that goes to the heart of filmmakers’ and others’ freedom to use technology to enhance creative process, for the ultimate benefit of audiences. Digital replicas are not something that Congress should simply aim to regulate out of existence. To the contrary, digital replicas have the potential to be an important new technological means for creators to tell their stories. Digital replicas have myriad entirely legitimate uses, ones that are fully protected by the First Amendment, and which must remain outside the scope of any digital-replica statute for it to survive strict scrutiny. Digital replica technology follows in a long line of technological innovations in depictions of individuals that allow creators to achieve their visions. Examples include using a real person’s actual image (e.g., clips of interviews with real individuals in the end credits of *I, Tonya*); or using prosthetics, makeup, and visual effects to make an actor more resemble the real person he or she is portraying (e.g., Gary Oldman as Sir Winston Churchill in *The Darkest Hour*; Nicole Kidman as Virginia Woolf in *The Hours*). No one questions that the First Amendment protects a creator’s ability to use these and similar techniques to bring verisimilitude to their work. Technology simply allows the filmmaker to do the same thing with greater realism. It supports the audience’s immersion in the story and suspension of disbelief, which are critical elements of cinematic storytelling. Realism, whether enhanced by technology or not, can bring to life the historical and cultural markers that enrich a story and bring the creator’s artistic vision to life.

Digital replicas could also be highly effective tools for parody and satire, forms of social or political commentary that the Supreme Court has held deserve high levels of protection.¹⁹ Imagine, for example, a late-night comedy show using digital replicas to poke fun at celebrities, politicians, athletes, or whoever happens to be in the news that week. Or documentarians could use digital replicas to re-create scenes from history where no actual footage exists, to enhance the visual appearance and verisimilitude of the scene (with disclosures where appropriate).

IV. BASELINE CONSIDERATIONS FOR CONTOURS OF ANY POTENTIAL FEDERAL DIGITAL-REPLICA RIGHT

A. Scope of the Right

Given the high hurdles erected by the Supreme Court’s strict-scrutiny requirements, it is imperative that any statute establishing a digital-replica right clearly and expressly avoid encroaching on First Amendment-protected uses such as those described above. Thus, any

¹⁸ *Id.*

¹⁹ See *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994); *Hustler Mag., Inc. v. Falwell*, 485 U.S. 46 (1988).

right must be limited to the use of a digital replica to replace a living professional performer, where the replica is used to perform a role, and the performance is of the type that the professional performer regularly engages in for gain or livelihood. Such a formulation would ensure that the MPA's members and other creators remain free to use technology to *depict* (rather than *replace*) individuals engaging in performances—including musicians and actors—in expressive works such as biopics and parodies. And, as with existing state right-of-publicity laws, any federal digital-replica right must include clear statutory exemptions to provide certainty to both creators and depicted individuals, which helps avoid unnecessary litigation as well as constitutional vagueness and overbreadth concerns. At a minimum, a bill establishing a federal digital-replica right must include exemptions where the use is in a work of political, public interest, educational, or newsworthy value, including comment, criticism, or parody, or similar works, such as documentaries, docudramas, or historical or biographical works, or a representation of an individual as himself or herself, regardless of the degree of fictionalization, and for uses that are *de minimis* or incidental.²⁰

B. Limitation to Living Performers

Additionally, to survive strict scrutiny, any digital replica-right must apply only to living performers. Some state right-of-publicity laws apply *post mortem*. However, those laws generally cover only commercial speech (i.e., advertising and merchandising uses)—not uses in fully First Amendment-protected speech like motion pictures. As discussed above, a court could determine that certain unconsented uses of digital replicas to replace living actors or recording artists could interfere with their ability to earn a living, establishing a compelling state interest sufficient to satisfy the constitutional requirement. However, that employment-based interest does not exist for deceased individuals. And other purported justifications for protecting deceased performers are unavailing. Any interest in a performer's reputation or dignity is already governed by defamation and privacy law. But recognizing dignitary interests of deceased individuals would represent a radical change in centuries of American law, under which “there can be no defamation of the dead.”²¹

As to financial interests, while it is understandable that heirs of deceased actors would like additional money, that desire is insufficient to overcome fundamental First Amendment rights.²² Indeed, the Supreme Court has held that the First Amendment must prevail over interests—such as national security concerns—that are orders of magnitude greater in

²⁰ States that have enacted statutes regulating the use of digital replicas have included such statutory exemptions. See N.Y. Civ. Rights Law §50-f(2)(d)(ii); La. Stat. Ann. § 51:470.5.

²¹ Restatement (Second) of Torts § 560 (1977); see also, e.g., *Bradt v. New Nonpareil Co.*, 108 Iowa 449 (Iowa Sup. Ct. 1899) (“The rule that an heir may recover for a libel of one deceased does not seem to have gained a foothold in this country, and we know of no principle that will sustain such an action.”); *Meeropol v. Nizer*, 381 F. Supp. 29, 34-35 & n.3 (S.D.N.Y. 1974), *aff'd*, 560 F.2d 1061 (2d Cir. 1977) (rejecting defamation and invasion of privacy claims by children of convicted and executed spies Julius and Ethel Rosenberg over statements in book because such claims expire upon the death of the subject of the statements at issue).

²² See *Sarver*, 813 F.3d at 905–06.

importance than the financial position of performers' heirs or their corporate successors.²³ To put it in the frame of strict scrutiny, heirs' desire to profit here is not a governmental interest at all, much less a compelling governmental interest required to satisfy incursions into a fundamental constitutional right.

C. Preemption

Any federal statute establishing a digital-replica right must preempt existing state laws to the extent that they apply to the use of digital replicas in expressive works. While many state right-of-publicity statutes contain express statutory expressive-works exemptions, not all do, and the case law regarding the proper test for evaluating First Amendment defenses in this context is in disarray.²⁴ If there is to be a federal digital-replica right, it must be carefully crafted to avoid interference with First Amendment rights, and should provide national uniformity.

MPA thanks the Subcommittee for its attention to this issue of critical importance to our members and stands ready to work with Congress and stakeholders to arrive at a solution that respects performers' concerns, while safeguarding First Amendment protections and our members' creative freedoms.

July 19, 2023

²³ See *New York Times Co. v. United States*, 403 U.S. 713 (1971) (Pentagon Papers case); see also Diane Leenheer Zimmerman, *Money As A Thumb on the Constitutional Scale: Weighing Speech Against Publicity Rights*, 50 B.C. L. Rev. 1503, 1514, 1524 (2009) (stating that the strict scrutiny test "normally requires a clear showing of necessity for the protection of human life, public safety, or national security—not profit aggrandizement" and arguing that economic interests do not justify encroachment on First Amendment rights, except in highly unusual circumstances where denying a celebrity's claim would result in "significant, highly particularized, and non-speculative" risk of "failure in the primary market for his talents").

²⁴ See Amicus Brief of 31 Constitutional Law And Intellectual Property Law Professors as Amici Curiae in Support of Petitioner in *Electronic Arts, Inc. v. Davis* (Supreme Court Case No. 15-424) (identifying five different tests: 1) the transformative use test, 2) the transformative work test; 3) the relatedness test; 4) the predominant purpose test; and 5) the balancing test)), <https://www.scotusblog.com/wp-content/uploads/2015/11/15-424-Amici-Brief.pdf>.



July 12, 2023

Screen Actors Guild – American Federation of Television and Radio Artists (“SAG-AFTRA”) submits the following comments for the record concerning the hearing *Artificial Intelligence and Intellectual Property – Part II: Copyright*. We thank the subcommittee for its attention to these important issues, particularly in the face of rapid developments in artificial intelligence (“AI”).

A number of recent hearings in the House and Senate have brought attention to potential risks and harms AI poses to working class Americans. It is critical that we recognize this as more than a fight between multi-billion dollar global companies — technology versus content — that some envision. This is an existential issue for Americans.

SAG-AFTRA members’ lives and careers depend on their ability to protect the fruits of their labor. We represent voice actors whose familiar voices read our favorite books, give voice to our favorite characters, tell us what music we are listening to, sing our favorite songs, or even narrate the actions of penguins living in remote climates. Others are the familiar faces who report the news, play our favorite characters, or host our favorite shows.

SAG-AFTRA’s members invest considerable effort into building their professional careers. While most are never “famous,” their names, voices, images or likenesses – their very persona – attain commercial value. The very cornerstone of their careers is their ability to protect their voice and likeness rights, and exploit their rights in these intangible, but often very valuable, assets, if they so choose.

The value of one’s name, voice, image, or likeness, among other aspects of one’s persona, is protected at the state level by laws such as the right of publicity. The right of publicity is the inherent right of every human being to control the commercial use of his or her identity.¹ Half a century ago, the Supreme Court described this right as “closely analogous to the goals of patent and copyright law, focusing on the right of the individual to reap the reward of his endeavors and having little to do with protecting feelings or reputation.”² The “rationale for [protecting the right of publicity] is the straightforward one of preventing unjust enrichment by the theft of goodwill” because “[n]o social purpose is served by having a defendant get free some aspect of the plaintiff that would have market value for which he would normally pay.”³

¹ Thomas J. McCarthy, *The Rights of Publicity & Privacy*, § 1:3 (2015).

² *Zacchini v. Scripps-Howard Broadcasting Co.*, 433 U.S. 562, 573(1977)

³ *Id.* at 576

But, this critical state intellectual property right varies from state to state, both in its scope and duration. The lack of federal protection leaves many without adequate protection, particularly against online platforms and AI giants who have far more access to legal resources than the average individual.

Researchers are rapidly introducing new, innovative, and sometimes abusive, AI technologies that implicate these important rights. Among the particularly concerning technologies are those that realistically clone individuals' images and voices, particularly when technology can make

people say and do things they never said or did. Those same technologies can also be used to create entire performances.

People are predisposed to trust what we see and hear and are even less inclined to seek verification of video and audio accounts of events.⁴ People are bad at detecting fake stories, even when they are purely text-based.⁵ Deepfakes likely will exacerbate this phenomenon because of the persuasive power of audio and video, which allow people to be first hand witnesses to an event without the need to trust someone else's account.⁶

This puts SAG-AFTRA's members at heightened risk from technology that allows the creation or manipulation of digital clones of their images, voices, and performances. Not only can their digital clone be made to perform against the individual's will, it can steal the work that is the cornerstone of that person's profession. Many areas of work are at risk from unchecked technology, but few so personally as the work performed by SAG-AFTRA's members and future members, who can be replaced by digital versions of *themselves*. This is very different from technology reducing the number of available jobs, this is your digital self replacing your human self. Times have changed.

The right of publicity ensures that individuals have the sole right to control how they are exploited in the public eye. It protects the individual's proprietary interest in their act, in part as an inducement to creating entertainment, but also to protect its economic value.⁷ Much like the act at issue in *Zacchini*, AI is usurping individuals' entire acts by creating digital clones that can act in their stead. It is taking away the value of their "own talents and energy, the end result of much time, effort, and expense."⁸

Although the right of publicity has long been recognized as an intellectual property right, including by the US Supreme Court, it remains a state-by-state right. Consequently, the courts have treated it as a lesser right, chipping away at it in the name of speech. With advances in technology, the need for robust protection has become even more critical.

⁴ Robert Chesney and Danielle Citron, 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics' (Foreign Affairs, Jan-Feb 2019)

<www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>

⁵ Holly Kathleen Hall, 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27 The Catholic University Journal of Law & Technology 51, 56

⁶ *Id.*

⁷ *Zacchini v. Scripps-Howard Broadcasting Co.*, 433 U.S. at 573.

⁸ *Id.* at 576

Accordingly, we believe the time is now to consider federal legislation to protect against voice and likeness misappropriation.

A handwritten signature in black ink, appearing to read 'JP Bennett', with a long horizontal flourish extending to the right.

JEFFREY P. BENNETT
General Counsel



July 18, 2023

The Honorable Chris Coons
218 Russell Senate Office Building
Washington DC 20510

The Honorable Thom Tillis
113 Dirksen Senate Office Building
Washington DC 20510

Re: SJC IP Subcommittee June 12 Hearing: "Artificial Intelligence and Intellectual Property – Part II: Copyright"

Dear Chairman Coons and Ranking Member Tillis,

On behalf of the tens of thousands of independent American music creators whom our organizations represent, The Songwriters Guild of America ("SGA"), the Society of Composers & Lyricists ("SCL"), and the Music Creators North America ("MCNA") coalition applaud and thank the Senate Judiciary Committee's IP Subcommittee for its steadfast engagement on copyright issues. We especially appreciate the Subcommittee's efforts on July 12, 2023, at its hearing titled [Artificial Intelligence and Intellectual Property—Part II: Copyright](#), to ensure that the voice and viewpoint of a working artist and creator (concept artist Ms. Karla Ortiz) was heard loud and clear on the urgent issue of ensuring adequate protections for creators at the very threshold of the generative AI revolution.

Ms. Ortiz's passionate articulation of the dangers posed to the U.S. creative community --and to the future of American culture and commerce-- by the unlicensed and unregulated ingestion of copyrighted works into generative AI systems was perhaps the most valuable contribution to the public's understanding of global, generative AI issues so far. Through this letter, we seek to respectfully have our own voices heard on behalf of songwriters, composers and lyricists on the importance of fairness in the generative AI discussion, including our emphatic endorsement of the "Three C's" concept of "Consent, Credit and Compensation," that must serve as the bedrock protections for creators and copyright owners on which U.S. generative AI policy rests.

We also join with those hearing witnesses, including Ms. Ortiz and General Counsel Jeffrey Harleston of UMG, who expressed an *urgent* need for legislative safeguards to be expeditiously put into place in order to encourage the orderly institution of fair licensing practices by AI system operators-- *before* the massive infringements we've seen in the past regarding unauthorized digital uses of music repeat themselves. The music industry is well positioned through its recent and past experiences and voluntary licensing practices to fulfill any large-scale needs that may arise among generative AI system operators. To be succinct, the "defense" offered by Stability AI at the hearing --that the scope of its unauthorized takings is too broad to give rise to infringement liability due to the hardship of the licensing process --is as specious as it sounds.

Further along those same lines, we wish to impress upon the Subcommittee the crucial need for members of the independent music community (as well as independent creators working in other artistic disciplines) to have the ability to collectively negotiate, issue and administrate licenses with generative AI system operators. An antitrust exemption that allows the establishment and operation of opt-in, voluntary

collective licensing organizations below a certain market size will serve not only the interests of individual, independent creators and copyright proprietors currently unable to compete on a practical basis with the world's three major music conglomerates (which together represent nearly eighty percent of the world's musical recordings and compositions), it will also better serve --through streamlining the licensing process-- the needs of the technology sector and the public market it seeks to satisfy.

The following information more fully enumerates our positions, and is offered in support of the above points:

1. We first wish to emphasize the importance of the following comments ("Alliance Comments") recently submitted by the US Copyright Alliance (of which our organizations are members) to the United States Office of Science and Technology Policy ("OSTP") and rooted in Article I Section 8 of the US Constitution: "[t]he marketplace should continue to properly value and incentivize creativity, and AI policy should not interfere with the right of copyright owners to license, or choose not to license, their works for AI uses." (Alliance Comments at 2). In other words, the generative AI-related principles of "consent, credit and compensation" are the watchwords of the entire, united entertainment communities of America, which together contribute well over a trillion dollars to the US gross national product and provide immense cultural prestige for the United States throughout the world.

Moreover, we believe it is equally important to make clear our perspective that the full scope of current and future generative AI uses, inputs and applications --whether now known or hereafter created-- are intended to be included within these principles. As such, the exact nature and scope of many of those current and prospective uses of our musical works, our names and our likenesses remain unforeseeable, or have not yet been fully disclosed or understood. That reality may require future consideration of the codification of additional, *sui generis* rights and protections for creators and copyright owners that of necessity go beyond current copyright laws and concepts. Further discussion of this issue *prior to the widespread, unlicensed ingestion and incorporation of our works and other aspects of our creative lives into AI systems* is to us a matter of great importance.

2. We likewise seek to underline the important assertion in the Alliance Comments that "[a]dequate transparency regarding ingestion of copyrighted works also goes a long way in helping to ensure that copyright owners' rights are respected. In particular...it is vital that AI developers maintain records of which copyrighted works are being ingested and how those works are being used, and make those records publicly accessible as appropriate...." (Copyright Alliance Comments at 3).

While such Comments omit from disclosure obligations those AI developers that are also the copyright owners of the works being ingested by the AI system, we believe it is essential to the integrity of any system designed to ensure fairness through transparency that information concerning *all* ingestive uses of copyrighted musical works be mandated for disclosure. Going this further step will serve both the interests of songwriters, composers and lyricists whose works are being incorporated into any and all AI systems,

and accrue to the benefit of a public that maintains the right to know the nature and sources of the AI-generated entertainment and information it is consuming.

3. One of the most compelling points made in the Alliance Comments pertains to concerns over the frequent attempts of the technology sector to expand the exception of “fair use” to the point of swallowing whole the general rules of copyright and creator protections. As the Alliance Comments state, “[s]ome AI developers assert that AI innovation will be impeded if the ingestion process is not deemed to be categorically fair use. This false narrative purposely omits the fact that in many cases licenses are available for ingestion. As history has shown, creators and copyright owners are usually willing to license their works when the parties can agree on appropriate terms and compensation; that is, of course, how creators typically earn a living. Copyrighted works provide immense value to AI developers, and they can and should pay for that value—as many are already doing. In other words, when properly applied, copyright law sets the conditions for the market to prevail.” (Alliance Comments at 3).

Again, our independent music creator groups wish to emphasize our concurrence, and would again go a step further by respectfully requesting that Congress act swiftly to make clear that ingestion of copyrighted works by AI systems without copyright owner permission is presumptively unfair and thereby infringing.

4. Another crucial point made in the Alliance Comments that we wish to highlight is the obvious conclusion that “[i]n some cases, AI-generated outputs may also infringe works ingested by the AI system.” (Alliance Comments at 3). We respectfully suggest that Congress should re-emphasize in law the black letter principle that permission to allow AI systems to ingest copyrighted works for the purpose of generating algorithmically machined content *in no way exempts those systems from liability for copyright infringement in the event that such generative output includes works substantially similar to one or more of the works ingested by the system.* Moreover, even in rare instances (if any) in which unauthorized ingestion of a copyrighted musical work is judged to be non-infringing, such finding should still not serve as a shield against copyright infringement liability for the owner or controller of the AI system concerning substantially similar generative *output.*
5. We similarly wish to focus attention on the fact that “[q]uestions about infringement, copyrightability, and liability related to the ingestion of copyrighted works and the output of works created using AI are topics that the Copyright Office has already begun exploring ... as its study on copyright and AI progresses...” (Alliance Comments 5-6).

In that regard, we believe we would be remiss in not specifically conveying our strong support for a necessary increase in funding for the Copyright Office. The heavy reliance on the Office’s expertise and enterprise that will be necessary to ensure the success of a variety of new and anticipated legislative mandates to help protect the rights of music creators and other authors and copyright owners --including the provision of

knowledgeable insight regarding the myriad, economically crucial issues concerning generative AI and potentially related *sui generis* rights-- requires that the Office be given the proper resources to carry out its responsibilities. From our perspective, the Office is currently severely under-staffed and under-funded, and we ask both the legislative and executive branches to study this issue carefully on an immediate basis and take whatever remedial action it deems appropriate given the importance of the creative IP sector to the US economy. (See, Alliance Comments at 4). We repeat, however, that the consultation process should not interfere with the expeditious action of Congress on these issues.

6. Finally, while we wholeheartedly concur with the Alliance Comments concerning the need for careful consideration prior to formulation and implementation of US Government policies related to generative AI, we also want to make clear our strong belief that addressing the matters raised in our letter today represents an urgent financial, cultural and trade priority requiring the immediate attention of Congress. As noted above in our introduction, we are especially firm in our belief that the need to establish a collective licensing option for generative AI to serve the independent creator and copyright owner communities is of paramount importance.

Specifically, maintaining competition in the marketplace is predicated on the ability of all participants in the value chain to have an opportunity to engage in commerce on a level playing field (or as close to level as consolidation and vertical integration within the music industry and other segments of the entertainment industries will effectively permit). Independent music creators not affiliated with the three major, global music conglomerates (which as previously noted together control a vast majority of the world's music copyrights), for example, must somehow be afforded the practical ability to compete fairly in the AI licensing space. Consideration of an opt-in system for voluntary, collective negotiation, licensing and administration by the independent creator community should therefore be the subject of immediate consideration by Congress as perhaps the best way to ensure that our independent musical works are neither shut out of the AI licensing market --nor infringed with impunity-- due to the lack of a viable, voluntary system of licensing, collective rate negotiation, and administration that the major music conglomerates already individually possess *on an even larger scale*.¹

¹ <https://musically.com/2023/04/26/music-copyright-publishes-its-market-share-analysis-for-2022/> The world's largest music conglomerate already controls more musical composition copyrights than the entire universe of independent songs, with the combined market shares of the next two largest music conglomerates *surpassing* the holdings of both the largest and the combined independents. Suggestions by these corporations that seeking an exemption letter from the U.S. Department of Justice (a process that could take years) would be a better choice for addressing collective licensing needs of independents represents a blatantly ingenuous attempt to derail the process in the pursuit of commercial advantage. Similarly, suggestions that unofficial representatives of the independents first attempt to negotiate deals with AI systems to which independent creators and copyright owners might later opt-in, is a fools' errand guaranteed to result in under-market licensing offers to independent creators and copyright owners, producing unfair market advantages that would again accrue in numerous ways to the major corporations.

In closing, we thank the Subcommittee for the opportunity to present our comments for consideration, and look forward to responding to any further requests for information or clarification.

Respectfully submitted,



Rick Carnes
President, Songwriters Guild of America
Officer, Music Creators North America



Ashley Irwin
President, Society of Composers & Lyricists
Co-Chair, Music Creators North America



Eddie Schwartz
President, Music Creators North America

cc:
Members of the Senate Judiciary Committee
Members of the Senate AI Caucus
Senate Majority Leader and Minority Leader
US Copyright Office
Charles J. Sanders, Esq.

ABOUT OUR ORGANIZATIONS:

The Songwriters Guild of America (SGA) is the longest established and largest music creator advocacy and copyright administrative organization in the United States run solely by and for songwriters, composers, and their heirs. Its positions are reasoned and formulated independently and solely in the interests of music creators, without financial influence or other undue interference from parties whose interests vary from or are in conflict with those of songwriters, composers, and other authors of creative works. Established in 1931, SGA has for over 90 years successfully operated with a two-word mission statement: "Protect Songwriters," and continues to do so throughout the United States and the world. SGA's organizational membership stands at approximately 4500 members. For more information: <https://www.songwritersguild.com/site/index.php>

The Society of Composers & Lyricists (SCL) is the premier U.S. organization for music creators working in all forms of visual media. With chapters in Los Angeles, New York and Nashville, and members in every state of the U.S. as well as over 80 countries around the world, the SCL operates as the primary voice for over 3,000 members who work as creators of scores and songs for film, television, video games, and theatre. The SCL is a founding co-member -- along with SGA and other independent music creator groups -- of Music Creators North America (MCNA). For more information: <https://thescl.com>

Music Creators North America (MCNA) is an alliance of independent songwriter and composer organizations that advocates and educates on behalf of North America's music creator community. As the only internationally recognized voice of American and Canadian songwriters and composers, MCNA, through its affiliation with the International Council of Music Creators (CIAM), is part of a coalition that represents the professional interests and aspirations of more than half a million creators across Africa, Asia, Austral-Oceania, North and South America, and Europe. For more information: <https://www.musiccreatorsna.org/>

SGA, SCL and MCNA are also founding members of the international organization **Fair Trade Music International**, which is the leading US and international advocacy group for the principles of transparency, equitable treatment, and financial sustainability for all songwriters and composers. For more information: <https://www.fairtrademusicinternational.org/>

