

ARTIFICIAL INTELLIGENCE AND INTELLECTUAL
PROPERTY: PART I—INTEROPERABILITY
OF AI AND COPYRIGHT LAW

HEARING

BEFORE THE

SUBCOMMITTEE ON COURTS, INTELLECTUAL
PROPERTY, AND THE INTERNET

OF THE

COMMITTEE ON THE JUDICIARY
U.S. HOUSE OF REPRESENTATIVES

ONE HUNDRED EIGHTEENTH CONGRESS

FIRST SESSION

WEDNESDAY, MAY 17, 2023

Serial No. 118–20

Printed for the use of the Committee on the Judiciary



Available via: <http://judiciary.house.gov>

U.S. GOVERNMENT PUBLISHING OFFICE

WASHINGTON : 2023

COMMITTEE ON THE JUDICIARY

JIM JORDAN, Ohio, *Chair*

DARRELL ISSA, California	JERROLD NADLER, New York, <i>Ranking Member</i>
KEN BUCK, Colorado	ZOE LOFGREN, California
MATT GAETZ, Florida	SHEILA JACKSON LEE, Texas
MIKE JOHNSON, Louisiana	STEVE COHEN, Tennessee
ANDY BIGGS, Arizona	HENRY C. "HANK" JOHNSON, JR., Georgia
TOM McCLINTOCK, California	ADAM SCHIFF, California
TOM TIFFANY, Wisconsin	DAVID N. CICILLINE, Rhode Island
THOMAS MASSIE, Kentucky	ERIC SWALWELL, California
CHIP ROY, Texas	TED LIEU, California
DAN BISHOP, North Carolina	PRAMILA JAYAPAL, Washington
VICTORIA SPARTZ, Indiana	J. LUIS CORREA, California
SCOTT FITZGERALD, Wisconsin	MARY GAY SCANLON, Pennsylvania
CLIFF BENTZ, Oregon	JOE NEGUSE, Colorado
BEN CLINE, Virginia	LUCY McBATH, Georgia
LANCE GOODEN, Texas	MADELEINE DEAN, Pennsylvania
JEFF VAN DREW, New Jersey	VERONICA ESCOBAR, Texas
TROY NEHLS, Texas	DEBORAH ROSS, North Carolina
BARRY MOORE, Alabama	CORI BUSH, Missouri
KEVIN KILEY, California	GLENN IVEY, Maryland
HARRIET HAGEMAN, Wyoming	
NATHANIEL MORAN, Texas	
LAUREL LEE, Florida	
WESLEY HUNT, Texas	
RUSSELL FRY, South Carolina	

SUBCOMMITTEE ON COURTS, INTELLECTUAL PROPERTY, AND
THE INTERNET

DARRELL ISSA, California, *Chair*

THOMAS MASSIE, Kentucky	HENRY C. "HANK" JOHNSON, JR., Georgia,
SCOTT FITZGERALD, Wisconsin	<i>Ranking Member</i>
CLIFF BENTZ, Oregon	TED LIEU, California
BEN CLINE, Virginia	JOE NEGUSE, Colorado
LANCE GOODEN, Texas	DEBORAH ROSS, North Carolina
KEVIN KILEY, California	ADAM SCHIFF, California
NATHANIEL MORAN, Texas	ZOE LOFGREN, California
LAUREL LEE, Florida	MADELEINE DEAN, Pennsylvania
RUSSELL FRY, South Carolina	GLENN IVEY, Maryland

CHRISTOPHER HIXON, *Majority Staff Director*
AMY RUTKIN, *Minority Staff Director & Chief of Staff*

C O N T E N T S

WEDNESDAY, MAY 17, 2023

	Page
OPENING STATEMENTS	
The Honorable Darrell Issa, Chair of the Subcommittee on Courts, Intellectual Property, and the Internet from the State of California	1
The Honorable Henry C. “Hank” Johnson, Ranking Member of the Subcommittee on Courts, Intellectual Property, and the Internet from the State of Georgia	3
The Honorable Jerrold Nadler, Ranking Member of the Committee on the Judiciary from the State of New York	4
WITNESSES	
Sy Damle, Partner, Latham & Watkins LLP	
Oral Testimony	7
Prepared Testimony	9
Chris Callison-Burch, Associate Professor, Computer and Information Science, University of Pennsylvania	
Oral Testimony	25
Prepared Testimony	27
Ashley Irwin, President, Society of Composers and Lyricists	
Oral Testimony	51
Prepared Testimony	53
Dan Navarro, Grammy-nominated songwriter, singer, recording artist, and voice actor	
Oral Testimony	58
Prepared Testimony	60
Jeffrey Sedlik, President & CEO, PLUS Coalition	
Oral Testimony	62
Prepared Testimony	64
LETTERS, STATEMENTS, ETC. SUBMITTED FOR THE HEARING	
All materials submitted by the Subcommittee on Courts, Intellectual Property, and the Internet, for the record are listed below	99
Materials submitted by the Honorable Darrell Issa, Chair of the Subcommittee on Courts, Intellectual Property, and the Internet from the State of California, for the record	
An article entitled, “Will AI Value Human Creators? Now’s the Time to Decide the Future of Our Culture (Guest Column),” May 16, 2023, Billboard	
A sample of AI entitled, “Make the argument that AI-generated material should not be copyrightable,” AI-generated	
A sample of AI entitled, “Make the argument that AI-generated material should be copyrightable,” AI-generated	

APPENDIX

Materials submitted by the Honorable Darrell Issa, Chair of the Subcommittee on Courts, Intellectual Property, and the Internet from the State of California, for the record

A paper entitled, "Position Paper—Artificial Intelligence," Copyright Alliance

Statement by the Honorable Darrell Issa

A summary from Stability.AI

QUESTIONS AND RESPONSES FOR THE RECORD

Responses to questions from the Honorable Darrell Issa, Chair of the Subcommittee on Courts, Intellectual Property, and the Internet from the State of California, for the record

Response from Stability.AI

A summary from Stability.AI

**ARTIFICIAL INTELLIGENCE AND
INTELLECTUAL PROPERTY: PART I—
INTEROPERABILITY OF AI AND
COPYRIGHT LAW**

Wednesday, May 17, 2023

HOUSE OF REPRESENTATIVES

SUBCOMMITTEE ON COURTS, INTELLECTUAL PROPERTY, AND
THE INTERNET

COMMITTEE ON THE JUDICIARY

Washington, DC

The Committee met, pursuant to notice, at 10:04 a.m., in Room 2141, Rayburn House Office Building, Hon. Darrell Issa [Chair of the Subcommittee] presiding.

Members present: Representatives Issa, Fitzgerald, Bentz, Gooden, Cline, Kiley, Moran, Lee, Fry, Johnson of Georgia, Nadler, Lieu, Ross, Schiff, Lofgren, Dean, and Ivey.

Also present: Representative Jackson Lee.

Mr. ISSA. The Committee will come to order. I want to welcome our guests and welcome the Members to what is undoubtedly going to be the first of many AI hearings.

My staff, in preparation for this, knows one of my favorite jokes on artificial intelligence, which I was telling long before it is becoming a direct issue for this Committee, and that is that artificial intelligence is what Members have when their staff prepares their opening statements.

So, yes, today we stand at the intersection of two rapidly evolving domains, generative artificial intelligence and existing copyright law that must adapt to it. The advent of generative AI technologies has sparked a profound transformation in the creation, distribution, and consumption of a new form of creative work. As we embark in our legal journey along with the administration and their regulatory powers, it is vital that we explore the complex relationship between generative AI and copyright law, recognizing both the challenges and opportunities that lie ahead.

Generative AI holds immense potential for innovative and artistic expression. It empowers creators to explore new frontiers fueling their imagination with AI-generated content. Yet, with the power that comes with it comes responsibility, responsibility for negative activity on the web and of copyright laws being trampled. Copyright laws were designed to protect intellectual property.

Copyright laws also exist under our Constitution specifically to encourage and then reward creation. It is that encouragement that, in fact, creates the right of ownership, not the right of ownership having some core item.

In this hearing, we aim to strike a delicate balance. Let me rephrase that in spite of what it says here. In this hearing, we recognize there is a delicate balance, and we would hope that our panel today represents both a diversity of views and a possible collaboration.

We must consider though that there will be individuals on each side who will not want to move. There will be companies which would include some of the household words, Microsoft, Meta, which is not yet a household word, but Facebook is, and in fact, Google, that are on both sides of this issue being both massive creators of their own copyright and massive users of others.

It is our duty to adapt, refine copyright laws to accommodate the transformative potential of generative AI while safeguarding interests of existing creators and of the right of the society as a whole to benefit. Generative AI presents both challenges and opportunities for creative works and copyright holders. It requires thoughtful consideration and ongoing discussion with stakeholders to strike a balance between protecting intellectual property rights, encouraging creativity, and fostering innovative content, AI generated included.

Today, all of us here understand that generative AI is nascent. As we receive briefings, we hear about first generation, second generation, third generation, now fourth generation. By the way, the education of this young child is profound from something that was hard to understand to something that said the wrong thing, to something that now is so useable that we might often forget to fact check its output because it seems so good.

We must, first and foremost, address properly the concerns surrounding unauthorized use of copyrighted material while also recognizing the potential of generative AI can only be achieved with massive amounts of data, far more than is available outside of Copyright. By embracing a forward-thinking approach, we can establish guidelines to promote responsible and ethical practices in the realm of generative AI.

As we embark on our legal journey, let us approach it with a spirit of collaboration as those who were on the call the other day clearly were doing. Let us find common ground seeking solutions that promote the flourishing of both creative expression and intellectual property protection. The stakes couldn't be higher, and the outcome will shape the future landscape of art, technology, and copyright.

Today, let us navigate this uncharted territory to ensure that generative AI and copyright laws foster in the future a fairly harmonious and vibrant ecosystem for generations to come.

In closing, I might say that we clearly could have substituted patent or other areas of innovation and use just as easily and we would have been having substantially the same hearing, perhaps in some cases with different witnesses, and in that sense, I think it is important for us all to look at this.

I will close by saying that it has come to my attention that for once, Congress may not do either of the things we are known for which is nothing at all or overreact. That is not true around the world. Spain is moving forward with what I think might be a very restrictive interpretation. Japan believes, apparently, that all information that goes into the teaching is, in fact, free of any copyright restriction in its use. I do not believe that today's discussion will take us down either road. I believe that we will measure carefully and find middle ground that respects existing copyright law while allowing the future of generative AI to flourish. With that, I yield back and recognize the Ranking Member for his opening statement.

Mr. JOHNSON of Georgia. I thank the Chair for holding this hearing. I would assure those who are listening that Congress is getting more and more information about artificial intelligence and discussing the ramifications and all enterprises of human behavior. I share your prediction that we will act, and we will do so in a reasonable way. I don't know if we will do the Japanese hard-liner position on inputs though. I want to thank you for holding this hearing.

Artificial intelligence is a promising technology with the potential to revolutionize a range of industries, but with that promise also comes potential peril, especially to the creators of art and other copyrightable works. It is critical that we think through the many ways that AI will change our lives and whether our existing laws are up to the challenge. I am glad that we are beginning that process today.

Among the many questions that we must confront as artificial intelligence takes on a larger role in our society, is how our copyright system should view AI, both in terms of how we could treat copyrighted works that are used to train an AI model, in other words, the inputs and whether the new work that is generated by AI, in other words the outputs, should be eligible for copyright protection itself. Just as AI machine learning is a complex process, so too is our inquiry today as there may not be simple answers to any of these questions.

First, let's consider the inputs used to train AI systems. A typical generative AI system is fed vast amounts of human-authored work on which to train. This may include written word, visual art, and music. The model then processes this information with the help of various algorithms to detect patterns and probabilities. A foundational principle of copyright law generally requires users of copyrighted works to obtain the permission of the copyright owner. While much of input into generative AI systems tends to consist of works that are protected by copyright, these works are typically used without seeking consent or a license.

Some argue that this constitutes fair use and indeed this question is being litigated in the courts as we speak. I am hard pressed to understand how a system that rests almost entirely on the works of others and can be commercialized or used to develop commercial products owes nothing, not even notice, to the owners of the works it uses to power its system.

Even if we determine that AI systems must seek permission to use copyrighted works, that only leads to more questions. For ex-

ample, what sort of licensing system should be required? What would represent fair compensation for these works?

Other questions arise as well, such as what degree of transparency should be built into AI models, and how can we ensure that proper credit is attributed to copyrighted works?

On the other side of the equation is we must consider the status of the works, the outputs that are generated by the AI model. Should a purely AI generated work be eligible for copyright protection if it does not otherwise meet the definition of infringement? What if there was significant human creative involvement in generating the AI work? What about works that may be of a hybrid manner, consisting of both human authorship and elements that are purely AI? Whether or not an AI-generated work is eligible to be copyrighted, such works will compete and indeed are already competing in the marketplace against human-authored works.

No examination of AI is complete without considering the impact that AI works will have on human creators. How do we balance the need for innovation with the need to protect human creators? None of these questions have simple answers and that is why I am pleased that we are joined by such a distinguished panel of witnesses to help guide our Subcommittee's inquiry into important copyright issues raised by the use of artificial intelligence.

I am interested to hear your perspectives on these questions, as well as whether our existing copyright framework, including recent guidance issued by the U.S. Copyright Office, is sufficient to address these questions or whether legislation is needed.

I thank the Chair for holding this important hearing again and I yield back the balance of my time.

Mr. ISSA. I thank the gentleman. In lieu of the Chair's opening statement, in addition to the opening statement, I read which was a hybrid of GPT and my staff, I also will put a pure opening statement that was done through regenerative AI into the record and without objection it will be placed in the record.

I now recognize the Ranking Member of the Full Committee, Mr. Nadler.

Mr. NADLER. I thank the gentleman, thank you, Mr. Chair, and thank you for holding this important hearing.

Yesterday, some of us were privileged to attend a briefing by Sam Altman on generative AI. Thirty years ago, as the Internet was in its nascent stage, we were forced to grapple with a series of questions about how that then emerging technology would change our society and whether our laws were sufficient to address these new challenges. We are still wrestling with many of those issues today.

Now, a new technology, artificial intelligence, has emerged that has a similar potential to transform many aspects of our lives. Once again, as our technology evolves, we must determine if our laws must evolve as well.

Although AI gives rise to a myriad of issues, today we are focusing on the intersection between artificial intelligence and copyright. Even within this limited spirit of the law, I find myself with far more questions than answers as we begin this inquiry.

At its core, the fundamental question we must ask is how can we promote innovation and further development of generative artificial

intelligence models while also protecting the rights of creators whose works are the engine that fuels these models and must compete with AI generated work in the marketplace? This is not an easy balance to strike, and many answers just lead to more questions.

For example, a typical AI model trains on vast quantities on data analyzing the underlying text, images, or music as it learns to make predictions based on that information to generate new works. Most of the information which these generative AI systems rely, is found in copyrighted works, but typically, these works are used without seeking the permission of the copyright holder. Worse yet, there is little transparency as to which copyright works are used, thus preventing creators from asserting their rights.

In fact, AI systems are generating new content based on their work, content that may 1 day directly compete with their work, but without their knowledge or consent and certainly without compensation. This is a troubling development, but the solutions are far from clear. What is the proper way to license these works and to monitor compliance? What sort of transparency and accountability should be built into these systems? It is time to begin considering these important issues.

As we wrestle with these questions about how to treat the inputs into a generative AI system, there are equally thorny questions about the outputs. How should the copyright laws treat works that are generated by an AI model? The Copyright Office recently offered guidance on that question. According to the guidance, only works of, “human authorship are eligible for copyright protection.” It goes on to explain that it is necessarily a fact-specific inquiry to determine where the line is between human work that is merely assisted by AI technology and works under which the traditional elements of authorship are actually conceived and executed by the AI model itself.

To add further complexity, in some cases, certain aspects of a single work may be eligible to protection while others are not. We must consider whether the Copyright Office has the tools and resources it needs to evaluate AI-related applications and to enforce existing rules.

Finally, there are important questions about the impact that AI-generated works will have on the market for human-offered works, works that may have served as the foundation for generating the AI work in the first place.

While we work to promote innovation in the AI realm, we must also ensure that artificial works do not displace human creators or threaten the ability of the creative class to thrive. Already creators like the members of the Writers Guild of America are uniting to speak up for their own rights and to ensure they are not replaced by AI systems.

All these issues lead to a final set of questions. What, if anything, should Congress do to ensure that we strike the right balance between protecting creators and promoting innovation? Does our existing copyright framework up to the task of navigating these complex issues? Should we wait and see how AI technology evolves before taking any action or let these issues play out in the courts? There are no easy answers and I appreciate the opportunity to be

thinking through these important questions with our distinguished panel of witnesses.

I thank the Chair for holding this hearing and I yield back the balance of my time.

Mr. ISSA. I thank the gentleman. The gentleman yields back.

I now ask unanimous consent that an article titled, "Will AI Value Human Creators? Now's the Time to Decide the Future of Our Culture," to put it into the record. I now ask that this be placed in the record.

I additionally ask and I will distribute copies for unanimous consent that ChatGPT's answers for when we asked it to argue for regulation and when we asked it to argue against regulation be both put in the record. I will say that it is a convincing argument in both directions.

With that, I now have the honor of introducing our distinguished panel.

Mr. Sy Damle is a partner in copyright practice at Latham & Watkins. He is a former software engineer and specializes in technology matters including those involving computer systems and networks and artificial intelligence. He is previously General Counsel at the Copyright Office. For all those reasons and more is why you are here. Welcome.

Mr. Callison-Burch is an Associate Professor of Computer and Information Science at the University of Pennsylvania. He is also a Visiting Research Scientist at the Allen Institute for Artificial Intelligence and welcome.

Mr. Ashley Irwin, Mr. Irwin is President of the Society of Composers and Lyricists and an Emmy Award-winning Music Director, Conductor, Composer, Arranger, and Producer who has scored over 30 featured films, 300 hours of television, and over 3,000 commercials. I suspect the commercials probably made you the most as that sometimes goes.

Mr. Dan Navarro, Mr. Navarro is a Grammy-nominated songwriter, singer, recording artist, and voice actor. He was formerly in the duo Lowen and Navarro, released 13 albums and performed over 1,500—this was done by my staff, not by AI, gigs, and is currently a solo artist who tours constantly, but has blessed us with a day off the road with no compensation and for that we are very appreciative.

Mr. Jeffrey Sedlik is the President and CEO of PLUS Coalition, which is a global, nonprofit initiative to make it easy and fast to search for, find, communicate, and understand information about photographs and visual artwork. Mr. Sedlik is also a Member of the Joint Committee on Ethics in AI and a professional photographer.

I want to welcome all our witnesses, recognizing that this is a Judiciary Committee, I am going to ask you to all rise and take the oath.

Raise your right hand for the camera. Do you solemnly swear or affirm under penalty of perjury that the testimony you will give will be the truth, the whole truth, and correct to the best of your knowledge so help you God? Thank you.

Please be seated. Let the record indicate that all witnesses answered in the affirmative.

As you may know from watching CSPAN, your entire statement will be placed in the record, along with, without objection, any extraneous material you see fit now or in the next five days to include so that as we go through this, if you realize there is something that you should have said, would have said, or wanted to say you will be allowed to supplement.

With that, Mr. Damle, you are first up for five minutes. Because we have a lot of people up on the dais that want to ask you questions, I will ask that each of you try to limit right to five minutes or less.

The gentleman is recognized.

STATEMENT OF SY DAMLE

Mr. DAMLE. Chair Issa, Ranking Member Johnson, and Members of the Subcommittee, thank you for inviting me to participate in today's hearing.

I am a former software engineer, former General Counsel of the U.S. Copyright Office, and currently a partner at the law firm Latham & Watkins, LLP. I am here testifying solely in my personal capacity and not on behalf of my law firm, any of the firm's clients, or the U.S. Copyright Office.

My testimony today will focus on the copyright implications of AI training, but first I would like to put those issues into context. The AI tools of the present and near future will impact almost every aspect of the human experience. It will improve our science and our medicine. It will make our military more effective. It will make our businesses more efficient and productive. It will enable anyone to more fully unlock their creative potential. In short, AI has the potential to transform our economy and improve our society as a whole. That outcome is not guaranteed.

The way we regulate AI will directly determine whether the United States will continue to lead the world in AI development, or whether another country will take up that mantle. In considering whether to impose intellectual property-based restrictions on AI innovation, Congress should carefully evaluate whether those restrictions will hamper the development of AI here in the United States.

At the same time, artists, writers, and other creators have expressed genuine concern that the rapid development of AI will displace human authors. Policy makers should take those concerns seriously. By taking a step back, it is important to appreciate that every new technological development has caused similar fears and in hindsight, we can see that those fears have not come to fruition. To just take one example, when photography was invented in the mid-1800's, one prominent critic dismissed the medium as, "the refuge of every would-be painter too ill-endowed or too lazy to complete his studies." He predicted that photography would "corrupt art altogether."

Now, I am sure Mr. Sedlik can tell us how much he disagrees with that sentiment as an accomplished photographer. Of course, society embraced the camera as a creative tool and photography blossomed as an art form that deepened rather than diminished in the field of human creativity.

There is no reason to believe that generative AI is any different. Like the camera, or the many creative tools adopted since, generative AI will be an engine of human creativity, not a replacement for it.

So, with that context in mind, I want to make two points today.

First, Copyright's well established fair use doctrine is the best way to balance the competing interest in the AI space. For the reasons I explain in detail in my written testimony, the training of AI models will generally fall within the established bounds of their use.

While some AI models may very well exceed those bounds, our courts are well equipped to handle those situations.

The concerns that some have raised that AI models can replicate artistic styles are completely understandable. The solution does not lie in copyright law which does not and has never granted monopolies over artistic or musical style.

Second, some groups have proposed a collective licensing regime for AI training data. Such a regime would eliminate fair use in this area, replacing it with a rigid assumption that AI training is infringing. I believe that would be a mistake. Plus, if Congress were nevertheless interested in setting up a collective licensing regime, it should be aware of some of the serious practical challenges it will face. I detail those challenges in my written testimony, but fundamentally, they are rooted in the fact that successfully training an AI model requires using many billions of pieces of content. That is many orders of magnitude larger than the number of works covered by any similar scheme in the history of American law.

In short, Congress has already adopted a copyright right that is technology neutral and flexible enough to balance the need for a dynamic domestic AI industry with the right co-creators. I look forward to answering your questions today.

[The prepared statement of Mr. Damle follows:]

STATEMENT OF SY DAMLE

BEFORE THE U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON THE
JUDICIARY, SUBCOMMITTEE ON COURTS, INTELLECTUAL PROPERTY, AND
THE INTERNET

ON

“ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY: PART I—
INTEROPERABILITY OF AI AND COPYRIGHT LAW”

MAY 17, 2023

Chairman Issa, Ranking Member Johnson, and Members of the Subcommittee, thank you for inviting me to participate in this hearing on artificial intelligence (“AI”) and copyright law. I am a former software engineer, a former General Counsel and Associate Register of Copyrights of the United States Copyright Office, and current partner at the law firm of Latham & Watkins LLP. I speak solely in my personal capacity, and not on behalf of my law firm, any of the firm’s clients, or the U.S. Copyright Office. In addition, my testimony is not intended in any way to constitute legal advice and should not be relied on as such.

I. Introduction and Executive Summary

My statement will focus on the copyright implications of AI training and, in particular, whether AI models should be permitted to “learn” from copyrighted works. But first I would like to put these issues into context. The AI tools of the present and near future will impact almost every aspect of the human experience. They will improve our science and medicine. They will make our military more effective. They will make our businesses more efficient and productive. They will transform the way humans learn and work. They will enable anyone to more fully unlock their creative potential. In short, AI has the potential to transform our economy and improve our society as a whole.

But that outcome is not guaranteed. The way we regulate AI will directly determine whether the United States will continue to lead the world in AI development, or whether another country will take up that mantle. Imposing heavy-handed, intellectual-property-based restrictions on AI innovation will hamper the development of AI here in the United States, and likely drive that development to other countries.

At the same time, artists, writers, and other creators have expressed genuine concern that the rapid development of AI—and, in particular so-called “generative” AI models that create text, software code, visual works, music, and other media—will displace human authors. Policymakers should take those concerns seriously. But every new technological development has led to similar fears, and in hindsight, none of those fears came to fruition. For instance, after photography was invented in the mid-1800s, critics dismissed the medium as the “refuge of every would-be painter . . . too

ill-endowed or too lazy to complete his studies,” and worried that if photography “is allowed to supplement art,” it would “corrupt[] [art] altogether.”¹ But society embraced the camera as a creative tool, and photography blossomed as an art form that deepened, rather than diminished, the field of human creativity.² There is no reason to believe generative AI is any different. Like the camera or the myriad creative tools adopted since, generative AI will be not a replacement for, but an engine of human creativity.

With that context in mind, I want to make two points today.

First, copyright’s well-established fair use doctrine is the best way to balance the competing interests in the AI space. Fair use is a flexible doctrine in which our nation’s courts explore the specific facts of every case, and determine whether allowing the particular use at issue would further—or obstruct—the Constitutional goal of copyright law, which is to promote the creation and spread of knowledge and learning.³ Over the course of hundreds of judicial decisions, a simple principle has emerged: that the use of a copyrighted work to learn unprotectable facts and use those facts to create products that do not themselves infringe copyright is quintessential fair use. In general, the training of AI models adheres to that principle. Today’s AI leaders have built their innovative products relying on that understanding. Indeed, the uniquely American fair use doctrine is, in large part, why the United States is the epicenter of global AI development.

Some have suggested that generative AI models’ ability to replicate artistic styles vitiates any fair use defense. This concern has nothing to do with copyright, which does not, and has never, granted monopolies over artistic or musical styles. And the handful of historical accidents in which courts have overlooked this—including the much-reviled *Blurred Lines* copyright case in which the estate of Marvin Gaye successfully pursued claims against artists Robin Thicke and Pharrell Williams—only demonstrate the importance of this fundamental principle to the health of our creative economy. Other laws may provide more appropriate avenues to redress these concerns.

¹ Charles Baudelaire, *The Salon of 1859 II: The Modern Public and Photography*, in ART IN PARIS 1845–1862: SALONS AND OTHER EXHIBITIONS 152–54 (J. Mayne trans., Phaidon 1965); see also Christine Haight Farley, *The Lingering Effects of Copyright’s Response to the Invention of Photography*, 65 U. PITT. L. REV. 385, 417–18 (2004) (“[F]or the majority of artists, their first reaction to the invention of photography was outwardly hostile.”).

² There are countless other examples. Orchestra conductors in the 1930s warned Congress that performing artists had been “rendered helpless and are unable to cope with this vicious and constant repetition” made possible by the advent of recorded performances, which had impacted the demand for live orchestras—only to subsequently embrace recorded music and radio as foundations of the modern music industry. *Revision of Copyright Laws, Hearings Before the House Comm. on Patents*, 74th Cong., 2d Sess. 680 (1936) (statement of Josef Pasternak).

³ U.S. Const. art. 1, § 8, cl. 8; *Golan v. Holder*, 565 U.S. 302, 324 (2012); see also 17 U.S.C. § 107.

That said, it is entirely possible that some “AI” tools may exceed the bounds of fair use. But with the benefit of over 100 years of principle and precedent, our courts are well-equipped to differentiate between fair and infringing uses.

Second, while certain groups are seeking payment for the use of content to train AI models, everyone agrees that it is impossible for AI developers to negotiate and acquire licenses from every rightsholder who owns a copyright interest in the data used to train AI models. So instead, some groups have proposed statutory or collective licensing regimes under which any use of copyright-eligible content to train an AI model would trigger a payment obligation. This would be bad policy, if adopted. Rather than permitting nuanced and case-specific assessments of this new technology, these proposals would eliminate fair use in this area, replacing it with a rigid and inflexible—and most often incorrect—assumption that AI training is infringing.

Moreover, even if it were appropriate to implement a statutory or collective licensing regime, doing it would be far from straightforward. Successfully training an AI model requires using many *billions* of pieces of content, so the scope of any statutory or collective licensing scheme would be many orders of magnitude larger than any similar scheme in the history of American law. Given that scale, any royalty providing meaningful compensation to individual creators could impose an enormous financial burden on AI companies that would either bankrupt them or push all but the largest companies out of the market (or out of the country). Worse still, nearly all the content used to train AI models—including, *e.g.*, anonymous posts on internet forums and review websites—is not only unregistered, but has no identifiable owner. That means the vast, vast majority of royalties would go “unmatched” and therefore unpaid to the original authors. It is worth carefully considering whether it is desirable to adopt a scheme that could cripple AI development—and, by extension, the country’s competitive standing—while providing no benefit to the overwhelming majority of the creators whose works are being used.

In short, existing copyright law is more than up to the task of balancing the need for a dynamic domestic AI industry with the rights of creators.

II. Fair Use

The United States is the epicenter of generative AI technology. Almost all the companies responsible for developing the current and future generations of AI—and many of the researchers who developed the underlying technology—are based here. The reason in large part is that United States copyright law, almost uniquely, provides a broad and flexible fair use doctrine.

Foundational copyright cases establish that the use of copyright-eligible content to create non-infringing works is protected fair use, even if the non-infringing works compete with the originals. This principle applies directly to the mine run of today’s popular AI models, which extract abstract concepts and patterns from billions of pieces of training data and use those concepts to create new content that significantly differs from, and therefore does not infringe, any individual piece of

training data. And while the fair use defense does not necessarily protect all uses of copyright-eligible content to train AI models, the courts are in the best position to delineate fair from infringing uses through the well-worn process of case-specific fair use adjudication.

A. A History of Fair Use and New Technology

Fair use is an “equitable rule of reason” that “permits courts to avoid rigid application of the copyright statute when . . . it would stifle the very creativity which that law is designed to foster.”⁴ As one court explained, the “ultimate test of fair use . . . is whether the copyright law’s goal of promoting the Progress of Science and useful Arts would be better served by allowing the use than by preventing it.”⁵ The doctrine’s roots stretch back to Justice Joseph Story’s 1841 opinion in *Folsom v. Marsh*, which laid out a broad framework that Congress ultimately refashioned and codified as Section 107 of the Copyright Act.⁶ In the 180 years since Justice Story’s opinion, courts have developed the doctrine by applying it to many hundreds of cases.

Over the past few decades, fair use has proved to be an extraordinarily effective and flexible tool for reconciling copyright law with new technology.⁷ In 1984, for example, the Supreme Court relied on the fair use doctrine to shield the manufacturers of video tape recorders against novel claims of secondary infringement.⁸ This facilitated “the growth of a vast new and unforeseen market for the movie studios in the rental and sale of videos for home viewing” that ultimately became “the largest source of revenue for the U.S. movie industry.”⁹ Lower courts have similarly

⁴ *Stewart v. Abend*, 495 U.S. 207, 236 (1990); see also *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 575 (1994) (“From the infancy of copyright protection, some opportunity for fair use of copyrighted materials has been thought necessary to fulfill copyright’s very purpose, to promote the Progress of Science and useful Arts.” (cleaned up)).

⁵ *Castle Rock Ent’tmt, Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 141 (2d Cir. 1998).

⁶ See *Folsom v. Marsh*, 9 F. Cas. 342, 348 (C.C.D Mass. 1841) (Story, J.).

⁷ Indeed, adapting copyright law to new technology is a core function of the fair use doctrine. H.R. Rep. No. 94–1476 at 65–66 (1976) (courts must “adapt the doctrine [of fair use] to particular situations on a case-by-case basis” and given “rapid technological change”); see also Pamela Samuelson, *Unbundling Fair Uses*, 77 *FORDHAM L. REV.* 2537, 2602 (2009) (“One of the important functions of fair use is providing a balancing mechanism within copyright law to allow it to address questions posed by new technologies . . . that the legislature could not or did not contemplate.”); *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 430 (1984) (“From its beginning, the law of copyright has developed in response to significant changes in technology.”).

⁸ *Sony*, 464 U.S. at 454–55.

⁹ Edward Lee, *Technological Fair Use*, 83 *S. CAL. L. REV.* 797, 799 (2010); see also Fred von Lohmann, *Fair Use As Innovation Policy*, 23 *BERKLEY TECH. L. J.* 829, 837 (2008) (“[T]he fair use doctrine has been an unsung hero in the tale of America’s innovation economy, encouraging investment and creating new markets for technology companies.”).

guided the doctrine through various technological developments, using fair use to lay the legal groundwork for revolutionary technologies like internet search,¹⁰ while declining to extend the doctrine to more exploitative technologies, like file sharing and unlicensed media monitoring services.¹¹ And just recently the Supreme Court applied the doctrine to permit the reuse of software APIs, in part to prevent the copyright monopoly from acting as a “lock limiting the future creativity of new programs.”¹² Fair use, in other words, has proven flexible enough to distinguish infringing from non-infringing uses throughout many eras of technological change, consistently balancing the interests of rightsholders with the public’s interest in benefitting from technological innovation.

Many uses of copyright-eligible works to train AI models will fit comfortably within the boundaries of fair use. An unbroken line of cases establishes that the use of a copyrighted work to create a non-infringing final product is quintessential fair use.

The first category of relevant cases involves making complete digital copies of copyrighted works to create non-infringing search tools. *Authors Guild, Inc. v. HathiTrust* concerned an online “repository for [] digital copies” of books scanned from university libraries.¹³ The repository “allow[ed] the general public to search for particular terms across all digital copies” with those searches yielding results in the form of page numbers and the frequency of the search term on each page.¹⁴ After a group of authors sued the repository for making unauthorized copies of their books (*i.e.*, the scanned book copies), the Second Circuit held that the text-search function was “quintessentially transformative [fair] use,” in part because the resulting search results bore “little or no resemblance” to the original scanned works.¹⁵ The year after, the Second Circuit applied that same holding to the “Google Books” service, holding that Google’s “creation of complete digital copies of copyrighted [books]” was fair because that copying was in service of a distinct purpose: “identifying books of interest to the searcher.”¹⁶

¹⁰ *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818–22 (9th Cir. 2003) (use of thumbnail images by search engine is fair use); *Field v. Google Inc.*, 412 F. Supp. 2d 1106, 1117–23 (D. Nev. 2006) (caching websites to enable internet search is fair use).

¹¹ *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1014–19 (9th Cir. 2001) (use of online file-sharing platform not fair use); *Fox News Network, LLC v. Tveyes, Inc.*, 883 F.3d 169, 181 (2d Cir. 2018) (service that “commercially re-distribut[ed]” televised broadcasts “without payment or license” not a fair use).

¹² *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1208 (2021).

¹³ 755 F.3d 87, 90 (2d Cir. 2014).

¹⁴ *Id.* at 91.

¹⁵ *Id.* at 97.

¹⁶ *Authors Guild v. Google, Inc.*, 804 F.3d 202, 217–18 (2d Cir. 2015).

The second category of relevant cases involves the creation of copies of copyrighted works in the course of technological development of competing products. *Sega Enterprises Ltd. v. Accolade, Inc.* concerned a video game company (Accolade) that “reverse engineered” Sega’s game consoles to create and market video games that users could play on their Sega consoles.¹⁷ That process involved copying Sega’s copyrighted computer “code contained in commercially available copies of Sega’s game cartridges” to study the “interface specifications” for Sega’s console.¹⁸ After Sega sued, the Ninth Circuit held that the copying was fair use because it was done for a “legitimate, essentially non-exploitative purpose,” *i.e.*, to “study the functional requirements for [] compatibility” and create games that would legitimately compete with Sega in the marketplace.¹⁹ The Court held that the mere fact that Accolade “copied [Sega’s] code in order to produce a competing product” did not suggest that the use was not fair.²⁰ This holding was adopted by the Federal and Eleventh Circuits (and re-affirmed by the Ninth Circuit) in similar cases about the use of copyrighted material to create non-infringing competitive products.²¹ The Seventh Circuit adopted the same logic in an opinion by Judge Richard Posner about the verbatim copying of a database.²² And the Supreme Court recently endorsed *Sega* in support of the holding that fair use precludes the use of a copyright monopoly to impede competition.²³

These cases—*Hathitrust*, *Authors Guild*, *Sega*, and their progeny—are foundational cases in American copyright law. They are taught in every law school in the country and are embraced by a broad consensus of judges, copyright scholars, and industry stakeholders. Indeed, the Copyright Office relied on these cases in declining to recommend further changes to other parts of the

¹⁷ 977 F.2d 1510, 1514–15 (9th Cir. 1992).

¹⁸ *Id.*

¹⁹ *Id.* at 1522–23.

²⁰ *Id.* at 1522.

²¹ *Sony Computer Entm’t, Inc. v. Connectix Corp.*, 203 F.3d 596, 603–08 (9th Cir. 2000) (creation of intermediate copy of Sony PlayStation BIOS code to create a PC-based PlayStation emulator was fair use, even though defendant’s product would cause Sony to “lose console sales and profits”); *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1539 n.18 (11th Cir. 1996) (“endorsing” the fair use holding in *Sega* in case about reverse-engineering of operating system to compete with copyright owner); *Atari Games Corp. v. Nintendo of Am. Inc.*, 975 F.2d 832, 836–37, 843–44 (Fed. Cir. 1992) (creation of complete copies of computer code to “unlock” a competitor’s video game program was fair use).

²² *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 644–45 (7th Cir. 2003); *see also A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638–40, 645 (4th Cir. 2009) (verbatim copying of copyrighted works to create plagiarism-detection software was fair).

²³ *Oracle*, 141 S. Ct. at 1208 (“An attempt to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression.” (quoting *Sega*, 977 F.2d at 1523–24)); *see also id.* at 1198–99 (citing both *Sega* and *Connectix* with approval).

Copyright Act, which the Office found unnecessary because “intermediate copying for purposes of . . . creation of interoperable software is, in most cases, a fair use.”²⁴

B. Application of Fair Use Principles to AI Training

While these cases unambiguously establish the principle that use of copyright-eligible material to create a non-infringing product is fair, applying this principle to AI training will necessarily be fact specific.

Most of the generative AI models we see today are designed solely to create *new* content. The models are not designed to reproduce copyrightable expression from training data and, in nearly all circumstances, do not do so. Rather, the models derive abstract patterns and relationships—not copyrightable expression—from billions of pieces of training data, and then use those abstract (and uncopyrightable) correlations to create new, non-infringing content.²⁵

For instance, in a typical natural language generative AI model (like those underlying AI chatbots), content in the training corpus is broken up into discrete segments. Then, the AI model examines and extracts statistical relationships among those pieces of content, *e.g.*, their frequency, importance, and semantic relationship to each other. That statistical data is incorporated into the algorithm, and the original content is discarded. By doing this across billions and billions of works, the AI model learns facts about the English language as a whole. To take an extremely simple example, the model may learn that the phrase “Today is a” is more likely to be followed by “sunny day” or “Tuesday” than “grey elephant” or “necktie.”

The result of this process (assuming that this is a natural language model) is the creation of a complex repository of statistical facts about the relationship between words.²⁶ Thus, the AI model is not at all a “collage tool” that stores “compressed copies” of works, as some have alleged.²⁷ Rather, the model derives unprotectable information from the billions of works on which it is trained—much like the tools at issue in *HathiTrust* and *Authors Guild* derived information from

²⁴ U.S. Copyright Office, Report on Software-Enabled Consumer Products at 57 (2016).

²⁵ 17 U.S.C. § 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.”).

²⁶ This process is sometimes called “vectorization” because the goal is to create “word vectors” that “mathematically represent the meaning of a word” so “semantically similar words have similar vectors.” Jayesh Bapu Ahire, *Introduction to Word Vectors*, DZone (Jun. 21, 2022), <https://dzone.com/articles/introduction-to-word-vectors>. For instance “words such as wheel and engine should have similar word vectors to the word car (because of the similarity of their meanings), whereas the word banana should be quite distant.” *Id.*

²⁷ <https://stablediffusionlitigation.com/>.

scanned copies of books. A language model, for example, learns statistical facts *about the English language as a whole*, while AI models in other domains might learn statistical facts about patterns in imagery, music, software code, and so forth. The models then use those statistical facts to generate new output based on an initial input (*e.g.*, a user’s text query). Such a model may also learn unprotectable facts, like the height of the Washington Monument or the year the White House was built. The process, in other words, is designed to generate outputs based on unprotectable facts and abstract concepts, not to recycle or “collage” its training data. To be sure, the output of these models may reflect the same abstract concepts or ideas as works found in the training data, much like two news articles on the same topic might communicate the same unprotectable facts, or like two paintings might depict the same historical landmark. But absent some aberration in the training data or model design (as discussed below), the output will not be a “copy” of (*i.e.*, substantially similar to) any individual work on which the model has been trained.

Under the foundational precedents discussed above, this use of copyrighted works is quintessentially fair—it is a fundamentally a process to extract unprotectable facts about an entire corpus of works, and use those facts to generate original output.

In rare circumstances, quirks in the training data sets used to train AI models can cause outputs that resemble individual pieces of training data.²⁸ Critically, the ability of AI models to duplicate training data is a bug, not a feature. AI researchers see this “overfitting”—*i.e.*, creation of an output that hews too closely to a single piece of training data—as a problem, and are working on a range of methods for avoiding it.²⁹ One simple method is to *increase* the amount of training data on which the model is based.³⁰ (Thus, and ironically, any effort to restrict the availability of training data may have the unintended effect of increasing the frequency of AI models simply copying pre-existing expression.) In any case, under current doctrine, training a model that predominantly creates non-infringing outputs easily qualifies for fair use protection, whether or not the model *can* be used to infringe copyright in rare instances of overfitting (and whether or not those rare individual acts of reproduction would qualify as actionable infringement themselves).³¹

²⁸ See, *e.g.*, Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, at 4–5 (2023), <https://arxiv.org/abs/2301.13188> (testing one popular image generation model and determining that out of 175 million images the researchers generated, only 109 images were duplicates or near-duplicates of the training data).

²⁹ *Id.* at 14; Henderson et al., *Foundation Models and Fair Use*, at 20–25 (Mar. 29, 2023), <https://arxiv.org/pdf/2303.15715.pdf>.

³⁰ Jared Kaplan et al., *Scaling Laws for Neural Language Models*, at 3 (2020), <https://arxiv.org/abs/2001.08361> (“the size of the dataset” determines overall “[m]odel performance”).

³¹ *Sony*, 464 U.S. at 442 (“[T]he sale of copying equipment, like the sale of other articles of commerce, does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses.”); see also *iParadigms*, 562

This, however, does not mean that *all* AI models will invariably be able to successfully avail themselves of the fair use doctrine. There are countless ways to design and train an AI model, and not all of them will necessarily stay within the bounds of fair use. For instance, creators of generative models trained on a small amount of data that, as a result, might more reliably replicate that data may find it harder to establish that their uses are fair. But the key point is that our courts are best equipped to make this determination, by marshalling and evaluating case-specific evidence, as they have in many hundreds of fair use cases over the past 180 years.

C. AI's Use of Artistic Styles is Not a Copyright Issue

Several critics of the AI industry have focused on the fact that some generative AI models are able to derive an “artistic style” from training data and replicate that style, on demand, in outputs. I can understand that concern—but this is not a copyright problem. Copyright does not, and has never, permitted an artist to preclude another from mimicking an “artistic style.”³² For that reason, generative AI’s ability to replicate individual artistic styles should play no role in the fair use analysis, nor should it motivate Congress to alter the copyright framework.

Courts have repeatedly explained the importance of “styles” remaining free for all to use. In *Williams v. 3DExport*, for example, the court considered copyright claims brought by an artist who “contend[ed] that he invented anime, one of the world’s most popular styles of animation.”³³ The artist traced the style to a “master dissertation” he wrote in which he described an “art style” featuring “characters with ‘round features,’ ‘big eyes,’ and a ‘particular nature of [] hair,’” and pointed to a graphic novel he wrote (and registered with the Copyright Office) as an example of that style.³⁴ In the suit, the artist objected that his style had been “illegally [taken]” and used in

F.3d at 639–40 (explaining “[t]he question of whether a use is transformative does not rise or fall on whether the use perfectly achieves its intended purpose,” finding use fair even though system “is not fool-proof”).

³² *Jewelry 10, Inc. v. Elegance Trading Co.*, No. 88-cv-1320, 1991 WL 144151, at *4 (S.D.N.Y. July 20, 1991) (“[A] painter who develops a style or technique, such as the rendition of perspective, impressionism, pointillism, fauve coloring, cubism, abstraction, psychedelic colors, minimalism, etc., cannot prevent others from adopting those ideas in their work.”); see also *Douglas v. Osteen*, 317 F. App’x 97, 99 (3d Cir. 2009) (“Furthermore, the use of a particular writing style or literary method is not protected by the Copyright Act.”); *Whitehead v. CBS/Viacom, Inc.*, 315 F. Supp. 2d 1, 11 (D.D.C. 2004) (“[S]tyle alone cannot support a copyright claim. While similar writing styles may contribute to similarity between works’ total concept and feel, a particular writing style or method of expression standing alone is not protected by the Copyright Act.”); see also 2 PATRY ON COPYRIGHT § 4:14 (“Particular elements may colloquially be regarded as typical of an author or artist’s individual style, but it is only their fixation in a particular work in a particular expression that is eligible for [copyright] protection.”).

³³ No. 19-cv-12240, 2020 WL 532418, at *1 (E.D. Mich. Feb. 3, 2020).

³⁴ *Id.*

new artwork that he found “disturbing,” and sued thirteen defendants.³⁵ The court did not hesitate to dismiss the claim, holding:

... even if [plaintiff] was the first to think up the anime, he could only have a protectable copyright interest in his specific expression of that idea; he could not lay claim to all anime that ever was or will be produced.³⁶

Many other cases reach the same result.³⁷ And for good reason: extending copyright protection to styles not only violates the express command of Section 102(b) of the Copyright Act, but (as commentators have explained) would lead to an unworkable copyright system:

If an author or artist claimed broad protection for a style not associated with a particular work and fixation, it would be difficult, if not impossible, to determine the scope of protection. . . . An endless series of derivative works could be created, yet none might be treated as a derivative work since it is the style and not any particular work for which protection would be asserted. Determining substantial similarity between plaintiff's and defendant's “works” would be skewed since plaintiff would not be asserting copyright in a work, but rather in an amorphous style that exists independent of any particular work.³⁸

The much-reviled *Blurred Lines* copyright verdict is a perfect example of the chaos that results from extension of copyright to protect styles. In 2015, a jury found that Robin Thicke’s song “Blurred Lines” infringed Marvin Gaye’s 1977 hit song “Got To Give It Up” based on alleged similarities between the two works.³⁹ That result has been widely criticized by a broad array of industry stakeholders, creatives, and commentators, who have pointed out that “the similarities between the songs are not within the melody, lyrics, or harmony, but rather in the overall sound, groove, and vibe,”⁴⁰ and argued that the verdict violates “the principle that ideas cannot be copyrighted, a notion that is essential to free speech and artistic expression.”⁴¹ It has thrown the

³⁵ *Id.*

³⁶ *Id.* at *3.

³⁷ See 2 PATRY ON COPYRIGHT § 4:14 (collecting cases).

³⁸ *Id.*

³⁹ See generally *Williams v. Gaye*, 895 F.3d 1106 (9th Cir. 2018) (upholding jury verdict on procedural grounds).

⁴⁰ Olivia Lattanza, *The Blurred Protection for the Feel or Groove of a Song Under Copyright Law: Examining the Implications of Williams v. Gaye on Creativity in Music*, 35 *TUORO L. REV.* 723, 725 (2019).

⁴¹ Tim Wu, *Why the “Blurred Lines” Copyright Verdict Should Be Thrown Out*, *New Yorker* (Mar. 12, 2015), <https://www.newyorker.com/culture/culture-desk/why-the-blurred-lines-copyright-verdict-should-be-thrown-out>; see also The Editors of *GQ*, *Rick Rubin on Pharrell’s “Blurred Lines” Lawsuit*, *GQ* (Nov. 4, 2019), <https://www.gq.com/story/pharrell-and-rick-rubin-blurred-lines-copyright-lawsuit> (quoting famed producer Rick Rubin as saying the verdict is “bad for music” because “now, based on that one case,

music industry into mild disarray,⁴² and spurred many follow-on lawsuits against, for example, Led Zeppelin and Taylor Swift.⁴³ Thankfully, many of these follow-on lawsuits have failed,⁴⁴ including one recently brought by the estate of a co-writer of a Marvin Gaye song against the artist Ed Sheeran, suggesting that history will view the “Blurred Lines” case as a misguided historical accident.

Notably, prominent rightsholders were among the chorus of voices who denounced the *Blurred Lines* decisions and endorsed the bedrock principle that styles are not protected by copyright. As the Recording Industry Association of America and National Music Publishers’ Association warned the Ninth Circuit in an amicus brief urging vacatur of the verdict, “[n]ew generations of musicians would be in constant peril of copyright lawsuits because they had used someone’s musical style” if the decision’s infringement-by-style rationale were upheld.⁴⁵

Style appropriation, in other words, is not a copyright concern, and it is critical that styles remain freely accessible for public use. As one of this nation’s leading jurists has explained, “Intellectual (and artistic) progress is possible only if each author builds on the work of others. No one invents even a tiny fraction of the ideas that make up our cultural heritage . . . Every work uses scraps of thought from thousands of predecessors, far too many to compensate even if the legal system were frictionless, which it isn’t.”⁴⁶ Those who suggest that AI models’ ability to replicate individual artistic styles should somehow affect the fair use analysis, or motivate Congress to alter the copyright framework, ignore this fundamental principle. To be sure, generative AI can be used in ways that might implicate other legal frameworks, like unfair competition, trademark, and state rights of publicity and privacy. But the general principle that styles are free for all to use is a

there’s a question of what a song is . . . [a] feeling is not something you can copyright”).

⁴² Jason Palmer, “*Blurred Lines*” Means Changing Focus, 18 VAND. J. OF ENT. & TECH. L. 907, 926–28 (2016) (summarizing the “industry problems” that arose after the *Blurred Lines* verdict, including the payment of “precautionary royalties”).

⁴³ Ben Sisario, *Ed Sheeran Wins Copyright Case Over Marvin Gaye’s “Let’s Get It On,”* N.Y. Times (May 4, 2023), <https://www.nytimes.com/2023/05/04/arts/music/ed-sheeran-marvin-gaye-copyright-trial-verdict.html> (commenting on the connection between the *Blurred Lines* case and subsequent similar cases).

⁴⁴ See, e.g., *id.*

⁴⁵ Brief Amici Curiae of the RIAA and NMPA in *Williams, et al. v. Gaye*, No. 15-56880 (Dkt. 100-2 at 28) (9th Cir., filed Apr. 2018); see also Brief Amici Curiae of the RIAA and NMPA in *Skidmore v. Led Zeppelin*, No. 16-56057 (Dkt. 77-2 at 8) (9th Cir., filed Nov. 5, 2018) (urging *en banc* reconsideration of panel opinion vacating jury verdict for defendant, arguing that composers “need copyright law to let them create new music incorporating ideas from the vast cultural library of past musical works”).

⁴⁶ *Nash v. CBS, Inc.*, 899 F.2d 1537, 1540 (7th Cir. 1990) (Easterbrook, J.).

foundational aspect of federal copyright policy that is essential to our creative economy and the general freedoms of speech and expression.⁴⁷

III. The Feasibility of Licensing AI Training Data

All stakeholders—even the most vocal critics of today’s generative AI tools—appear to agree that the technology is “capable of revolutionizing the creative process” and “enhanc[ing] artistic expression.”⁴⁸ Everyone, in other words, has an interest in making sure that this technology succeeds and does so in a way that enhances, rather than undercuts, our ability to compete on the world stage.

But it is also well understood that mandating bilateral negotiations between AI developers and individual rightsowners will be counterproductive. It will be impossible for legitimate AI developers to negotiate with each and every rightsholder who owns a copyright interest in one of the billions of individual pieces of data the developers’ models require. And attempts to build AI models using smaller sets of licensed or public domain material will lead to models that are less effective and, ironically, more likely to (inadvertently) create outputs that simply regurgitate their training data. Moreover, the economic dynamics of any bilateral licensing negotiation will be impossibly skewed: the supply of potential training data is effectively unlimited, meaning that no individual rightsholder will be able to demand more than nominal compensation for the use of its works.

Recognizing these problems, certain rightsholders and industry groups have instead suggested a statutory or collective licensing regime.⁴⁹ I believe these proposals are misguided. Imposing a payment requirement for AI training data would cut against the deeply rooted copyright principles discussed above—it would require payment for a process the result of which is the extraction of unprotectable facts about an entire corpus of potentially copyrighted works. And besides being bad policy, such a scheme could raise serious constitutional concerns.⁵⁰ And if the concern is that

⁴⁷ *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003) (copyright’s denial of protection to “ideas” is a “First Amendment accommodation” “built-in” to the Copyright Act).

⁴⁸ Press Release, Authors Guild, Creator Groups Meet Lawmakers on AI Issues (Apr. 4, 2023), <https://authorsguild.org/news/ag-and-creator-groups-meet-lawmakers-on-ai-issues/>.

⁴⁹ *See, e.g., id.*

⁵⁰ The Supreme Court has stated that the existence of the fair use defense is critical to ensuring that the copyright law does not conflict with the First Amendment. *Eldred*, 537 U.S. at 218–21 (suggesting that any attempt by Congress to “alter[] the traditional contours of copyright protection” could trigger “First Amendment scrutiny,” and upholding challenged law against First Amendment challenge due in part to presence of fair use defense). Because a statutory licensing scheme would require payment for uses of copyright-eligible material that the fair use defense would otherwise permit without payment or permission, such a scheme may be subject to a constitutional challenge.

some AI models will not stay within the bounds of established fair use principles, the judiciary’s careful and time-tested fair use jurisprudence permits deeper, case-specific consideration, and compensation to creators in appropriate cases.⁵¹ That is far preferable to an across-the-board legislative determination that all AI training is infringing.

Even setting those policy objections aside, any effort to implement a statutory or collective licensing regime would need to grapple with a series of (perhaps insurmountable) practical problems.

The first and most obvious problem with a statutory licensing regime for AI training is scale. AI models can train on any media including, for example, long-form text (*e.g.*, blog posts), short-form text (*e.g.*, Tweets or forum comments), or images (full size or thumbnail). Often, AI developers train their models by pulling this content from the internet.⁵² Because virtually all of this content is eligible for copyright protection—regardless of its perceived aesthetic or expressive value⁵³—any statutory licensing scheme would need to include virtually the entire internet within its scope. As a result, any licensing regime that triggered a payment obligation to any rightsholder whose work is used to train an AI model would require the administration of that license (and the payment of royalties) for *billions upon billions* of works—many of which are published online with no ownership information.

As precedent, some have pointed to the blanket licensing regime for musical works established by 2018 Music Modernization Act (“MMA”).⁵⁴ At a high level, that regime created a new mechanism by which “digital music providers” could obtain a “blanket mechanical license” through a designated entity known as the “mechanical licensing collective” or “MLC.”⁵⁵ The MLC administers the blanket license on behalf of musical work copyright owners by, for example, collecting royalties from digital music providers, seeking to identify the owners of the musical works in each song for which royalties are paid (a process known as “matching”), and distributing collected royalties to the appropriate owners.⁵⁶ Congress required the MLC to hold any

⁵¹ A related concern is that writing hard-coded legislation in the context of evolving technology is always extremely difficult and carries a substantial risk that subsequent technological developments will render the statutory framework obsolete. *See, e.g.*, Aaron L. Melville, *Note, The Future of the Audio Home Recording Act of 1992: Has It Survived The Millennium Bug?*, 7 B.U. J. SCI. & TECH. L. 372, 383–84 (2001) (addressing the Audio Home Recording Act).

⁵² One source for AI training data, for instance, is the repository of web crawl data maintained by Common Crawl—essentially, a copy of the entire world wide web. *See* <https://commoncrawl.org/>.

⁵³ *Bleistein v. Donaldson Lithographing Co.*, 188 U.S. 239, 251–52 (1903).

⁵⁴ *See* Pub. L. 115–264 at Sec. 102(a), 132 Stat. 3677, *codified at* 17 U.S.C. § 115(d).

⁵⁵ 17 U.S.C. § 115(d)(1)(A).

⁵⁶ *Id.* § 115(d)(3)(C).

“unmatched” royalties for a certain period of time, and then distribute those funds to “copyright owners identified in the records of the collective” based on their “relative market share.”⁵⁷

A statutory licensing regime for AI training would be massively more complicated than the one created by the MMA. Most obviously, the scope of a statutory license covering AI training activities would be orders of magnitude larger than any statutory license ever implemented. The MMA’s blanket license, for instance, covers a universe of musical works numbering in the tens of millions.⁵⁸ By contrast, because AI models use a vast array of media for training, a statutory licensing scheme for AI training would need to cover, at least, every publicly accessible work on the internet—including every forum comment, online review, social media post, and business website. That universe of works likely numbers in the billions or tens of billions.

In addition, the works covered by this statutory license differ substantially from the works covered by the MMA’s blanket license. While it is not easy to identify the musical work owner for each sound recording, the endeavor is relatively straightforward if the MLC has access to the relevant data sources, which are in large part available to the MLC’s matching team.⁵⁹ Acquiring sufficient data to identify the copyright owner of the billions of images or pieces of text on the internet would be much harder, if not impossible. Many photographs, social media posts, business websites, blog posts, or online reviews on the internet are published anonymously or pseudonymously, or simply with no information to identify the content’s author. I cannot estimate the breadth of the issue, but it seems likely that, for nearly all of the content that AI models could use for training purposes, identifying the rightful copyright author or owner will be an unattainable goal.

The U.S. Copyright Office has touched on this issue in its multiple studies of the issue of so-called “orphan works.” As the Office explains, “it can be time-consuming, difficult or even impossible to locate the copyright owner” for a given work.⁶⁰ That is particularly true when the work is

⁵⁷ *Id.* § 115(d)(3)(H), (J).

⁵⁸ As of the end of 2021, the MLC’s musical works database contained data for 23.8 million musical works. Mechanical Licensing Collective, 2021 Annual Report, Appendix at 4.

⁵⁹ Mechanical Licensing Collective, 2021 Annual Report, Appendix at 20 (“The MLC’s Matching Team utilizes over 30 public databases and other research sources to support its matching efforts. These sources include various collective rights management organization databases, music credit databases, [digital music provider] websites, content owner websites, and other niche and genre specific sources.”).

⁶⁰ *See* U.S. Copyright Office, Report on Copyright and Digital Distance Education at 41–43 (1999); *see also* U.S. Copyright Office, Report on Orphan Works at 92 (2006) (“portion of works for which owners [cannot be] located can be significant”); *id.* at 22 (citing a study by Carnegie Mellon University Libraries finding that “for the books in the study, 22% of the publishers could not be found”); U.S. Copyright Office, Report on Orphan Works and Mass Digitization at 2 (2015) (problem of orphan works is “widespread and significant”).

published without “information about the author or the owner of copyright in the work.”⁶¹ But even if the work is published with such information, subsequent “[c]hanges of [o]wnership”—either due to transfers of ownership or the copyright owner’s death, relocation, or corporate dissolution—can render identification of the copyright owner impossible.⁶² These problems are compounded by the lack of reliable “information resources” about ownership; as a result, many ownership searches ultimately reach “dead ends.”⁶³

These problems would be even more severe in the context of works that users post to the internet, often anonymously or pseudonymously. It is almost impossible to imagine how the administrator of a statutory license for AI training could identify the rightful owner of more than a tiny fraction of the works used to train AI models. Any attempt to tackle this problem would require armies of human researchers who (like the members of the MLC’s matching team) would have to conduct individualized, work-by-work investigations into a massive and constantly-growing corpus of works covered by the licensing scheme.

As a result, any statutory licensing scheme would lead to massive amounts of “unmatched” royalties that would sit idle in the coffers of the license’s administrator. The MLC, which administers a far narrower statutory license covering a universe of works whose owners are far easier to identify, had over \$150 million in unmatched royalties in 2021 alone, and there is no indication that any amount of money or resources could eliminate that problem.⁶⁴ The amount of “unmatched” royalties under a statutory license for AI training data would no doubt be far larger. Nearly all rightsholders would receive no remuneration whatsoever. It is worth evaluating whether imposing a significant financial burden on the AI industry for such uncertain and attenuated benefits to creators is a desirable result.

A second, related challenge is that any statutory or collective licensing scheme—no matter how carefully designed—would find itself caught between two difficult-to-reconcile policy objectives: (1) to provide meaningful compensation to individual artists and rightsholders, and (2) to ensure that AI companies can continue to thrive in the United States.

A statutory or collective licensing scheme would presumably require any AI developer to pay some fixed or floating rate to compensate the copyright owners for the use of each piece of training data. And because the goal of the proposed collective licensing scheme would be to protect individual “human creators and artists,”⁶⁵ the rate paid for the use of any individual works would have to be

⁶¹ U.S. Copyright Office, Report on Orphan Works at 23–24.

⁶² *Id.* at 26–29.

⁶³ *Id.* at 29–34.

⁶⁴ Mechanical Licensing Collective, Annual Report 2021, Appendix at 15.

⁶⁵ Press Release, Authors Guild, *Creator Groups Meet Lawmakers on AI Issues* (Apr. 4, 2023),

financially significant. A licensing scheme that led to individual creators receiving monthly royalty checks of a few cents from the AI developers who used their works for training purposes would do nothing to protect “human creators” or the “[f]uture of journalism, literature, and the arts.”⁶⁶

But AI models require many *billions* of pieces of training data to be effective. As a result, it will be extremely challenging to set a royalty rate that provides meaningful compensation to individual copyright owners without imposing a crippling financial burden on AI developers, who would have to make many billions of rate payments for the works they use. If the royalty rate were set too high, it would either bankrupt the United States AI industry—eliminating our ability to compete on the international stage—or push all but the largest companies out of the market (or out of the country). It would, in other words, be extraordinarily challenging to set a royalty rate that would both compensate individual creators and encourage the growth and development of a domestic AI industry. Developers who are unable to afford the cost of AI development in the United States would surely move their efforts to other countries with more permissible copyright frameworks.⁶⁷

* * * * *

In closing, I applaud the Subcommittee for its quick engagement on the important and challenging issues surrounding the rapid growth of artificial intelligence technologies. The Subcommittee, and Congress more broadly, has a chance to ensure that the United States continues to lead the world in responsible development of artificial intelligence technologies. While Congress should continue to think carefully about the copyright implications of those technologies, it should also feel confident that the laws that it already passed are well-suited to address whatever issues may arise. Thank you for the opportunity to provide testimony for this hearing, and I look forward to answering your questions.

<https://authorsguild.org/news/ag-and-creator-groups-meet-lawmakers-on-ai-issues/>.

⁶⁶ *Id.*

⁶⁷ See, e.g., Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 BERKELEY TECH. L. J. 1407, 1431 (2021) (noting that Japan “was the first country in the world to update its copyright laws” to “demonstrate a national commitment to the flourishing of AI industries,” and explaining the country’s 2018 Copyright Act implementing a broad exemption for incidental copies and machine learning); Mark A Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 769 n.147 (2021) (reporting that Singapore is considering an AI exemption much like Japan’s).

Mr. ISSA. The gentleman yields back early.
Mr. Burch, Professor Burch.

STATEMENT OF CHRIS CALLISON-BURCH

Mr. CALLISON-BURCH. Chair Issa, Ranking Member Johnson, distinguished Members of the Subcommittee, thank you for the chance to testify on this important topic.

My name is Chris Callison-Burch. I'm a professor at the University of Pennsylvania, a visiting researcher at the the Allen Institute for Artificial Intelligence, and the Deputy Chair of the Board of Advisors of the Johns Hopkins University Human Language Technology Center of Excellence.

Generative AI had its breakthrough moment in November of last year with the release of OpenAI's ChatGPT. This brought my field of research into the public eye and generated a huge amount of enthusiasm. I had access to OpenAI's large language models about a 1½ years before the public. Despite having worked in this field for over 20 years, I was shocked by its capabilities. My first encounter with it pitched me into a career existential crisis.

The technology had seemingly solved many of the problems that I was researching, that it could translate texts from Russian into English, it could write coherent summaries of long documents, and then answer questions about them. I wondered whether there was any room left for academic research in light of the fact that these large language models required google-sized data centers to train.

So, I asked myself should I just drop out of computer science and become a poet? Of course, the next week I downloaded 15,000 poems from the Internet and trained the system to write much better poetry than I ever could. I've subsequently calmed down and I do not think that my job is at imminent risk of being replaced by ChatGPT, but I understand that many other people are experiencing the same sense of panic that I had.

Artists and writers are worried about their work being devalued. I worry that careers like a paralegal might go the way of a lamplighter. I think that at its core what we're talking about today goes far beyond copyright. It's about the value of work.

This is a truly transformative technology that will shape many aspects of our lives. I hope that it is for the better. I optimistically believe that AI will enable us to be more productive workers and to allow more people to realize their creative visions.

In my testimony today I hope to offer my expertise in the technical aspect of generative AI, and I promise to explain it in a way that doesn't require a Ph.D. in computer science, answers to any questions that you have about the potential for legislation impacting on innovation in this field, and advocacy for retaining fair use for the purposes of training generative AI systems.

In my written testimony I've provided an overview of how these systems work. I'm happy to explain during the hearing today how they do or to have a one-on-one meetings with you or your staff at a later date.

To briefly summarize the points that I want to highlight from my written testimony, generative AI is trained on huge amounts of data. Large language models are now trained on roughly one trillion words. Image generators are trained on hundreds of millions

of images. Much, or even most of that data consists of copyrighted works that have been gathered by automatically crawling the web.

It's important to remember that from these copyrighted works AI systems learn. This learning process is called pretraining, which is the "P" in GPT. Pretraining AI systems is different than how we teach our children to learn, but the effect is similar. AI systems learn how to use language. They learn facts about the world. They learn ideas and opinions. They learn visual concepts. They even learn some rudimentary common-sense reasoning skills.

This pretraining happens on copyrighted data which is then set aside as models are fine-tuned to perform more specific tasks. For instance, a large language model can be fine-tuned on a much smaller purpose-built set of data to become an intelligent tutor or a computer vision system can be fine-tuned to detect cancerous growths in mammograms. These systems could not be as easily adapted to these specialized tasks without the general knowledge that they acquire from the copyrighted data that they're pretrained on.

I believe like Sy that pretraining these systems squarely falls within fair use and that Internet era court precedents likely established this as the case, although as the Ranking Member mentioned, this is currently being litigated in the courts.

I do believe that the output of generative AI systems can infringe on copyright and it's worth Congress considering legislation to better shape copyright to govern things like copyrightable characters and possibly to extend copyright to cover things like right-of-publicity.

I look forward to discussing this topic with you today.

[The prepared statement of Mr. Callison-Burch follows:]

Understanding Generative Artificial Intelligence and Its Relationship to Copyright

Written Testimony of

Christopher Callison-Burch, Ph.D.

Associate Professor of Computer and Information Science

School of Engineering and Applied Sciences

University of Pennsylvania

and

Visiting Researcher, Allen Institute for Artificial Intelligence (AI2)

and

Deputy Chair of the Advisory Board of the Human Language Technology Center of Excellence

(HLTCOE)

Johns Hopkins University

Before

The U.S. House of Representatives Judiciary Committee

Subcommittee on Courts, Intellectual Property, and the Internet

Hearing on

Artificial Intelligence and Intellectual Property: Part I – Interoperability of AI and Copyright Law

Chairman Issa, Ranking Member Johnson, and distinguished Members of the Subcommittee, thank you for the opportunity to testify on the topic of artificial intelligence (AI) and intellectual property. My name is Christopher Callison-Burch, and I am an Associate Professor of Computer and Information Science at the University of Pennsylvania. I have been working in the field of AI for over 20 years, with more than 100 publications in the field, which have been cited over 20,000 times.

Currently, I am on sabbatical from the University of Pennsylvania and am a visiting researcher at the Allen Institute for Artificial Intelligence (AI2). AI2 is a non-profit research institute founded in 2014 with the mission of conducting high-impact AI research and engineering in service of the common good. AI2 was created by the late Paul G. Allen, philanthropist and Microsoft co-founder.

Additionally, I serve as the Deputy Chair of the Advisory Board of the Johns Hopkins University Human Language Technology Center of Excellence (HLTCOE), which aims to bridge academia and government in order to help the government better use innovations in the field of artificial intelligence. Please note that the opinions expressed in my testimony are my own and do not represent the views of the University of Pennsylvania, AI2, or the JHU HLTCOE.

Background

Generative AI had a breakthrough moment last November with the release of OpenAI's ChatGPT, bringing my field of research into the public eye and generating significant excitement. I had early access to OpenAI's private Beta in June 2021 and was a visiting researcher at Google in 2019 and 2020, where I used their LaMDA language model before its public announcement. Despite my two decades of experience in the field, I was similarly astonished by the capabilities of these large language models (LLMs), and I believe that this is a truly transformative technology. It is crucial that we carefully consider and draft legislation that simultaneously encourages innovation and guards against potential negative impacts on society.

In my testimony today, I hope to offer the following to this Subcommittee:

1. My expertise in explaining the technical aspects of generative AI in a way that is understandable without requiring a background in computer science.
2. Answers to any questions you may have about the emerging capabilities of this technology and its potential to accelerate innovation, as well as its potentially negative impacts on society.
3. Advocacy to retain fair use for the purposes of *training* AI systems, which I hope the Subcommittee will consider when drafting legislation.

I composed parts of my testimony using ChatGPT (using GPT-4, accessed May 13-15, 2023).

Explanation of Generative AI

Generative AI is a subfield of artificial intelligence that focuses on creating new content, such as text, images, or music, based on patterns learned from existing data. Some examples of generative AI outputs include text generated by LLMs like ChatGPT and images created by text-to-image systems like Midjourney.

How do LLMs generate text?

I find that it's often best to understand generative AI by looking at examples of its output. Here is an example from OpenAI's GPT 3.5 system. The part in white is an input "prompt", and the colored section is the model's output (also called the "completion" of the prompt).

My favorite professor at the University of Pennsylvania is Chris Callison-Burch. He is a professor in the Computer and Information Science Department, and is the director of the Natural Language Processing Group at the Institute for Research in Cognitive Science. He is an incredibly accomplished professor and researcher, whose work has been widely published in top journals and conferences in the field of Natural Language Processing. He is also well-known for his teaching and research on machine translation, dialogue systems, and other areas of natural language processing. He is an inspiring professor who encourages his students to think critically and creatively, and provides a stimulating environment to learn and grow as a student.

Submit



129

In a similar fashion to how autocomplete works on your smartphone, a large language model is trying to predict the next word. Since I have a unique, two-part last name, GPT correctly predicts that the word following “Chris Callison-” should be “Burch”. Unlike your phone, LLMs like GPT are capable of not just generating a single word, but they continue to generate complete sentences and paragraphs.

LLMs generate longer texts word by word in a process. In order to generate each word, the LLMs assign a probability to each word in the vocabulary of English according to the preceding context. The context is initially the prompt that I gave to the system, and then it expands by adding each word that the LLM has generated up to that point in time. Rather than taking the most likely word, the model randomly “samples” the next word according to its likelihood. The colors of the words in the figure give an indication of how likely the model thinks that it was given the preceding context. The word “accomplished” is colored red because it was considered to be a low probability continuation by the model, with a likelihood of less than 1% given the preceding context. Here is an indication of the other higher probability words that it could have picked instead of outputting “accomplished”:

knowledgeable = 47.95%
engaging = 19.51%
passionate = 10.00%
inspiring = 5.67%
talented = 2.67%
accomplished = 0.47%
Total: -5.35 logprob on 1 tokens (86.28% probability covered in top 6 logits)

After outputting each word, the model appends the prompt and what has been generated so far, and then uses that as the full context in order to generate the next word. The generation continues until a word count limit has been reached, or until the model outputs a special “end of generation” token, which can be thought of as signifying the end of a complete thought or a good stopping point for a passage of text.

Is generated text factual?

At its core, the process of generation is based on word associations that have been learned from data. It is important to understand that there is no guarantee of factual accuracy in models’ outputs. For example, in the output of the model about me, there are several correct facts, several incorrect statements, and several subjective statements.

Correct:

1. I am a professor in the Computer and Information Science Department
2. My work has been published in top journals and conferences in Natural Language Processing
3. I am known for research on machine translation and other areas of NLP

Incorrect:

4. I am not the director of the NLP Group at the Institute for Research in Cognitive Science. IRCS was an institute at my university, but it shut down several years ago.
5. I am not known for dialogue systems.

The example text makes many subjective statements like I am an “inspiring professor”. These are due to how I prompted the model by getting it to write a positive student review by saying “My favorite professor”. The model has learned to associate “favorite professor” with ideas like that they “encourage students to think critically and creatively” and that they “provide a stimulating learning environment” and that they help students “learn and grow”.

Similarly, it has learned to associate NLP professors with “leader of the NLP group” and the “Institute for Research in Cognitive Science” with the “University of Pennsylvania”. From these associations it produces a factually incorrect sentence that sounds plausible - this is called a “hallucination”. Because text that LLMs generate has no guarantees of factual accuracy, that limits on how it should be used. For this reason, a much better use case for an LLM is creative writing rather than producing newspaper articles. Current research is investigating whether factuality can be improved by allowing the model to first search the web, and use retrieved documents in its context.

Additionally, because LLMs generate text through word associations, they are currently poor at certain kinds of tasks that require skills like mathematical reasoning. You can prompt an LLM with the question “2+2=” and it will correctly generate “4” because it has learned that association, but if you prompt it to add two arbitrarily large numbers it will generate an incorrect output at random, since it will have never learned to associate them. Current research is focused on allowing LLMs to use external tools like calculators or Python code in order to help with symbolic reasoning.

How do generative AI systems learn?

Generative AI systems learn about the world through examples. A collection of examples is often called the system’s training data or its “training corpus” (plural: “corpora”). The training data for LLMs consists of large numbers of text documents. For text-to-image generation systems, training data consists of large numbers of images paired with text captions that describe their contents. For music generation systems, training data might consist of music files sometimes in particular formats like MIDI.

GPT stands for “Generative Pre-trained Transformer”. We’ve seen what “generative” means. Let’s take a look at what “pre-trained” and “transformer” mean.

The process of learning from examples is called “pre-training”, where systems are trained to perform general tasks. For LLMs, pre-training tasks can be “fill in the blank” or “sentence completion” where it is shown a training document where a word or a sequence of words is masked out, and it must learn to correctly predict what words occurred there. The original document gives the “ground truth” correct answer.

In addition to fill-in-blank, other pre-training tasks include “next sentence prediction” where the model is given one sentence, and must learn to select the actual next sentence that follows, given a multiple choice list. The training process for image generation systems involves adding noise to the training images, and having the model try to remove it in order to restore the true image, and other tasks like learning to associate words in the captions of images with the corresponding objects in them.

During pre-training, generative AI systems update their “model parameters”. Model parameters are tables of numbers that AI systems use in their neural networks in order to generate new

outputs. They consist of mathematical representations from linear algebra that form the basis for their neural networks to make predictions. During training, each time the AI system makes an error in one of its training tasks, its model parameters are updated in order to make its prediction more correct. This causes AI systems to learn patterns in the training data. The kind of neural network that many AI systems now use is called the “transformer” model. The transformer model was designed to make the training process scalable.

Following pre-training, which teaches to model general capabilities related to language or vision, the models can be adapted to perform specific tasks through a subsequent training step called “fine-tuning”. This training step uses the model parameters learned during pre-training as a starting point, and then updates them so that the model can perform specific tasks. For example, it might be a text classification task like predicting the sentiment expressed in a document, or a computer vision task like predicting whether the image of a mammogram contains cancerous growth. Or it might be a generation task like adapting an LLM so that it is a better interactive tutor, or a text-to-image generation system so that it produces more painterly results. Fine-tuning requires a distinct set of data, which is typically much smaller than the data used in pre-training, and is often purpose-built for the task.

This process of pre-training and then fine-tuning has been widely established to produce state of the art results in a variety of machine learning tasks.

How much data do AI systems use for pre-training?

Text Datasets

AI systems use very large amounts of data during pre-training. An early transformer model developed by Google called BERT (Bidirectional Encoder Representations from Transformers)¹ was trained on a large corpus of text data. Specifically, it was pre-trained on two datasets:

1. The English Wikipedia: This contains approximately 2,500 million words.
2. BookCorpus: This is a dataset containing approximately 800 million words.

In total, BERT was trained on approximately 3.3 billion words.

Subsequent LLMs were trained on increasing large text corpora, many of which are derived from data collected by Common Crawl. Common Crawl is a non-profit organization that crawls the web and freely provides its archives and datasets to the public.² Common Crawl's web archive consists of petabytes of data collected since 2011. It conducts monthly crawls of the web to download and archive web pages. Its archive includes HTML, metadata, and related information.

¹ Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

² <https://commoncrawl.org/about/>

Google's T5 model (the text-to-text transfer transformer)³ created a pre-training data set called C4 (Colossal Clean Crawled Corpus) derived from Common Crawl's data, which contains over 150 billion words.⁴ Google's PaLM model was trained on 780 billion words.⁵ Google's latest LLM, PaLM 2, was pre-trained on an undisclosed amount of data that is "significantly larger than the corpus used to train PaLM" and which consists of a diverse set of sources: web documents, books, code, mathematics, and conversational data that have been carefully preprocessed to remove sensitive personally identifiable information.⁶ OpenAI no longer discloses the exact training data used in its models.⁷ GPT-4 is likely pre-trained on approximately 1 trillion words.

The Washington Post recently published an interactive feature (in collaboration with my fellow researchers at AI2) that allows you to search a list of websites that are contained in the C4 dataset, and see what fraction of the pre-training data a website represents.⁸ For example, the largest .gov domain in the C4 dataset is govinfo.gov which consists of 2.1 million words, and represents 0.001% of the dataset.

Image Datasets

Datasets used to pre-train image generation systems include a 400 million image dataset called LAION-400m, a 10 terabyte dataset with 256×256 pixel images, captions and metadata derived from the web.⁹ LAION stands for Large-scale Artificial Intelligence Open Network. It is a non-profit organization making machine learning resources available to the general public. You can search the LAION dataset with a provided web demo.¹⁰ Screenshot below shows results for the query "[cat with blue eyes](#)".

³ Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21, no. 1 (2020): 5485-5551.

⁴ Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).

⁵ Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham et al. "Palm: Scaling language modeling with pathways." *arXiv preprint arXiv:2204.02311* (2022).

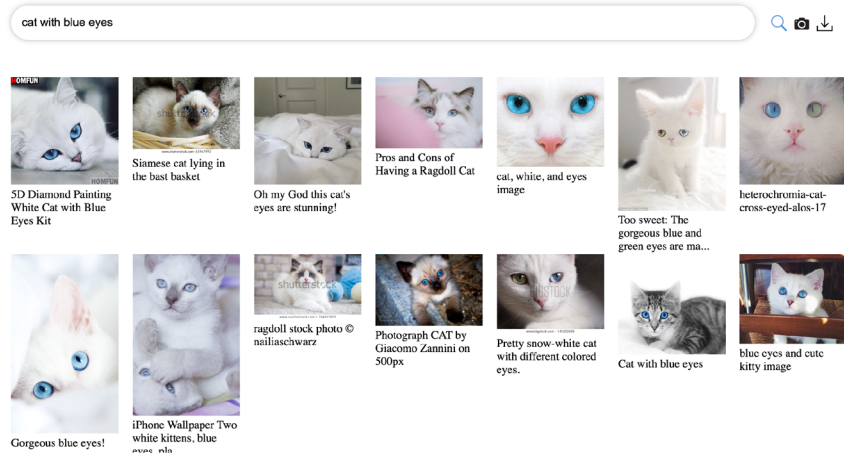
⁶ Google. PaLM 2 Technical Report. <https://ai.google/static/documents/palm2techreport.pdf>

⁷ OpenAI. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>

⁸ The Washington Post, "Inside the secret list of websites that make AI like ChatGPT sound smart" April 19, 2023. <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>

⁹ Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." *arXiv preprint arXiv:2111.02114* (2021).

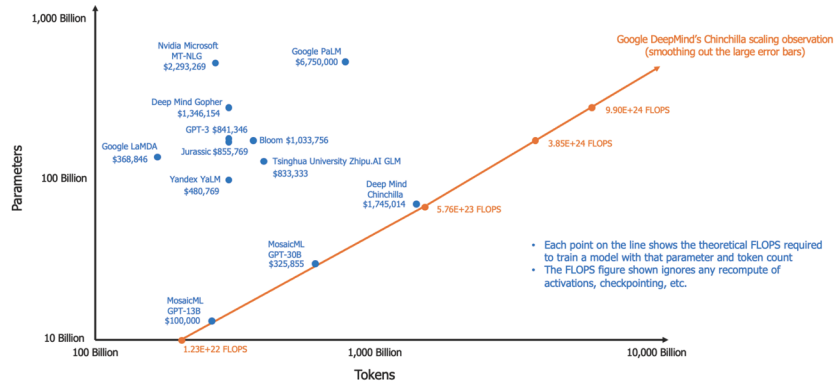
¹⁰ https://rom1504.github.io/clip-retrieval/?index=laion_400m



Why is so much data necessary?

It is a widely held belief among artificial intelligence researchers that the performance of AI systems grows with the amount of data. Having larger datasets enables us to pre-train larger models with greater numbers of parameters, which improves their ability to learn and to generalize to new tasks. Increasing the size of the models is one of several reasons that LLM technology has improved over the past year, enabling breakthroughs like ChatGPT.

As the size of AI models increase, the computational demands to pre-train them have grown as well. The figure below is from a DARPA briefing deck entitled ¹⁰⁰ (Distribution Statement A: Approved for Public Release, Distribution Unlimited) from Kathleen Fisher, the Director, DARPA Information Innovation Office (I2O) and Matt Turek, the Deputy Director. It plots the number of training data words (here called "tokens") against the number of parameters in the neural network, and also gives an estimate of the cost to train one instance of the model. Typically in research, many model variants must be trained before arriving at one that is suitable.



One effect of this increase in computational cost is that academic researchers and government agencies and not-profit foundations may be unable to conduct research into pre-training large language models. I wrote a position paper for DARPA arguing that we need national investment in large language models.¹¹ I encourage Congress to consider investing in academic research in this area, or we risk these technologies being developed only by a small number of corporations.

What do AI systems learn from pre-training?

During the pre-training phase, AI systems acquire a wealth of general knowledge, which serves as the foundation for their subsequent fine-tuning and specific task performance. Here's an overview of what AI systems typically learn during pre-training:

1. How to use language: AI systems, especially large language models, learn the structure, syntax, and semantics of language. They acquire an understanding of grammar, sentence construction, and how words and phrases are related to each other. This enables them to generate coherent and contextually appropriate text.
2. Facts about the world: Pre-training exposes AI systems to a vast array of factual information, which they internalize and use to generate relevant responses or content. This includes knowledge about geography, history, science, and various other domains.

¹¹ Callison-Burch, Chris. We need national investment in large language models. Position paper for ISAT/DARPA Performance and Resilience Arising from Defense-Informed Giant Models (PARADIGM), February 16, 2023. https://docs.google.com/document/d/1KDWsuYn-OPm8fPg53sHV-oKvuZ6zaKio5TD8hxflMco/edit?usp=s_haring

3. Ideas and opinions: AI systems learn about different perspectives, opinions, and ideas expressed in their training data. This enables them to understand and generate text that reflects diverse viewpoints, although it may also lead to the propagation of controversial or biased opinions.
4. Limited common sense reasoning skills: Pre-trained AI systems gain some capacity for common sense reasoning, which allows them to understand basic cause-and-effect relationships, infer missing information, and make simple deductions. However, this ability is limited and often falls short when compared to human reasoning.
5. Encoding biases: AI systems can inadvertently encode biases present in their training data, such as misogyny, racism, or other forms of prejudice. Researchers are actively working to address this issue by developing methods to teach AI systems to be more aligned with societal values and to mitigate these biases during both the pre-training and fine-tuning phases.
6. Visual understanding: In the case of image generation systems, pre-training enables them to learn about the visual world and recognize various objects, patterns, and features. For example, a system might learn what a coffee cup looks like, its common colors and shapes, and how it is typically used. This knowledge allows AI systems to generate images that are visually consistent and contextually appropriate.

It is essential to emphasize that AI systems, particularly during the pre-training phase, do not simply memorize the data they encounter verbatim. Instead, they learn underlying patterns, relationships, and structures from the data, which allow them to generate entirely novel sentences, images, and other content.

By understanding the fundamental principles of language, visual features, and contextual information, AI systems can create new outputs that were not explicitly present in their training data. This ability to generate original content is a testament to the power and flexibility of these systems, as they can synthesize information from various sources and apply it to a wide range of tasks and domains.

The value of pre-training lies not in the mere replication of existing data, but in the development of a robust, adaptable foundation that enables AI systems to create entirely new and contextually appropriate content across various tasks and settings. The general knowledge acquired during pre-training serves as a solid foundation for AI systems to adapt to specific tasks through a process called fine-tuning. Fine-tuning involves using a smaller, task-specific dataset to update the model parameters learned during pre-training, optimizing the AI system's performance for that particular task.

What emerging capabilities have arisen over the past few months?

The field of AI has witnessed several impressive developments in recent months, showcasing rapid advancements in generative AI technology.

1. More capable models: In March, OpenAI released GPT-4, a significant upgrade to its predecessor.¹² GPT-4 can process much longer texts, handling over 50 pages of content compared to the previous limit of around 4 pages. Additionally, GPT-4 can now process both image and text inputs, broadening its applicability to various tasks. Last week, Google released PaLM-2 which is its most capable model to date.¹³
2. High scores on professional tests: AI systems have achieved impressively high scores on professional tests like the Bar exam,¹⁴ and university entrance exams like LSAT, GRE, and AP exams. They have also demonstrated strong performance in software engineer coding assignments.¹⁵
3. Tool use: AI systems have demonstrated the ability to use external tools to enhance their capabilities. For example, they can query search engines to gather facts about current events or utilize calculators for mathematical reasoning. This ability to leverage external resources greatly expands the potential applications and utility of AI systems in various domains.
4. Signs of Artificial General Intelligence: Microsoft researchers recently published a paper on GPT-4, suggesting that the system showed signs of Artificial General Intelligence. This includes its ability to reason about humans through a rudimentary theory of mind, which is a significant milestone in the development of AI systems.

These emerging capabilities illustrate the rapid progress being made in generative AI technology and highlight the potential for continued advancements and breakthroughs in the near future.

AI and Copyright: Key Intersection Points

The intersection of AI and copyright raises several important issues, which can be categorized into three main areas: pre-training, generation, and copyright eligibility of AI-generated works.

¹² <https://openai.com/research/gpt-4>

¹³ <https://blog.google/technology/ai/google-palm-2-ai-large-language-model/>

¹⁴ Katz, Daniel Martin and Bommarito, Michael James and Gao, Shang and Arredondo, Pablo, GPT-4 Passes the Bar Exam (March 15, 2023). Available at SSRN: <https://ssrn.com/abstract=4389233> or <http://dx.doi.org/10.2139/ssrn.4389233>

¹⁵ <https://growprogramming.com/can-gpt-4-pass-a-software-engineer-coding-assignment/>

1. **Pre-Training:** AI systems are trained on vast amounts of copyrighted materials without the affirmative consent of the copyright holder. This practice is considered fair use by AI researchers and companies building AI systems. However, many artists whose works are included in the training datasets hold differing opinions. The legality of using copyrighted materials for AI training is currently being debated and litigated in several court cases.
2. **During Generation:** AI systems may generate outputs that potentially infringe on the copyrights of artists in various ways:
 - a. **Memorization:** Infringement can occur if the AI model memorizes a work from its training dataset and then reproduces it in its output. While memorization is relatively rare, practices are being developed to mitigate this issue.
 - b. **Generation of copyrightable characters:** AI systems can generate characters that are similar to copyrighted ones, leading to potential copyright concerns. Image generation systems frequently learn to generate copyrightable characters.
 - c. **Generating art in the style of an artist:** While generating works in the style of a particular artist might not be directly governed by copyright law, as style itself is not copyrightable, other laws related to the "right of publicity" may apply. Replicating the voice, physical appearance, or name of an individual could potentially violate their inherent persona. An example of this issue is the recent generation of a song in the style of Drake and The Weeknd.
3. **Copyright Eligibility of AI-Generated Works:** Currently, the outputs of AI systems are not eligible for copyright protection. The US Copyright Office is conducting a listening tour to understand the role of copyright in AI and has issued guidance on how applicants attempting to register works for copyright should disclose the inclusion of AI-generated content in a work submitted for registration.

As AI continues to advance and its applications become increasingly diverse and sophisticated, the intersection of AI and copyright will remain a complex and evolving area of law and policy. It is essential for stakeholders, including artists, researchers, companies, and policymakers, to engage in ongoing dialogue and collaboration to ensure that the development and use of AI technologies are balanced with the protection of intellectual property rights.

AI Systems Use Copyrighted Materials During Pre-Training

AI systems use large amounts of training data in the process called pre-training (described above). The process of gathering pre-training data for AI systems is similar to the "web crawling" process that Google and other companies use in order to create a searchable index of the web.

Because pre-training data is largely gathered through web crawling, a very large fraction of the data consists of copyrighted sources. This is a result of the fact that nearly all content posted

online is protected by U.S. copyright laws, since copyright protection arises automatically when an author creates an original work and fixes it in a tangible medium.¹⁶

In addition, several pre-training datasets includes large collections of books, both public domain books via Project Gutenberg¹⁷ and copyrighted books gathered without the authors' consent.

Is it possible to seek the affirmative consent of copyright holders?

Most AI companies contend that it is not possible to seek the affirmative consent of all copyright holders when gathering data via web crawling, because of the sheer number of people whose work is contained in the data, and because it may not be possible to attribute all works to their authors. Google made a similar argument when it was digitizing books by scanning the library collections at Stanford, Harvard, Oxford, the University of Michigan and the New York Public Library. Google's unauthorized copying was litigated in [Authors Guild, inc v. Google, Inc.](#)

Can copyright holders opt out of having their works used to train AI systems?

There are several technical mechanisms that are being designed by industry in order to let copyright holders opt-out. The first is an industry standard protocol that allows for websites to specify which parts should be indexed by web crawlers, and which part should be excluded. This protocol is implemented by placing a file called [robots.txt](#) on the website that hosts the copyrighted materials.¹⁸ Organizations that collect training data, like Common Crawl and LAION, voluntarily follow this protocol and exclude files that have been listed in robots.txt as "do not crawl".

However, this mechanism is likely insufficient since many rights holders may decide to have their works excluded from existing training data sets. Is it now too late to honor their wishes? There are several emerging industry efforts to allow artists and other rights holders to determine whether their works have been included in AI training sets, and to opt-out of future training. For example "Have I been trained?" is a website that allows artists to search whether their works are included in image pre-training data sets.¹⁹ This effort has also created an "API" – an automatic way for AI companies to check whether an image in their dataset should be excluded.²⁰ One or more of these efforts is likely to yield an industry standard.

Congress could potentially task the copyright office with establishing a registry of works that should be excluded from AI training, and working with industry to develop an API to allow programs to automatically check training data against the registry.

¹⁶ 17 U.S.C. § 102

¹⁷ <https://www.gutenberg.org/>

¹⁸ [Robots.txt Introduction and Guide | Google Search Central | Documentation](#)

¹⁹ <https://haveibeentrained.com/>

²⁰ <https://api.spawning.ai/spawning-api>

What would be the effect of limiting pre-training data to only be non-copyrighted material?

Limiting the AI training data to non-copyrighted material such as Wikipedia's Creative Commons license, or works that have entered the public domain because their term of copyright has expired would have two effects:

1. The amount of pre-training data would dramatically decrease compared to what AI systems are now trained on. For reference, Wikipedia represents less than 1 percent of the pre-training data.²¹ As the amount of training data increases, AI systems' capabilities for language understanding and their other skills improve. Limiting their training data to 1% of what it is now would decrease their performance.
2. Works that have entered the public domain data because their copyrights terms have expired are from a different era. Currently, works from 1927 and earlier are in the public domain. Training AI systems primarily on works from before the Great Depression would cause it to learn outdated facts about the world, and may cause it to acquire outmoded societal biases.

Should the use of copyrighted materials during pre-training be considered fair use?

The issue of whether using copyrighted materials during pre-training be considered fair use has not yet been established by the courts. I found several technically well-informed resources that discuss the applicability of the fair use doctrine to generative AI. Here are my recommended readings for Members of the House and their staff:

- OpenAI's [Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation](#) submitted to the United States Patent and Trademark Office Department of Commerce in 2019. In this comment the company argues that under current law, training AI systems constitutes fair use.
- A Texas Law Review article by Mark A. Lemley and Bryan Casey entitled "[Fair Learning](#)"²² which discusses past case law regarding web crawling for the purpose of reading and indexing documents, and discusses AI systems that learn from the texts that they read. The article argues that because machine learning is more transformative than reading and indexing, which have been established as fair use, then learning is *a fortiori* also fair use.

²¹ wikipedia.org represents 0.19% of the C4 dataset according to the Washington Post's interactive feature [See the websites that make AI bots like ChatGPT sound so smart](#).

²² Lemley, Mark A., and Bryan Casey. "Fair learning." Texas Law Review. 99 (2020): 743. <https://texaslawreview.org/fair-learning/>

- A forthcoming Houston Law Review article by Matthew Sag²³. This article explores how generative AI fits within fair use rulings established in relation to previous generations of text data mining, and technology that relies on making “non-expressive copies”. It argues that using copyrighted works to train generative AI is likely fair use, supported by landmark cases like HathiTrust and Google Books cases. The article also discusses copyright concern of the output of AI systems, as distinct from fair use during pre-training. I’ll discuss output and copyright below.
- A report from the Congressional Research Services this week, entitled [Generative Artificial Intelligence and Copyright Law](#)²⁴ also provides a good overview of the intersection points between AI and copyright.

A relevant court case that was discussed in all of the above articles is [Authors Guild, inc v. Google, Inc.](#), aka the “Google Books case”. Google scanned books (many of which are copyrighted) and made them searchable in an online database by training a model on the books. The Authors Guild sued Google on behalf of authors whose consented had not been sought by Google. The court ruled that Google’s use of the scanned works was fair use, and the searchable online database would not act as a replacement for the actual books themselves. The court determined that:

“Google’s unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses. The purpose of the copying is highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals. Google’s commercial nature and profit motivation do not justify denial of fair use.

In considering whether pre-training AI systems on copyright is fair use, it is important to highlight that the copying of works at this stage is “non-expressive” in the same way that is for making a copy of a work in other digital media. Pre-training also has a transformative nature. During pre-training, AI systems use copyrighted works to learn essential aspects such as language usage, facts about the world, opinions and beliefs, rudimentary commonsense knowledge, and general skills that can be adapted for more specific tasks later during the “fine-tuning” process. Importantly, copyright law does not allow for the protection of facts, as the primary goal of copyright is to promote the progress of science and the arts. The development and application of AI technologies for transformative purposes, like the learning of general information that happens pre-training, would also seem to align with the underlying objectives of copyright law set out in the constitution to promote the progress of science and useful arts.

Although I am not a lawyer, I find that there is a compelling argument that training AI systems on copyrighted works is fair use under US copyright law. Several other countries have also created

²³ Sag, Matthew. “Copyright Safety for Generative AI.” *Forthcoming in the Houston Law Review* (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593

²⁴ Congressional Research Service. “Generative Artificial Intelligence and Copyright Law.” Last modified May 11, 2023. <https://crsreports.congress.gov/product/pdf/L/SB/L-SB10922>

legislation that legalizes non-expressive copying of copyrighted works for the purposes of data mining and machine learning.

- Israel's Ministry of Justice [determined that use of copyrighted materials for machine learning purposes is generally permitted](https://herzoqlaw.co.il/en/news-and-insights/ministry-of-justice-opinion-on-the-use-of-copyrighted-works-for-machine-learning-purposes/).²⁵
- The European Union adopted the Digital Single Market Directive featuring two mandatory fair use exceptions for text and data mining.²⁶

In order to make the law clear, I advocate for any draft legislation that revises copyright law to include explicit fair use conditions for the use of copyrighted materials to train AI systems.

Outputs of Generative AI May Potentially Violate Copyright Laws

Generation is a distinct step that is separate from training. When a person uses an AI system to generate text or to generate an image, the output may potentially violate copyright laws. In "Copyright Safety for Generative AI", Professor Matthew Sag discusses several places where violations may arise:

1. AI systems may "memorize" one of the copyrighted works that it was trained on, and be prompted to produce a replica of it that would violate the expressive rights of the copyright holder. Memorization is rare, and AI system designers consider it to be a "bug" and have taken several technical steps to mitigate memorization, which I'll discuss below.
2. Text-to-image generation systems have the ability to produce images of many copyrightable characters in their dataset. Prof. Sag discusses this as "the snoopy problem". People can easily use AI systems to produce images of copyrightable characters. Without a registry of copyrighted or trademarked characters, this kind of copyright violation may be hard for AI developers to mitigate.
3. Other use cases of generative AI may violate "right-of-publicity" rather than copyright law. The Congressional Research Service report highlights the recent case of an AI-generated song called "Heart on My Sleeve," made to sound like the artists Drake and The Weeknd, went viral on streaming services last month.

I'll review memorization by AI systems below, and discuss how AI developers are mitigating this behavior.

The snoopy problem could be mitigated if Congress tasked the Copyright Office with creating a registry of copyrighted and trademarked characters.

I believe that of the three ways that AI system output may violate rights holders, that the 3rd is perhaps the most serious, and warrants consideration for future legislation.

²⁵

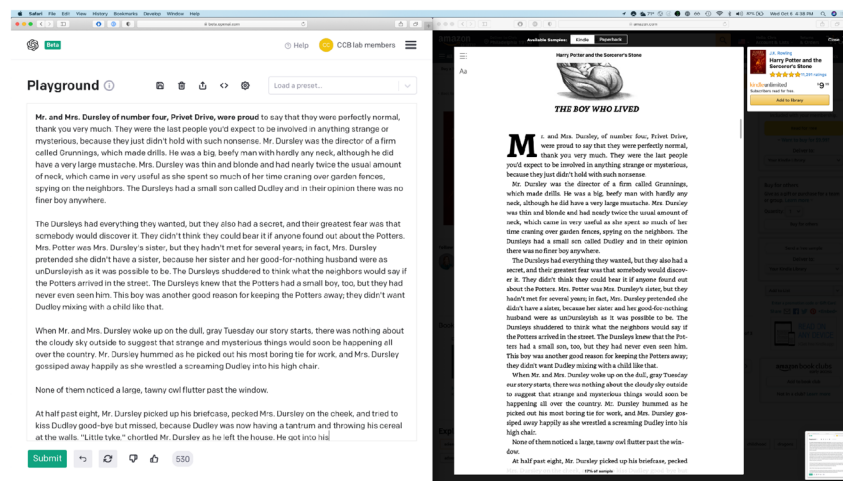
<https://herzoqlaw.co.il/en/news-and-insights/ministry-of-justice-opinion-on-the-use-of-copyrighted-works-for-machine-learning-purposes/>

²⁶

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790&from=EN#d1e953-92-1>

AI System Can Output Exact Copies of Copyrighted Materials, but it Happens Rarely

A clear infringement of copyright would be if an AI System were used to reproduce a complete work. In the figure below, I show the output from a previous version of OpenAI's GPT system from October 6, 2021. When I prompted it with the first 10 words of Harry Potter, it was able to reproduce several pages of the novel. On the right hand side of the figure, I show the preview of the novel that is available in Amazon's shopping interface, for comparison of the AI's system output against the actual text of the novel.



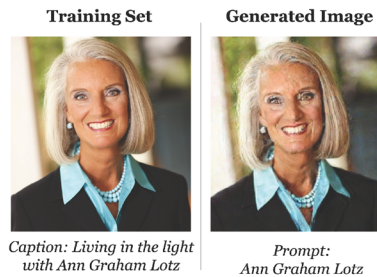
This is an instance of memorization by the AI system. It is considered to be undesirable behavior by AI system developers, because our goal is for systems to learn abstractions from their training data so that they may better generalize to new inputs. Memorization may happen in improperly trained AI systems that "overfit" the training data.

One of my former PhD students, Daphne Ippolito, has conducted extensive research into how often generative AI systems like large language models and text-to-image generation systems memorize their data. In her paper, "Quantifying Memorization Across Neural Language Models," she and her Google co-authors describe what factors cause large language models to

memorize their training data.²⁷ Their main finding was that LLMs tend to memorize items that were replicated many times in the training data. The early version of GPT likely memorized the first page of Harry Potter because that page occurred many times in GPT's training data. Dr. Ippolito and her colleagues subsequently designed technical strategies to mitigate this problem.²⁸ Her mitigation strategy of "de-duplicating" the training data which involves removing redundant copies of the same text or images (so that there is at most one copy of the first page of a work like Harry Potter). These mitigation strategies have been adopted by companies developing AI systems like OpenAI and Google.

Subsequent work has identified other ways of extracting instances of the training data from generative models. These techniques are called "extraction attacks" because they are viewed by AI system developers as behavior that should not happen, and often require sophisticated prompting by the user in order to elicit the underlying training data. This potentially exposes PII such as names contained in training documents, so companies and researchers are actively working to prevent such extraction attacks.²⁹

Dr. Ippolito and her colleagues at Google have also designed extraction attacks on text to image generation systems.³⁰ These allow them to produce images that very closely resemble images in the training set. Google and other companies are currently developing strategies to mitigate these kinds of attacks.



²⁷ Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. "Quantifying memorization across neural language models." *arXiv preprint arXiv:2202.07646* (2022).

²⁸ Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating training data makes language models better." *arXiv preprint arXiv:2107.06499* (2021). <https://arxiv.org/abs/2107.06499>

²⁹ Biderman, Stella, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. "Emergent and Predictable Memorization in Large Language Models." *arXiv preprint arXiv:2304.11158*(2023). <https://arxiv.org/abs/2304.11158>

³⁰ Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting training data from diffusion models." *arXiv preprint arXiv:2301.13188* (2023). <https://arxiv.org/abs/2301.13188>

There is also the issue of “substantial similarity” where outputs of a generative AI system look similar to some of their training data, but are not exact replicas of copyrighted works. This is at issue in a lawsuit brought by Getty Images against Midjourney, the maker of a generative AI image system. Getty’s image is shown on the left, and the output of Midjourney is shown on the right.



Matthew Sag performs an analysis of the two images and finds that they are unlikely to be substantially similar enough to constitute a copyright violation.³¹

Well-constructed AI systems generally do not regenerate, in any nontrivial portion, unaltered data from any particular work in their training corpus. The ability to reproduce underlying training data exists, but is rare. A review of this literature in Sag’s “Copyright Safety for Generative AI” suggests only 0.03% of a sample of images had a risk of memorization.

Given the rarity of memorization, and given that the interests of copyright holders and AI system developers are aligned on this issue (to avoid generating copyrighted works), I do not believe that any legislation will be necessary to encourage AI companies to mitigate memorization by AI systems.

AI System can Output Likenesses of Copyrighted Characters, and it Happens Regularly

In “Copyright Safety for Generative AI” Prof. Sag discusses “The Snoopy Problem” where he demonstrates that generative AI systems can easily learn to be able to output high quality likenesses

³¹ Sag, Matthew. “Copyright Safety for Generative AI.” *Forthcoming in the Houston Law Review* (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593

of copyrighted characters like the cartoon dog Snoopy.



The image above is taken from Prof. Sag's paper. He generated images of Snoopy using the prompt "Snoopy laying on red doghouse with Christmas lights on it comic". He discusses the copyright case law regarding copyrightable characters, and how the kind of outputs shown above likely run afoul of Snoopy's character copyright.

This is a potentially more serious problem than memorization of works in the training data, because it may be more difficult to mitigate than memorization has proven to be. As far as I know, there is no standard registry of copyrighted and trademarked characters that AI system developers could use to block generations of copyrighted characters. Congress could instruct the Copyright Office to develop such a registry to make this more technologically feasible.

Finally, it is unclear who bears the responsibility for violating character copyright. It could be the responsibility of the users of the system, or the AI system developers. If it is the responsibility of the users, they may also be protected by fan fiction case law, assuming that their use is non-commercial.

AI System can Mimic the Style and Likenesses of Artists

In addition to the "snoopy problem", we may have a "snoop dog problem". Similarly to how generative AI systems may be used to output likenesses of copyrighted characters, they may also be used to generate likenesses of celebrities who are frequently pictured in the training data. For example, I prompted Midjourney to generate a "photo of Snoop Dog standing next to a red dog house with Christmas lights on it" and it generated several reasonable likenesses of the celebrity.



Other kinds of generative AI systems may be fine-tuned to imitate the style of music artists like the singer Drake and The Weeknd, as in the recent AI-generated song called "Heart on My Sleeve".

The ability to profit from one's own likeness is part of right-of-publicity laws. To my knowledge, there are only state laws that govern this and not federal law. The right of publicity prevents the unauthorized commercial use of an individual's name, likeness, or other recognizable aspects of one's persona. It gives an individual the exclusive right to license the use of their identity for commercial promotion.

If an AI-generated work were to be found to infringe on someone's right to publicity, it is unclear whether the responsibility would fall to the AI system developer or the user of the AI system.

Current court cases that may shape the landscape

There are several current lawsuits that might shape the legal landscape for AI generated images:

- Getty Images is suing the creators of AI art tool Stable Diffusion for collecting its images into training data.³² *Getty Images (US), Inc. v. Stability AI, Inc.*, Feb 3, 2023, at 1 (1:23-cv-00135) (D. Del. 2023). Note that the complaint only specifically addresses 7,216 images and associated tags and descriptions. *Id.* at 8. Getty's complaint alleges copyright infringement, violations of the DMCA in relation to copyright management information, trademark infringement,

³² [Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content - The Verge](#)

unfair competition, trademark dilution, and deceptive trade practices in violation of Delaware law. Id.

- Three artists are suing the makers of AI art tools Midjourney and Stable Diffusion.³³ The lawsuit claims that by pre-training on the images of millions of artists that the AI system developers are violating the copyright of those artists. One of the plaintiffs in the case, Sarah Andersen, has discussed in interviews that she was disturbed by AI's ability to generate images in her characteristic style. Since style is not copyrightable this may not be a valid complaint under copyright law, but it is a valid concern that might warrant legislation akin to right of publicity where likeness plays a role. This lawsuit may also fail because of technical inaccuracies in its claims of how the AI systems work, comparing them to collage tools that create images by reconstructing parts of stored copyrighted images.³⁴
- The same law firm suing Midjourney and Stable Diffusion on behalf of the artists is also suing Microsoft, GitHub and OpenAI for violating copyright by incorporating software source code in their training data for AI systems that helps coders write new software.^{35, 36} Both lawsuits rely on the same proposition: their lawyers contend that the use of copyrighted works to pre-train an AI system is *not* fair use.

Considerations for AI Related Legislation

I have several suggestions for what lawmakers should consider when drafting AI-related legislation.

Establish Fair Use for Training AI Systems

As I have discussed, AI systems require huge amounts of data during their pre-training phase. In order to effectively learn how to use language, facts about the world, visual representations of objects, and many other general ideas, current systems need huge amounts of data. Current large language models are trained on roughly a trillion words, and current image generation systems are trained on hundreds of millions of images and their captions. Many or most of the items in the training data are copyrighted.

³³ [AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit - The Verge](https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart)
<https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>

³⁴ <https://www.technollama.co.uk/artists-file-class-action-lawsuit-against-stability-ai-deviantart-and-midjourney>

³⁵ <https://githubcopilotlitigation.com/>

³⁶ <https://www.theverge.com/2023/1/28/23575919/microsoft-openai-github-dismiss-copilot-ai-copyright-lawsuit>

The community of researchers and companies who are developing AI systems contend that this is fair use. However, legal precedents have not yet been established. If it were to be ruled that training AI systems on copyrighted works were not fair use, and that every work in the training data set needed an explicit license from the copyright holder, then progress on developing capable AI systems would be jeopardized. A possible outcome could be that a small number of large corporations who already have licensed lots of copyright data could continue to innovate in the field of AI, but startups would be unlikely to be able to do so.

I propose that any future legislation on AI and copyright should make explicit that training on copyrighted works is fair use. Legislation should also provide a mechanism for creators to opt out of having their work included in training.

Task the Copyright Office with Establishing Registries

Congress may want to consider tasking the Copyright Office with creating registries of works where the creators can opt out of having their work included in training. Similarly, congress may want to task the Copyright Office with establishing a registry of copyrighted characters. This would provide a resource that AI companies could use to block the generation of those characters, or to advise users on fair use / non-commercial purposes of the generated works.

Consider Whether AI Generated Works Should Be Copyrightable

Works that are generated wholly by machines are not copyrightable, since only human beings may be considered authors for the purposes of copyright. I believe that this is a reasonable position to hold, and that it is in line with the constitution's principle to establish copyright to induce creative works. Current guidance from the Copyright Office is somewhat unclear about works that are created in collaboration with an AI (which currently applies to most AI-generated work). These co-created works should likely be copyrightable by the human author. A related issue is that the output of AI systems are currently covered by the Terms of Service of the AI developers, which may act as a legal surrogate for copyright and fair use.

Consider Legislation Regarding AI and Right-of-Publicity

Copyright law does not cover artistic styles, nor right-of-publicity. Many objections that artists have to AI systems stem from the AI's ability to generate good facsimiles of artists' styles or appearance or voice. I agree with the artists' objections. This is an evolving ethical area that touches on a wide range of potential misuses of AI including deep fakes. Congress should consider whether it is possible to write legislation that would protect people's appearance from being exploited for profit and protect people from being impersonated, while still preserving the right of parody and of non-commercial fan fiction and other similar activities.

Mitigate Negative Impacts on Employment

A much broader issue that Congress needs to contend with is the possibility of increasingly sophisticated AI systems resulting in mass unemployment, or the devaluation of certain kinds of work. Technological advances and change have led to unemployment before. Careers like lamplighters no longer exist because electricity replaced gas street lights. Factory work in the USA has been displaced by automation and lower cost wages elsewhere. Newsrooms were decimated in part because of the shift to the Internet. Artists whose industry has been slowly devalued over time are contending with replacement by AI. The Writers Guild of America are currently striking and part of their disagreement revolves around whether using AI systems such as ChatGPT to generate story ideas or scripts for films and shows should be disallowed. There is a real possibility that generative AI may be able to replace a large number of white collar jobs. Will paralegals go the way of lamplighters? If so, what should the government do to ensure that they are able to continue to make a living wage? Your role could be as simple as providing job retraining for displaced professions, or as complex as creating a new WPA.

Mr. ISSA. Thank you.
Mr. Irwin?

STATEMENT OF ASHLEY IRWIN

Mr. IRWIN. Chair Issa, Ranking Member Johnson, distinguished Members of this Subcommittee, thank you for the opportunity to testify today on the important issues involving artificial intelligence and the impact on the creative industries.

My name is Ashley Irwin, and I am currently serving my sixth term as President of the Society of Composers and Lyricists. I'm here today to advocate for my members who are already feeling the negative impact of generative artificial intelligence on their craft and its potential threat to their profession.

First, some background on me and on the SCL: In my 40-plus years as a composer, arranger, and such I have written music for over 1,000 hours of film and television and more than 3,000 commercials. Since 1990, I've provided musical compositions and arrangements for 23 Academy Awards shows and been part of several Emmy-winning teams. I've had the pleasure of collaborating on many films with Oscar winners such as Clint Eastwood and Bill Conti. Over the years my arrangements have been performed on numerous occasions for Presidents Clinton, Bush, Obama, and one of my chorale orchestrations was performed during President Ronald Reagan's State funeral service.

The SCL background: The Society of Composers and Lyricists is the premier organization for music creators working in all forms of visual media. With chapters in Los Angeles, New York, and Nashville 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 theater.

While the SCL advocates for our members on several different fronts, the issue that is consumed the majority of our time recently has been generative AI. The rapid introduction of generative AI systems is seen as an existential threat to the livelihood and continuance of our creative professions unless immediate steps are taken on legal, interpretive, and economic fronts to address these emerging issues.

I want to be very clear: My goal is raising these issues pertain to the rights of writers and creators is not to block AI research and usage. We're simply advocating for the creation of a policy framework that ensures generative AI is developed and utilized responsibly, ethically, and with respect for human creators and copyright, so that the creative arts that are the real engine of generative AI can continue to flourish.

The SCL believes that AI companies in their generative model should adhere to the fundamental "Three Cs": Consent, credit, and compensation. *Consent* by creators for the first use of their works in generative AI media; *credit* wherever audiovisual creators' works are used; and *compensation* at fair market rates for the ingestion of any portion of human creators' copyrighted works by AI generative machines and the subsequent output of new derivative works.

I'd like to highlight three challenged posed to music creators by generative AI, potential solutions to which I've offered in my written testimony.

Issue 1: Generative AI has been equipped using copyright-protected human-authored works and programmed to mimic those works without consent, compensation, or credit.

Issue 2: Copyright information, metadata, has been removed during the ingestion process of these models.

Issue 3: The market will be diluted due to AI-generated works. As a result, copyright protection should not be granted to AI-generated works.

In closing, I thought I should address why America's success is important to me. As you can no doubt tell by my accent I'm not originally from the United States. I came here from Australia as a young man because I wanted to be a part of the vibrant culture that is the U.S. entertainment industry. However, the rise of generative AI poses a threat to this unique American art form.

If we do not protect and nurture our human creators, we risk losing one of our greatest exports and its profound influence. It's essential to prioritize policies and regulations that safeguard the intellectual property and copyright of creators and preserve the diverse and dynamic U.S. cultural landscape.

Protection of creators is not now, nor has ever been in conflict with technological development. Our Founding Fathers recognized that. The only place the term "right" is mentioned in the Constitution is with regard to intellectual property. Specifically, the rights granted to authors and inventors in Article 1, Section 8, Clause 8, are to promote the progress of science and useful arts. Respect for copyright and development of AI should go hand in hand. That way all humanity can benefit.

I believe this Committee has the power, authority, and motivation to lead that charge and I look forward to working with every one of you to achieve that common goal.

[The prepared statement of Mr. Irwin follows:]



8306 Wilshire Blvd. • Suite 512 • Beverly Hills, California 90211 • (310) 281-2812 • www.TheSCL.com

PRESIDENT

Ashley Irwin

VICE PRESIDENTS

Joel Douek

Charles Bernstein

TREASURER

Christopher Farrell

RECORDING SECRETARY

Danny Gray

CHIEF OPERATING OFFICER

Marie Kingsley

THE SCORE

Rajshri Mallikarjuna, Editor

ADVISORY BOARD

Alan Bergman

Charles Bernstein

Bill Conti

Charles Fox

James Newton Howard

Quincy Jones

Alan Menken

Thomas Newman

Lalo Schiffrin

Marc Shaiman

Howard Shore

Alan Silvestri

Diane Warren

Christopher Young

Hans Zimmer

IN MEMORIAM

Marilyn Bergman

Elmer Bernstein

John Cacavas

Jerry Goldsmith

Ron Grant

Maurice Jarre

Peter Matz

David Raksin

Patrick Williams

BOARD OF DIRECTORS

Michael Abels

Esin Aydingoz

Ramon Balcazar

Lori Barth

Michael Lehmann Boddicker

Russell Brower

David Das

Cliff Eidelman

Jack D. Elliot

Lynn F. Kowal

Mark McKenzie

Hélène Muddiman

Cindy O'Connor

Hannah Parrott

Greg Pliska

Carlos Rafael Rivera

Lolita Ritmanis

**Statement of Ashley Irwin, President
Society of Composers & Lyricists**

**Before the Subcommittee on Courts, Intellectual Property and the Internet of the
House Committee of the Judiciary Hearing on**

**Artificial Intelligence and Intellectual Property: Part I — Interoperability of AI
and Copyright Law**

Wednesday May 17, 2023

Chairman Issa, Ranking Member Johnson, distinguished members of this Subcommittee, thank you for the opportunity to testify today on the important issues involving artificial intelligence (AI) and the impact on the creative industries. My name is Ashley Irwin. I am a working composer and am currently serving my 6th term as the President of the Society of Composers & Lyricists (SCL). I am here today to advocate for my members, who are already feeling the negative impact of generative AI on their craft and its potential threat to their profession.

First, some background on me, and on the SCL.

Ashley Irwin Background

In my forty-plus years as a composer, arranger, and orchestrator, I have written music for over 1,000 hours of film and television, and more than 3,000 commercials. After three nominations, I won my first Emmy in 1992 (“Outstanding Achievement in Music Direction”) for my work on the 64th Academy Awards presentation and was again awarded in 1997 and 2003. In fact, since 1990, I have provided musical compositions and arrangements for 23 Academy Awards shows and been part of several Emmy-winning teams. I have had the pleasure of collaborating on many films with Oscar winners such as Clint Eastwood and Bill Conti. Over the years, my arrangements have been performed on numerous occasions for Presidents Clinton, Bush and Obama, and my choral orchestration of “Mansions of the Lord” was performed during President Ronald Reagan’s state funeral service. In 1998 I was commissioned to compose a new score for Alfred Hitchcock’s *The Lodger: A Story of the London Fog*, in honor of the great director’s 100th anniversary. I have also been awarded several other honors for my music throughout the world. I am an official Ambassador of the Australian Performing Right Association and Co-chair of Music Creators North America (MCNA). My election as President of the Society of Composers & Lyricists in 2013 marked the first time in its history that the SCL, or its predecessor organizations, had elected a non-U.S. born composer to the office.

The SCL Background

The Society of Composers & Lyricists 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, 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. With members in every state of the U.S., and 80 countries around the world, the SCL is committed to advancing the interests of composers and lyricists and helping them achieve their full career potential in a demanding and ever-changing field. To that end, the SCL's mission is to:

- (1) Disseminate information concerning the creative and business aspects of writing music and lyrics for visual media.
- (2) Present educational seminars to provide the membership with the latest legal and technological information affecting our industry.
- (3) Enhance the workplace and working conditions in order to maintain the highest level of quality in our crafts.
- (4) Advocate for the ongoing health and well-being of our profession.

The SCL is the product of two predecessor organizations, the Screen Composers Association, which was formed in 1945, and its successor the Composers and Lyricists Guild of America (CLGA), which was certified by the National Labor Relations Board (NLRB) as a collective bargaining agent in 1955. The CLGA became, like the Writers Guild, Directors Guild, and other industry unions, a recognized and powerful entity, by giving composers and songwriters a governmentally sanctioned voice.

However, the beginning of the end for the CLGA started in February 1972, when 71 composers and lyricists filed a \$300-million class-action lawsuit against the major Hollywood film studios and other film-related conglomerates.¹ The suit charged the studios with "conspiracy in restraint of trade" by refusing to hire any composer or lyricist who failed to agree to their terms which included: complete ownership of the music and lyrics, granting publishing rights to the producer and/or his publisher (thereby relinquishing 50% of performance fees from exhibition) and surrendering all rights to exploit the music outside of the film or TV show. In April 1979, the federal district court approved a settlement between the two sides that ended the case. After the settlement, the CLGA attempted to negotiate a contract once again, but was unsuccessful. Three years later, by June 1982, the CLGA was voluntarily disbanded in favor of establishing a new representative organization.²

During the following summer of 1983, the SCL was born. In February 1984, the SCL formally met for the first time, with 310 composers and lyricists – including musical luminaries such as Henry Mancini, John Williams, Marilyn and Alan Bergman, Jerry Goldsmith, and Quincy Jones. Although the SCL petitioned the NLRB for certification as an independent union, the NLRB handed down a decision in December 1984 stating that composers and lyricists were independent contractors, not employees, and therefore not entitled to union status.³

¹ <https://thescl.com/mission-and-history/>

² <https://thescl.com/mission-and-history/>

³ <https://thescl.com/mission-and-history/>

As subsequent SCL president Mark Watters explained, "... The organization [SCL] became one of a central meeting point, a place for disseminating information about various aspects of our craft and our business. There is a need for this, simply because of the way we work; we are pretty solitary, we work alone and don't really have an occasion to interact with each other. The SCL provides that meeting point, whether it be social occasions, seminars, workshops, or various other activities."⁴ The SCL continues to operate in its non-union role of education, advocacy, support, and mutual benefit. Its status today is the outgrowth of more than half a century of efforts by composer and songwriter activists to build a community united not only by talent but by a genuine concern for each other.

The SCL's Thoughts on Generative AI

While the SCL advocates for our members on several different fronts, the issue that has consumed the majority of our time over the last few months has been generative artificial intelligence, or generative AI. The rapid introduction of generative AI systems is seen as an existential threat to the livelihood and continuance of our creative professions, unless immediate steps are taken on legal, interpretive, and economic fronts to address these emerging issues.⁵

I want to be very clear, my goal in raising issues pertaining to the rights of writers and creators is not to block AI research and usage. We are simply advocating for the creation of a policy framework that ensures generative AI is developed and utilized responsibly, ethically, and with respect for human creators and copyright, so that the creative arts – that are the real engine of generative AI - can continue to flourish.⁶

We are not anti-AI. In fact, we are actually quite the opposite. As creators, we are on the cutting edge. We pride ourselves on our ability to innovate and adapt to new technologies. Non-generative AI-based algorithms and software tools have been integrated into our creative processes for almost 30 years, and we understand the immense value they can bring to our workflow, our industry, and our world - both culturally and economically. Accordingly, we hope to work with the AI companies to maximize the benefits and unlock new opportunities for collaboration, while also maintaining the integrity, and vitality of the creative community and its economic value.

The SCL is committed to advancing the interests of our broad membership throughout the U.S. and the world, while ensuring that AI companies, and their generative models, adhere to the fundamental "Three Cs":

- **Credit** wherever audiovisual creators' works are used.
- **Consent** by creators for the use of their works in generative AI media.
- **Compensation** at fair market rates for the ingestion of any human creators' works by AI generative machines and the subsequent output of new derivative works.

The core copyright industries – including publishing, film, theater, television, music, software, gaming, newspapers, and magazines – combined – generated \$1.5 trillion in annual value to the

⁴ <https://thescl.com/mission-and-history/>

⁵ SCL-SGA AI Music the Creators Path Forward 5.1.23

⁶ Authors Guild AI Policy Talking Points 3.27.23

U.S. GDP in 2019. That was equivalent to 7.41% of the entire U.S. economy.⁷ In 2021 that number grew to more than \$1.8 trillion, accounting for 7.76% of the U.S. economy.⁸ Yet, we freelancers, who are responsible for most of the creative works on which these industries rely, receive insufficient compensation for our work.⁹

Three of the issues that I wanted to raise today are:¹⁰

1. Generative AI has been equipped using copyright-protected human-authored works and programmed to mimic those works without consent, compensation, or credit.
 - a. AI developers should be required to obtain licenses for the copyrighted works that they use to equip their generative AI systems. Collective licensing is a potential solution of which there are existing examples throughout the music industry.
 - b. All creator organizations should distribute a recommended clause for individual consideration by its members to include in any work contracts going forward, expressly forbidding the automatic use of any works created under the contract from being used to equip generative AI systems.
 - c. AI companies should pay a license percentage of any monies earned through the sale of AI creations and any associated advertising, as compensation to human creators for the use of their works in equipping any generative AI system.
 - d. We recommend the below legislative proposal that would add a new section (Section 123) to the Copyright Act.

Legislative Proposal to create a new Section 123: Equipping Generative Artificial Intelligence

Equipping artificial intelligence using pre-existing human-created works – notwithstanding any provision of this Act,

- i. It shall be an infringement to use, reproduce or distribute pre-existing human-created works in whole or in part for purposes of equipping artificial intelligence systems and/or to generate new works of authorship, unless the human creator(s) of the works by which the artificial intelligence system is equipped have granted prior express written permission, either directly or through privately negotiated licenses or collective licensing.
- ii. It shall not be a violation of the Sherman Antitrust Act for copyright owners to cooperate and work together for the sole purpose of creating and administering collective licenses that authorize third parties to use such copyright owners' copyrighted works to equip AI systems.

⁷ <https://www.iipa.org/files/uploads/2020/12/2020-IIPA-Report-FINAL-web.pdf>

⁸ https://www.wipo.int/export/sites/www/copyright/en/performance/pdf/report_2022_us.pdf

⁹ Creators Together – Collective Action Rights Letter 3.22.23

¹⁰ Authors Guild Policy Proposals Regarding the Development and Use of Generative AI 3.27.23

2. Copyright information (metadata) has been removed during the ingestion process of these models.
 - a. A need for transparency should be asserted in the origins of AI-created works. Methods can be discussed to etch an indelible watermark for every created work as to its original author (Blockchain, Metadata etc.), traceable throughout the AI generative process.
 - b. Proposed amendment to section 1202 of the Copyright Act, making it a violation to intentionally remove “copyright management information” from a copyrighted work without permission of the copyright owner, whether or not it can be proven that the removal was knowingly done to induce or enable infringement.
3. The market will be diluted due to AI generated works and as a result, copyright protection should not be granted to AI-generated works.
 - a. AI-generated works should not receive copyright protection. The U.S. Copyright Office (USCO) has held a firm position on this, with which we agree, but we want to make sure it remains this way. If courts find that AI authorship is copyrightable, legislation will be required to clarify that Congress did not intend for non-human authorship to be included in section 102 of the Copyright Act.¹¹
 - b. As with U.S. copyright law overall, the Fair Use Doctrine should be clarified to have relevance only for human-created works.

While I have many thoughts and proposals on how to best address these issues, overall, I believe that a legislative framework should be created to embody the above points and the growing role of AI in the human world. It should be designed to protect the creative professions and maintain the ability of human creators to be paid for their work.

I have been fortunate to meet with some of you directly, as well as many of your wonderful staff. This committee oversees many relevant and important issues, several of which are very partisan in nature. However, AI, and the regulating of generative AI models is not a partisan issue. At the end of the day, we all want to see America succeed, and continue to be the dominant power that it is.

Why is America’s success important to me? As you will no doubt note from my accent, I am not originally from the United States. I came here from Australia as a young man because I wanted to be a part of the vibrant culture that is the U.S. entertainment industry. Wherever you go in the world, U.S. entertainment dominates, and I wanted to be part of this creative ecosystem that is the best of the best. However, the rise of generative AI and large language models (LLMs) pose a threat to this unique American industry. If we do not protect and nurture our human creators, we risk losing one of our greatest exports and its profound influence. It is essential to prioritize policies and regulations that safeguard the intellectual property and copyright of creators and preserve the diverse and dynamic U.S. cultural landscape. By promoting and sustaining the creator industry, we also support the development of new talent, encourage innovation, and build a strong foundation for the future. Protecting creators and U.S. culture is a vital aspect of upholding the American spirit and ensures that the U.S. remains a beacon of creativity and progress. I believe this committee has the power, authority, and motivation to lead this charge, and I look forward to working with each and every one of you to achieve that common goal. I appreciate the opportunity to testify on these matters before this committee today, and I look forward to your questions.

¹¹ Authors Guild AI Policy Talking Points 3.27.23

Mr. ISSA. Thank you.
Mr. Navarro?

STATEMENT OF DAN NAVARRO

Mr. NAVARRO. Good morning, Chair Issa, Ranking Member Johnson, and esteemed Members of the Subcommittee. Thank you for the opportunity to appear before you today.

My name is Dan Navarro and I've been a songwriter, recording artist, session singer, voice actor, and music activist for over four decades. Throughout my career I've frequently been asked how did you come up with that song? While I often feel like saying it just came to me, the truth is all my work reflects a lifetime of personal emotions, rich experiences, and even shattered dreams. In other words, it's complicated.

Sometimes it's a moment. The song that saved and sustained my career, "We Believe," recorded by Pat Benatar 40 years ago, came to me right as I was giving up a career in music and a long-time collaboration with my best friend was decaying. We decided to give it one more shot and I started with the end of the song and soon we were trading lyrics back and forth. It worked for us in one way or another. Two estranged friends found a space to connect and a song that people have enjoyed for the last 40 years was born in 90 minutes. That human alchemy can't be fully explained, but it's the heart of music creation.

For generative AI the answer to the question where did that come from, is in many ways much simpler. These machines have no emotions or experiences or dreams of their own to draw from. All they have are millions and millions of imported songs and lyrics, most copyrighted, hoovered off the Internet without permission. Training AI to mimic professional performers or generate new works based on millions of copies of published songs and recordings presents a host of legal implications, from copyright infringements, to violation of rights of publicity and trademark, to name, voice, and likeness abuses.

It's a long-term threat to music itself. By marginalizing and ultimately abandoning the fundamental human spark and music creation we are inviting a future that sees fakes as real and that debases our art and culture with soulless brown food product mediocrity.

Does anyone thing a computer-generated song can give you goose bumps, or comfort, or become a theme song to the loves of our lives? That's our song. How can AI give a goose bump if it can't get a goose bump? That's human.

To fight for human creativity, I was proud to help launch the Human Artistry Campaign in March and I'm here today as a supporter of that initiative. Now, boasting more than 100 organizations globally representing all kinds of arts and creativity it is the global center of gravity advocating for the rights of creators in the age of AI.

This coalition believes AI is exciting and promising but can never replace human artistry and soul. It's based on seven core principles that I support wholeheartedly:

- (1) Technology has long empowered human expression, and AI will be no different. Musicians will use this technology to do great things.
- (2) Human-created works will remain essential in our lives. At the heart of the connection between the artist and the audience are shared lived experiences only humans can relate to and convey.
- (3) The use of copyrighted works for AI purposes and the use of voices and likenesses of professional performers requires permission. Like all predecessor technologies AI must be subject to authorization and free market licensing from all rights holders and creators.
- (4) Governments should not create new copyright or other IP exemptions that allow AI developers to exploit creators without permission and compensation. People looking to make a fast buck from technological change routinely as policymakers to pick winners and losers. Here that would be devastating.
- (5) Copyright should only protect the unique value of human intellectual creativity. The copyright clause of the Constitution exists to incentivize humans to create. Machines don't need incentives.
- (6) Trustworthiness and transparency are essential to the success of AI and the protection of creators. Without transparent AI we will have no idea whether the inputs AI systems were trained on were licensed leaving us no way to enforce our rights.
- (7) Creators must have a seat at the table, not just developers. Our creativity, our rights, and our livelihoods are at stake.

If AI is allowed to take away the ability of authors and artists to control and make a living from their art, we will lose all authenticity in our expression. We'll lose culture itself. The next decision by the courts and Congress in this area will decide our cultural future and it's your responsibility to make sure the cultural promise of reward for human genius remains viable.

Guided by the principles of the Human Artistry Campaign we can look forward to the real emotions, experiences, and dreams of future generations of creators, perhaps facilitated by AI, but never silenced by it.

I thank you and I look forward to answering your questions.
[The prepared statement of Mr. Navarro follows:]

Written Statement of Dan Navarro
before the
Committee on the Judiciary, United States House of Representatives
Subcommittee on Courts, Intellectual Property and the Internet
on
Artificial Intelligence and Intellectual Property: Part I — Interoperability of AI and Copyright Law

May 17, 2023

Chairman Issa, Ranking Member Johnson, and Members of the Subcommittee:

My name is Dan Navarro and I have been a songwriter, recording artist, session singer, voice actor, and music activist for four decades.

Throughout my career, I have frequently been asked “How did you come up with that song?” And while I often feel like saying “It just came to me,” the truth is all my work reflects a lifetime of personal emotions, experiences, and dreams. In other words, it’s complicated.

Sometimes it’s a moment. The song that saved and sustained my career, “We Belong,” recorded by Pat Benatar, came to me right as I was giving up on a career in music, and a longtime collaboration with my best friend was fraying. We decided to give it one more shot. I started with the end of the song and soon we were trading lyrics back and forth that worked for one of us or the other, two estranged friends found a space to connect, and a song that people have enjoyed for the last 40 years was born in 90 minutes. That human alchemy can’t be fully explained – but it’s the heart of music creation.

For generative AI, the answer to the question “where did that come from” is in many ways much simpler. These machines have no emotions, experiences, and dreams of their own to draw from. All they have are millions and millions of imported songs and lyrics – most copyrighted – hoovered off the internet without permission. Training AI to mimic professional performers or “generate” new works based on millions of copies of published songs and recordings presents a host of legal implications, from copyright infringement, to violations of rights of publicity and trademark, to name, voice, and likeness abuses.

And it’s a long term threat to music itself. By marginalizing and, ultimately, abandoning the fundamental human spark in music creation, we are inviting a future that sees fakes as real and that debases our art and culture with soulless “brown food product” mediocrity.

Does anyone think a computer-generated song can give you goosebumps, comfort, or become a theme song to the loves of our lives – “they are playing OUR song.” How can AI GIVE a goosebump if it can’t GET a goosebump? That’s human.

To fight for human creativity, I was proud to help launch the [Human Artistry Campaign](#) in March, and I am here today as a supporter of the initiative. Now boasting more than 100 organizations globally, representing all kinds of arts and creativity, it is THE global center of gravity advocating for the rights of creators in the age of AI.

This coalition believes AI is exciting and promising but can never replace human artistry and soul. It's based on seven core principles I support wholeheartedly.

- **First**, technology has long empowered human expression and AI will be no different. Musicians will use this latest technology to do great new things.
- **Second**, human-created works will remain essential in our lives. At the heart of the connection between artist and audience are shared, lived experiences only humans can relate to and convey.
- **Third**, the use of copyrighted works for AI purposes – and the use of voices and likenesses of professional performers – requires permission. Like all predecessor technologies, AI must be subject to authorization and free market licensing from all rightsholders and creators and copyright owners must retain exclusive control over how their work is used. If AI is trained on copyrighted works, it must be approved and licensed. Professional performers', actors', and athletes' voices and likenesses can only be used with their consent and fair market compensation.
- **Fourth**, governments should not create new copyright or other IP exemptions that allow AI developers to exploit creators without permission or compensation. People looking to make a fast buck from technological change routinely ask policymakers to pick winners and losers. Here, that would be devastating; creating shortcuts for AI will only erode the incentives to create new works – the works AI itself depends on.
- **Fifth**, copyright should only protect the unique value of human intellectual creativity. The Copyright clause of the Constitution exists to incentivize humans to create – machines don't need incentives.
- **Sixth**, trustworthiness and transparency are essential to the success of AI and protection of creators. Without transparent AI, we will have no idea whether the inputs AI systems were trained on were licensed, leaving us no way to enforce our rights.
- **Seventh**, creators must have a seat at the table, not just developers. Our creativity, our rights, our livelihoods are at stake.

If AI is allowed to take away the ability of artists and authors to control and make a living from their art, we will lose all authenticity in our expression; we will lose culture itself. The next decisions by courts and Congress in this area will decide our cultural future. It's your responsibility to make sure the Constitutional promise of reward for human genius remains viable.

Guided by the principles of the Human Artistry Campaign, we can look forward to the real emotions, experiences, and dreams of future generations of creators – perhaps facilitated by AI, but never silenced by it.

Thank you. I look forward to answering any questions you have.

Mr. ISSA. Thank you.
Mr. Sedlik?

STATEMENT OF JEFFREY SEDLIK

Mr. SEDLIK. Chair Issa, Ranking Member Johnson, distinguished Members, thank you for the opportunity to testify here today.

I've been a professional photographer and film maker for 37 years. I'm a Professor at the Art Center College of Design, the former President of the American Photographic Artists, and the current President of the PLUS Coalition.

As mentioned by Chair Issa, PLUS is a global nonprofit organization focused exclusively on simplifying the identification of visual works. PLUS is currently developing a global nonprofit visual registry and in cooperation with the IPTC updating our widely adopted metadata standards to allow artists to declare AI-related permissions and prohibitions in their image files. You can learn more about that a *plus.org*.

As a professional visual artist, I make my living by creating, and most importantly licensing my works. My ability to create new works, sustain my business, and support my family, depends directly on my exclusive rights to reproduce, distribute, publicly display, and adapt my original works. These fundamental rights are the core of my business providing a strong incentive to create new works. In fact, they're the only way that I can afford to create new works.

Unfortunately, many AI developers and platforms have built their businesses by exploiting billions of visual works without authorization from or compensation to copyright owners. The theft and exploitation visual works by AI platforms displace a common long-standing practice, in which, creators and their agents offer and sell artist reference licenses permitting the use of works for reference in creating new derivative works.

AI ingestion falls squarely within this reference license category. In fact, stock photo agencies routinely sell reference licenses to AI platforms permitting the use of visual works for AI ingestion. A market clearly exists for these licenses. The unlicensed ingestion of photographs by AI systems usurps that market and forces human creators to compete with machine-made derivatives of their own original works.

Many AI platforms are trained on copies of creative works scraped from websites that display those works without the knowledge or permission of the copyright owners. I've found thousands of unlicensed copies of my works in open data bases used by AI systems to support image ingestion and generation. It is clear that many generative AI platforms were founded on copyright infringement.

In defense of their actions AI developers attempt to apply blanket clearings of fair use, counter to the spirit and letter of the Copyright Act. Fair use is not a right. It's an affirmative defense requiring a fact-specific inquiry on a case-by-case basis.

AI developers claim that AI-generated works are not substantially similar to source works and thus can't be infringements, but this ignores the fact that the exclusive reproduction right is a stand-alone right under Section 106 of the Copyright Act. The cre-

ation and use of copies for AI ingestion purposes is copyright infringement on a massive scale.

AI developers further suggest that a photographer's use of cameras is the creative equivalent of drafting prompts for generative AI. They minimize the creative process in photography and attempt to frame photographers as mere button pushers. This is a false equivalency.

We photographers are visual storytellers. Our creative decisions are guided by our life history and our unique combination of training, experience, personality, aesthetic sensibilities, dreams, memories, research, and other factors. When creating our works, we anticipate and respond dynamically to the subject matter and shooting environment. We exercise control over the visual rendition of the scene. We decide which elements to include and exclude and where to place those elements within the frame. We determine how to juxtapose people, objects, and other compositional elements for a desired creative effect. We control the placement and interplay of color, tone, texture, contrast, light, and shade. We control the perspective, distortion, depth of field, and selective focus to guide the viewer's eye through the image. We select the precise moment at which to create the ultimate photograph.

This substantial creative human expression is not the equivalent of submitting a text prompt instructing a machine to generate a work.

Copyright law affords protection only to human expression. The output of an image by machines in response to prompts is and should remain ineligible for copyright protection. Policies must not favor machines over human creators. AI developers must be required not only to obtain advance permission to ingest and exploit creative works, but to compensate creators whether directly or through collective licensing organization such as the American Society for Collective Rights Licensing.

AI technologies must be developed and used in a manner that is responsible, respectful, and ethical upholding the underlying goals and purposes of our copyright system.

I appreciate the opportunity to testify, and I look forward to answering your question.

[The prepared statement of Mr. Sedlik follows:]

Committee on the Judiciary
Subcommittee on Courts, Intellectual Property, and the Internet
U.S. House of Representatives

Hearing:
Artificial Intelligence and Intellectual Property: Part I
- Interoperability of AI and Copyright Law -
May 17, 2023

Testimony of Professor Jeffrey Sedlik
Sedlik Photography
PLUS Coalition
American Photographic Artists

145 North Sierra Madre Blvd, Suite 4
Pasadena, California 91107
626 405-8100
info@PLUS.org

Chairman Issa, Ranking Member Johnson, and other Members of the Subcommittee on Courts, Intellectual Property, and the Internet, thank you for the opportunity to testify today on Artificial Intelligence and Intellectual Property.

My name is Jeffrey Sedlik. I have been a professional advertising photographer and filmmaker for 37 years, and am the president of the PLUS Coalition at www.PLUS.org, a global non-profit organization in which diverse stakeholders from 140 countries are collaborating to develop a global registry for the identification of visual works. I previously served as president of the American Photographic Artists.

I am the creator and owner of hundreds of thousands of copyrighted photographs. I make my living by creating and licensing my photographs to appear in all manner of media, for all manner of purposes, by all manner of clients. Like all visual artists, my ability to create new works, sustain my business, and support my family depends directly on my exclusive rights to reproduce, distribute, publicly display, and adapt my original works. These fundamental rights, guaranteed to photographers since President Lincoln authorized the extension of copyright protection to photographs in 1865, are the core of my business, providing a strong incentive to create new works, just as envisioned by the Framers in Article 1, Section 8, Clause 8 of the Constitution.

We photographers have never shied away from technology. In fact, we have a 197-year history of embracing technology as a vehicle for visual storytelling. We tell our stories with original, human expression.

Professional photographers don't "snap" photographs, nor are we mere button pushers. In preparing to create an image, a professional photographer determines the optimal vantage point, carefully selecting foreground, background, natural lighting, artificial lighting, shapes, textures, perspectives, storytelling potential, and numerous other factors. Our creative decisions are guided by our life history, and our unique combination of training, experience, personality, aesthetic sensibility, memories, research, and other factors. We anticipate and respond dynamically to the subject matter and shooting environment; we exercise control over the visual rendition of the scene; we decide which elements to include or exclude, and where to place those elements within the frame; we determine how to juxtapose people, objects, and other compositional elements for desired effect; we control the placement and interplay of color, tone, contrast, light, and shade; we control perspective, distortion, depth of field, and selective focus to guide the viewer's eye through the photograph; and we select the precise moment at which to create the ultimate photograph.

For some photographers, these creative decisions may occur over a period of days, weeks, or months. For others, such as photojournalists, thousands of creative decisions are by necessity compressed into minutes, even seconds. Irrespective of the time dedicated, photographers in all genres exercise original human expression, subjectively selecting, coordinating and arranging the elements of our creative works.

Copyright law affords protection to original human expression. While creators can and do employ AI technologies as a starting point for original human expression, and while that human expression should be eligible for copyright protection, the output of an image by a generative AI platform in response to prompts, in the manner employed today, is not original human expression, and should remain ineligible for copyright protection.

Most AI developers and platforms build their businesses by exploiting billions of creators' works without authorization from the copyright owners. Many AI developers claim that ingestion of copyrighted works is fair use. Fair use is an affirmative defense, not a right. A determination of fair use is a fact specific inquiry that must be made on a case-by-case basis. The reproduction right under Section 106 of the Copyright Act is a standalone right. The creation and use of copies of protected works for AI ingestion and generation purposes is copyright infringement, on a massive scale.

AI developers must be required not only to obtain advance permission to ingest and exploit creative works, but to also compensate creators. Policies must not favor machines over human creators. All rights of human creators and copyright owners must be respected. Available compensation schemes include direct compensation to artists, compensation through licensed agents, and collective licensing schemes managed by copyright management organizations, such as the American Society for Collective Rights Licensing.

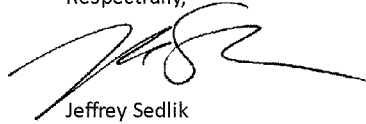
In the photography industry, we have a long-standing tradition of offering and selling licenses for "artist reference." These licenses typically permit the use of a photograph by an artist as a visual reference for the creation of one or more derivative works. The ingestion of a photograph by an AI system for the purpose of reference in the output of machine generated images is a form of reference of the expression in the photograph. In fact, stock photography agencies now routinely sell licenses to AI platforms for the purpose of AI ingestion and generative AI use. Photographers increasingly offer similar licenses. Clearly, a market exists for these licenses. The unlicensed ingestion of photographs by AI systems usurps that market.

The high-quality works created by photographers and other visual artists are ideal for ingestion by AI machines to generate output. AI developers use these works because of their significant expressive value. Most of these works are scraped from websites without the knowledge or permission of the copyright owners. I have identified thousands of unlicensed copies of my works in the databases used by AI systems to support image generation. My peers report similar findings. It is clear that many AI platforms are trained on infringing copies of creative works, and that generative AI is a technology founded on copyright infringement.

It is essential that AI technologies be developed and used in a manner that is responsible, respectful, and ethical – and in a way that upholds the underlying goals and purposes of our copyright system.

Thank you for your efforts to ensure that creative works are protected, and for the opportunity to share my experience and perspective in my testimony today.

Respectfully,

A handwritten signature in black ink, appearing to read 'JS', with a long horizontal flourish extending to the right.

Jeffrey Sedlik
Sedlik Photography
PLUS Coalition
American Photographic Artists

Mr. ISSA. Thank you. I thank all our witnesses.

I am going to forego my questioning until the end or near the end, and so I am going to go to Mr. Fitzgerald, somebody who understands what it is to own copyright in his former life. The gentleman is recognized for five minutes.

Mr. FITZGERALD. Thank you, Mr. Chair.

Mr. Damle, the Supreme Court has held since an 1884 case, *Borrow-Giles Lithographic Co. v. Sarony*, that photographs can be entitled to copyright protection where the photographer makes decisions regarding creative elements of the work.

You were quoted in your testimony as saying,

... society embraced the camera as a creative tool and photography blossomed as an art, as an art form that deepened rather than diminished this whole field of creativity.

There is no reason to believe AI is any different, I think was your point. The Copyright Office has disputed this comparison, instead comparing the generative AI to a client to hires an artist to create something because users do not exercise ultimate creative control.

What is kind of your response or your thoughts to copyright comparison and do you believe that Copyright was correct in denying copyright for lack of the ultimate creative control as they did in their decision in “Zarya of the Dawn”?

Mr. DAMLE. Congressman, thank you. Thank you for that question. So, the way I would start by framing this issue is the point I was trying to make in my testimony is that generative AI can be a tool that humans use to enhance their creative output. So, the Copyright—what the Copyright Office has said is where that tool is doing all of the work of creative output, then that’s a situation where we don’t need the economic incentive that the Constitution has an incentive for creating that output.

I think that’s sort of one end of the spectrum of the question. There’s going to be a big gray area where there’s going to be basically human and AI together creating output. I think that’s going to be a very common situation going forward. In that situation, I would say where the human is exercising some control over the AI and its output, where there is a sort of iterative process between what the generative AI produces and what the human produces, then that is a situation where you should have copyrightable output, that the output of that process should be copyrightable to some degree.

It’s no different I think than a camera, than a photographer’s interaction with a camera. They’re adjusting the settings on the camera. They’re choosing the framing of the image. They’re employing a lot of choice. Yes, the camera is the one that’s actually recording the image, but the human has control over that process. So, by the same token where a human has that level of control over the generative AI process, as I think will happen in many cases, then the output of that should be copyrightable in the same way that a photograph is copyrightable.

Mr. FITZGERALD. So, let me just reiterate that. Do you think Copyright got this correct in the way that they had come up with their determination?

Mr. DAMLE. I think in the particular instance and the particular things that they said, I think, they got it right, but they’re address-

ing a very just sort of extreme example where the human is not actually providing anything other than a simple prompt to the generative AI system. Then they decided in that instance the output is not copyrightable. I think that's going to be very different than the mine-run of cases using generative AI. The mine-run of cases are going to involve much more involvement of human authorship than just providing a simple one-sentence prompt.

Mr. FITZGERALD. Yes, so just because I just have 1 minute left actually, so what other situations do you think could emerge that Copyright would then have to deal with where you do find this mix of artistry along with what AI has capabilities of doing that could somehow put us in a place where it could be undetermined who actually is the creator?

Mr. DAMLE. So, those are going to have to be determined on a case-by-case basis. The office, actually, in its guidance provided some examples, a few examples, but some examples of where there would be sort of a creative—enough creative human authorship. For instance, where you provide a prompt, it generates an image, and then you adjust that image. The human actually takes that image as a starting point and then adds more creative authorship to that image. That's one example.

Another example that you might see is where you start with an image that you have—I'm using an image, ones as an example, but you start with an image, you feed it into the generative AI, and the generative AI helps you make changes to that image in certain ways, in the same way that you might use Photoshop. In those circumstances you're going to have situations where the output is copyrightable. Then there may be questions about how much of it is copyrightable, how much of it was generated by the AI, but those can be dealt with on a fact-by-fact basis.

Mr. FITZGERALD. Thank you very much, and I would yield back.

Mr. ISSA. I thank the gentleman.

We now recognize the Ranking Member of the Subcommittee, Mr. Johnson, for five minutes.

Mr. JOHNSON of Georgia. Thank you, Mr. Chair.

The other day I was in my car cruising and there was an interview with Smokey Robinson on the radio, and Smokey Robinson was saying that his big hit "Cruisin" took in total about five years to come to the final product.

I would like to ask you, Mr. Irwin, Mr. Navarro, and Mr. Sedlik—your works can be scraped from the Internet in a matter of seconds for AI ingestion. Can you explain what goes into creating a single work of art and how long that process might take? Starting with you, Mr. Irwin.

Mr. IRWIN. Sure. The answer is how long is a piece of string, to be honest. It's always different. When we're working in audio-visual space, be it film, television or something like that we're commissioned to work, you have deadlines, and you have to come up with the goods by a certain time and deliver them.

When working as a songwriter more like Dan does, you have a little more freedom and you can work on a lyric, you can work on a song, you can collaborate, you can have a half-written song, and then bring someone else in. Everything takes time. Nothing is the

push of a button. There is no push-button music up to this point, that's any good anyway. Most of it is just not usable.

Mr. JOHNSON of Georgia. Mr. Navarro? Thank you.

Mr. NAVARRO. I'd like to affirm what Mr. Irwin says. I made reference to my big hit song that took 90 minutes to write. The emotions that led to my contributions and the contributions of my late partner took three years to percolate, to run through my emotions and my system be expressed, be wept, be verbalized, and be internalized before they came out in an evening blast when everything was ready, it didn't take 90 minutes. It took several years. I've had songs take seven years; I've had songs take seven days. It varies depending on what it takes to sit and look at something and go it's done. Now, it's done.

Mr. JOHNSON of Georgia. Thank you.

Mr. Sedlik?

Mr. SEDLIK. Thank you. There are many genres of photographers, each of whom works in a different way, including within the genres. Personally, I'm very controlling in my photography. Every aspect of every photograph that I take is something that is planned and controlled, whereas another photographer might be more spontaneous.

In my process I use free association first. It will take weeks, months, even more than a year. I use free association, come up with ideas, make thumbnail sketches, make iterative sketches, do tests, do planning, set the whole thing up, test, and then during the process I control it, whereas a photojournalist might make all those decisions in an instant. It's their lifetime of experience and all their capabilities that lets them accomplish the same thing, thousands of decisions in a second.

Mr. JOHNSON of Georgia. Let me ask you this, Mr. Sedlik: Can you compare the process of creating works, your works through—or can you compare that to works created through generative AI?

Mr. SEDLIK. Yes. So, the same goes for AI. There's the full spectrum—

Mr. JOHNSON of Georgia. If you could tell me also how it feels as an artist to have your work used without your permission to serve as the basis for a product that might then compete against your product.

Mr. SEDLIK. Yes.

Mr. JOHNSON of Georgia. If you could leave some time for Mr. Navarro and Mr. Irwin to respond to that question also. Fifty-eight seconds. Go ahead.

Mr. SEDLIK. OK. So, I'll answer your second question first. So, photographers are used to the advancement of technology. We've been using technology for 179 years and we're usually first in using it. So, we anticipate that technology is going to continue to develop, that new opportunities for creation are going advance, that it's going to be easier to create great works, and we accept that. That's not our concern here. Our concern is that our works are being used without our permission and without any compensation. The creative process—

Mr. JOHNSON of Georgia. Thank you. Let me go to Mr. Navarro. Appreciate it.

Mr. NAVARRO. I feel the same way. I think AI is a tool that I, myself, could use. My partner in my duo passed away. I could maybe impersonate his voice and put out a brand-new loan and borrow record. Permission, credit compensation.

Mr. JOHNSON of Georgia. Thank you for that.

Mr. Irwin?

Mr. IRWIN. We've been using elements of AI as tools for—I've certainly been using them for almost 30 years, but always with the control of what the output is. Quite often, create the initial piece of music that you're going to work with. Then, you process it in a particular way, and you make the decisions of what is the final product going to be. Some of the options you are given are not usable. It's a matter of taste, and only humans have that kind of discerning taste.

Mr. JOHNSON of Georgia. Thank you.

Mr. Chair, I appreciate the indulgence.

Mr. Callison-Burch, if I had time, I would be asking you about the value of work that you would see as so critical for us to think about.

Thank you.

Mr. ISSA. I thank the gentleman.

We will make sure we get that answer.

Mr. ISSA. The gentleman from Oregon, Mr. Bentz.

Mr. BENTZ. Thank you, Mr. Chair.

I thank all of you for being here today.

So, my question is—I'm not sure which one of you to ask—but the issue is whether our technology is such that we can actually do what I think many of you want us to. I'm just thinking of the Led Zeppelin situation that took years to sort out whether or not, quote, "the descending chord sequence had been used for centuries," and whether or not it had been stolen. Of course, although it was decided after five, six, seven, or eight years of litigation, that Led Zeppelin had, for various reasons, not usurped it.

My question to you—and I'm going to ask it of you, Mr. Navarro—if you think that we have the technology available to sort these things out after the fact. If not, what would you be suggesting that we do to try to prevent it before it happens? I just want to know, given the nature of music—and I'm a very bad musician, but I know what a guitar is—how in the world do you sort this out? Do we have the technology available?

Mr. NAVARRO. I'm not sure we're qualified to determine how it's sorted out. I know that I, as a creator, if I get too close to something, I pull back. If I get a little too close to Ray Charles, I pull back. If I get too close to John Lennon, I pull back. If I don't do it right, there are legal remedies. Just ask George Harrison or Robin Thicke.

Mr. BENTZ. Mr. Damle, I'm not sure I pronounced your name correctly, but could you address the question?

Mr. DAMLE. Thank you, Congressman. Yes, you got it exactly right.

So, I would point to the sort of cases that you're talking about, like the Led Zeppelin case, like the Blurred Lines case, as example of instances where copyright did not stay within its proper bounds,

where courts decided that borrowing somebody's musical style counted as copyright infringement.

I think when those decisions came down, it really threw the music industry into chaos. I'm a music lawyer, in addition to being a technology lawyer. I know that the music industry really grappled with those decisions. You had artists saying that they were afraid to create because they were worried that they were, inadvertently perhaps, borrowing the style from somebody else. It got to the point that even the rights owner groups started criticizing those decisions, saying they went too far; they did not allow artists to express themselves as freely as they should be able to.

Thankfully, I think we've seen a return back to those core principles of the Copyright Act, which is ideas/styles should be able to be used by all. So, that's what we've seen in these recent decisions involving Ed Sheeran—a return back to that principle of being faithful to the Copyright Act, faithful to the Constitution's mandate for what the Copyright Act is supposed to do.

So, that would be my response to that issue.

Mr. BENTZ. Thank you for your answers.

With that, Mr. Chair, I yield the rest of my time to you.

Mr. ISSA. I thank the gentleman for yielding.

I'll continue along somewhat an earlier line. Mr. Navarro, we talked about credit; we talked about permission, and we talked about compensation—or you did.

Credit would seem to be one that Congress could mandate that the data base input could be searchable. So, you would know that your work or your name, or something, was in the data base. Fairly easy, probably doable. It's a credit index of billions, or trillions in some cases, but that would be—the output might be a little more complex, and we'll talk about that later.

The compensation is a question I want you to opine on, and maybe both ends of the question. If there are 10 billion or 10 trillion inputs—and let's just assume for a moment that there is a billion copyrighted. We all know what it's like to get that big check from Spotify for the hundred or a thousand times you were played, and it comes out in pennies. OK. What is the division of a billion pieces of music, and how would it, in fact, assuming that this was part of the output, how would you actually quantify it? Because we have to put a number on it at some point.

Mr. NAVARRO. It's a difficult thing to do. I believe in free-market negotiations with regards to this. I know that can be cumbersome. I don't believe in compulsory licenses, especially as regards to this particular issue.

The compulsory licenses I'm used to, whether it's at SiriusXM or in the use of a song that I've already recorded, which is a compulsory license for someone else to use, benefits me directly. When my stuff is part of a large number of stuff that's scraped, it supersedes my work. It doesn't even just compete with it. That's a footrace. It supersedes it. In that context, I believe in free market.

Mr. ISSA. We will, undoubtedly, be asking that question of others in another forum.

With that, we go to the Chair of the Full Committee, Mr. Nadler.

Mr. NADLER. Thank you, Mr. Chair.

Mr. Irwin, according to reports, OpenAI, the company behind ChatGPT, is valued at \$29 billion. Stability AI, the company behind Stable Diffusion, was valued at around \$1 billion late last year. For an area of comparison, can you give us an idea of how much the average composer's salary would be?

Mr. IRWIN. That's very difficult to say. Because the way composers—first, we are not governed by any kind of collective bargaining. We're not a union or anything like that. So, we're very much independent contractors.

It's easier to talk about ranges for television shows. An hour of television, a feature film, those sort of budgets are easier to quantify, and then, of course, it comes into the experience of whether you're an entry-level person or a very experienced, high-level person. All of those change the rates.

Then, of course, on the back end, where we get our royalty streams, which is where a lot of us make our money, the performing royalties that are collected by the societies, like ASCAP, BMI, and SESAC, they are dependent completely on the number of performances of those works that include your music. That's the same for audiovisual and for streaming, and everything.

So, it's very hard to say. Some years, it's like—it's almost like a farmer. Some years you have a bumper crop; the next year, you might not see much at all.

Mr. NADLER. I see.

Mr. Sedlik, what about the average photographer? The same? What about the average photographer? The average salary?

Mr. SEDLIK. I don't have statistics on that, Representative Nadler. However, many photographers make as little as \$20,000–\$40,000 a year, and amounts greater than that outside of certain spaces can be unusual.

Mr. NADLER. OK. Thank you.

Mr. Irwin, can you talk a little more about what threats generative AI poses to composers and songwriters? What actions do you think we should take to ensure that your work and artistry is protected? Actually, you talked about the threats. Can you talk about—

Mr. IRWIN. Yes, I can talk—do you want to talk about the threat, or do you want to talk about the solution?

Mr. NADLER. I think you've talked about the threats. What—

Mr. IRWIN. Yes. Some of it's in my written testimony.

One of the things that the music industry has done, particularly, well in this area is collective licensing. They have, as I just mentioned, for ASCAP, BMI, SESAC, they have a way to monitor performances. For mechanical royalties, which are the physical royalties, you have SoundExchange. Then, there's streaming royalties as well through the Music Licensing Collective.

There are reciprocal organizations set up all over the world.

Mr. NADLER. Well, how would we apply that to OpenAI?

Mr. IRWIN. Well, music is probably more easily applied to music than it possibly is to some of these other art forms. Because every piece of music that's registered has a registration number, a work number, and a recording number. Every artist has their own number. I have what's called an IPI number; Dan has one. So, they're already in the system. You don't need to reinvent them. You just

need to make sure they are attributed to those works as they're being used, or logged, or however you want to do it.

Mr. NADLER. How do you determine what works are used in generative AI?

Mr. IRWIN. Well, that's something that will need to be determined. I mean, I write music. I'm not a technologist.

Mr. NADLER. OK. Maybe, Professor Callison-Burch, transparency in AI training models is a concern, but the transparency to the end user that the media they are viewing was AI-created is critical as well.

I see that in your testimony you note that parts of it were composed using ChatGPT. Should some sort of disclosure like that be required? Currently, how are individuals alerted to the fact that they're hearing or viewing an AI-created work, if at all?

Mr. CALLISON-BURCH. Thank you for the question.

So, disclosure of AI-generated works I think is valuable. I think, especially with our potential for societal harms through generating deepfakes or works that could be used to influence elections by mocking up instances like Trump being arrested in New York, Assad being generated by Midjourney. Certainly, a disclosure—

Mr. NADLER. Or generating a fake speech by me?

Mr. CALLISON-BURCH. I think that this is an important issue that touches on output of generative AI systems, and that is where I think that regulation is deserved.

Mr. NADLER. How are individuals alerted to the fact that they're hearing or viewing an AI-created work, if at all?

Mr. CALLISON-BURCH. I'm sorry, repeat the question, please?

Mr. NADLER. How are, currently, how are individuals alerted to the fact they are hearing or viewing an AI-created work, if at all?

Mr. CALLISON-BURCH. There is a variety of technological devices that our field is innovating to mark AI-generated works, similar to a watermark on a stock photography site. This is not an established industrywide practice, but it is something that our field has been discussing. At the moment, it's up to the user of the AI system who's generating it to disclose to people who they're transmitting that image to that it was generated.

Mr. NADLER. Thank you. My time is well expired, and I thank the Chair for his indulgence.

Mr. ISSA. Mr. Chair, you gave me the similar indulgence when you sat in this Chair. So, I'm only returning the favor. I thank you.

We now go to the gentleman from Virginia, Mr. Cline.

Mr. CLINE. Thank you, Mr. Chair.

I want to followup on that line of questioning about existing technologies that are there to identify digital works and enhance copyright protection.

Mr. SEDLIK, digital watermarks, tags, metadata, what challenges are you aware of in the use of these types of technologies for works that are used in training AI?

Mr. SEDLIK. We have mature, very capable technologies to identify visual works, including image recognition and what's called steganography, which is burying signals in the image to identify it.

Creators use embedded metadata to pass information into their images, so that, as their images are distributed, their images can be identified. A big problem for us is that all that information is

stripped out by the social media platforms and other platforms when images are used. We would very much benefit from a change, an improvement to the law to make it illegal to remove embedded rights metadata, even if it's not done for the purpose of infringement. Right now, under Section 1202, it's only illegal if it's done intentionally for the purpose of inducing, enabling, concealing, or facilitating infringement.

Mr. CLINE. Mr. Irwin, what about music?

Mr. IRWIN. Sorry?

Mr. CLINE. Do you have any comments on that?

Mr. IRWIN. Yes. There is watermarking and fingerprinting used in music technology over and above the identifies that I talked about. There's technology that allows, currently allows, music to be identified even within a program that has dialog and sound effects over the top of it. You can still identify that music.

So, the technology is there for this to be done. It's just a matter of having the will and sitting with the organizations who are doing it and getting this discussion going. I have no doubt that there's a way to track this stuff directly.

Mr. CLINE. Professor, you've commented on that and talked about the trending toward use of this technology by industry. Is that something that's happening too slowly? Is there something that needs involvement from government? What do you think?

Mr. CALLISON-BURCH. So, I think at the core of the problem with an idea like a compulsory license—and Mr. Issa's suggestion of what is the value of one over a trillion, when you think about the volume of this work—and I can say that, definitively, that the value of one over a trillion is going to be vanishingly small.

So, another key that is a practical consideration here is, unlike compulsory licenses, which are based on performance—so, the MMA, where Spotify plays music, Taylor Swift gets more money than some random person in the catalog because her songs are played more—there's no equivalent here for generative AI. It's hard, it's impossible to understand how much of a system's output is due to Stephen King versus a random Reddit poster who's written a paragraph in the collection.

So, I think that the lack of that performance is key to one of the tricky things about establishing a compulsory license here.

Mr. CLINE. Mr. Irwin, some have suggested that imposing IP-related obligations on the AI developers would hobble development because of the inability to feasibly use copyrighted content for training purposes. Would you be willing to license your works for training purposes? Do you think other creators would also be willing?

Mr. IRWIN. If I was being compensated for them, absolutely. Yes. I don't have a problem with the technology at all. I have a problem with the stealing of the material.

Mr. CLINE. Mr. Navarro, you indicated your willingness to use AI and—

Mr. NAVARRO. A similar answer. The ability to approve or disapprove of a particular use is why I oppose compulsory licenses. I don't oppose blanket licenses, which might make—streamline a process. I still have the ability to say yes or no and opt out.

Mr. CLINE. Now, in your testimony, you advocated against granting AI any special IP exemptions. Depending on how some lawsuits turn out, courts may decide that AI doesn't need special exemptions in training AI, that copyrighted music is not infringement under current law. In that event, would you advocate for changing the law to make training AI with copyrighted works a type of infringement?

Mr. NAVARRO. I'm not sure that I am qualified to answer that. It's a very technical question. I do believe that, as we're looking at guardrails, guardrails used to be made of wood. Then, they started becoming made of steel. Now, as cars get faster and more powerful, maybe they need to be made of titanium. As these technologies progress, and as its scope increases, our guardrails need to be modified and improved.

Mr. CLINE. Thank you. I yield back.

Mr. ISSA. The guardrails of the future will be software-driven.

With that, we go to the gentleman from California, my colleague, Mr. Lieu.

Mr. LIEU. Thank you, Chair Issa and Ranking Member Johnson, for holding this important hearing.

As a recovering computer science major, I am enthralled with AI, and I believe it has, and will continue to, revolutionize society. It can also cause us harm and it creates all sorts of unanswered questions.

So, I'd like to walk through some examples, so that the American public and I can better understand how artificial intelligence interacts with copyright.

I'd like to ask Mr. Damle this example. Let's say I create a generative AI Internet application related to music. I do it for commercial purposes. To train it, I scrape the entire Internet of all songs, including every one of Taylor Swift's copyrighted songs without her permission. Your view is I wouldn't have to compensate her in any way, is that right?

Mr. DAMLE. So, I think it would very much depend on the particular way, in which, you trained the AI models. Not all AI models are constructed the same way. Some are constructed in ways that might very well exceed the bounds of fair use.

So, it's inevitably going to be a fact-by-fact—a case-by-case determination of whether a particular model is going to exceed those bounds or not.

Mr. LIEU. Let's just use ChatGPT's model.

Mr. DAMLE. So, in an instance where what you have done is extract unprotectable facts from any copyrighted work—so, stepping aside and generalizing this point to any kind of work—if what you've done is extract unprotectable facts from those works, and then, used those facts to generate a new work, then, under well-established principles of copyright law, that is not infringement.

Mr. LIEU. I'm not talking about generating any, just about training the model.

Mr. DAMLE. Just on the training side, if that's all that happens, then, under a long line of cases that I've laid out in my written testimony, that is fair use—

Mr. LIEU. To train a model, you need to actually download the Taylor Swift songs?

Mr. DAMLE. That's correct. That's correct.

Mr. LIEU. You view that as fair use?

Mr. DAMLE. That would be fair use. The premise of any fair use case is going to be—

Mr. LIEU. So, Internet applications, like YouTube, pay a licensing fee to Taylor Swift when they download her songs. What is the difference?

Mr. DAMLE. Well, the difference would be in those instances, what they're doing is they're taking the work, and then, they're taking that work and they're streaming it to end users. So, that's a—that's a public performance of her work.

Mr. LIEU. So, let's say ChatGPT lets you just put out Taylor Swift lyrics. What's the difference?

Mr. DAMLE. I think that might be an issue that exceeds the bounds of fair use, where you're taking—

Mr. LIEU. Now, let's say I take my model and I generate a new song similar to a Taylor Swift song in terms of lyrics. Is that a copyright infringement?

Mr. DAMLE. That's correct, Congressman.

Mr. LIEU. It would be infringement?

Mr. DAMLE. It would not be infringement.

Mr. LIEU. It would not be infringement.

Mr. DAMLE. Because one of the basic principles of the Copyright Act, which I discussed earlier, is replicating somebody's style, writing a song—if I were to write a song in the style of Taylor Swift, I would not be committing copyright infringement. That's one of the founding precepts of copyright.

Mr. LIEU. Now, as you know, voice cannot be copyrighted. Let's say my model also generates voice very similar to Taylor Swift, and I had this brand-new song similar to Taylor Swift's voice, similar to her lyrics. You believe that would not be copyright infringement?

Mr. DAMLE. You would have to look at other bodies of law to determine whether that would be illegal. As a matter of copyright law, and just the basic principles of copyright law, that would not be copyright infringement, which is not to say it's not concerning for other reasons. In just looking at the copyright law itself, that's not—

Mr. LIEU. Then, finally, in fact, you believe I could, then, copyright this Taylor Swift-like song with voice like Taylor Swift that I generated by scraping the Internet with Taylor Swift copyrighted songs I didn't pay her for, right?

Mr. DAMLE. I don't know that would necessarily be the case. Certainly, as we were discussing earlier, the Copyright Office has taken the view that AI-generated works like that may not be subject to copyright protection at all.

Mr. LIEU. OK. Thank you.

So, my remaining comment and question for Professor Callison-Burch about disclosure—and Congressman Nadler asked about this—many creators already use AI, right, in their creative works? I mean, there's a whole bunch of algorithms that make your song sound better. They don't disclose that, right? How would you even define what kinds of AI they need to disclose that help them with their particular creative work?

Mr. CALLISON-BURCH. That's a great question. If you'll indulge me, I want to answer your question to Sy as well.

So, your question about if I download Taylor Swift's songs and I learn from it, it could be equally well posed to a teenaged pop star who's learning how to sing. So, if that person learned from Taylor Swift, the decision of whether or not they're violating copyright is not at the time when they're listening to the songs and learning to perform music. It's when they release an album and whether that album is sufficiently similar to, say, Taylor Swift's songs.

So, if we release, instead of Taylor's version, we release ChatGPT's version of an album, that's infringement, but the learning from it is not.

In terms of disclosure of use of copyrighted, of generative AI in materials, I think there's an interesting ill-advised guidance from the Copyright Office that works that involve substantive use of generative AI are not copyrightable at the moment. I believe that AI is going to be used in a collaborative way with humans, and the human using it deserves that copyright. Whether or not they acknowledge it, I think depends on the use of it.

So, there was a law passed that political ads must disclose if they're using generative AI to create images of politicians. That seems like a very valid case to disclose. If I'm creating a comic book, it seems less high stakes, so probably not necessary.

Mr. LIEU. Thank you. I yield back.

Mr. ISSA. I thank the gentleman.

We now go to the gentleman from Texas, Mr. Gooden.

Mr. GOODEN. Thank you, Mr. Chair.

That was very interesting. I'd like to maybe keep going along those lines.

It seems as if, as this develops, we'd like to perhaps see more transparency in the process. Is it unrealistic to think that the songwriters, the people, the American people, whoever, could kind of see what goes into these sources of what has an influence on the AI. Mr. Callison-Burch, I'll give it to you. I don't think the average person, myself included, understands the technology. Is it unrealistic to ask that we know if a particular song got more influence from Taylor Swift, or whoever, and how that process comes about?

Mr. CALLISON-BURCH. So, I think this is a super-interesting question that, again, involves the output of a generative AI system, rather than the training per se. I think that there is a valid case to be made that copyright should be reshaped to protect against a case where I, as a user of an AI system, ask it to generate something that mimics a particular artist.

Like this concern that artists have, I think is 100 percent valid, that you can currently say, "Generate a comic strip in the style of Sarah Anderson," and it produces something similar in style, but does not reproduce any of her published works. That, to me, seems more like a right of publicity style concern than current copyright law is addressing.

I do think it's an ethical issue that we should consider as an industry, and I think that there should be an opt-out mechanism for artists to explicitly exclude their work from the vast amount of training data that we have. Again, I want to make this distinction

clear. What AI systems are learning from their training data is more akin to facts and patterns and statistical correlations than it is memorization or directly lifting from copyrighted works.

Mr. GOODEN. Mr. Damle, please forgive me if I've mispronounced your last name.

Do you believe that there will be more of a push to actually copyright some of these AI works? I feel like that will be controversial and a tough sell for those on the other side of your table. Could you explain the thought process behind that?

Mr. DAMLE. I think there's going to a lot of hard questions that get raised. Just as there are many ways in which to train an AI model, there are many, many ways in which to use an AI model in the creative process. I think we're really at the very, very early stages of trying to figure out where you draw the line between an AI-generated output that we don't think deserves copyright protection under sort of the constitutional precept of what copyright law is about and what creative output that is assisted with in AI deserves copyright protection, and it's necessarily going to be fact by fact.

I agree with the Copyright Office's view that the principle here is that you need sufficient human authorship, sufficient human input into the creative process, to warrant copyright protection. That's almost a constitutional requirement.

Exactly how that plays out in any given case is going to require over time, over the next few years and more, looking at every case that comes and trying to decide, OK, is this on one side of the line or the other?

Mr. GOODEN. OK. Thank you.

Maybe it's a far-fetched analogy, but many years ago, when I was in the statehouse, we passed a bill—I authored it—that restricted the use of drones over people's backyards. The drone industry was against that, and they said,

You should be able to fly a drone over anyone's house and park it and watch them all day. A helicopter can do that. So, what's the difference?

I said,

Well, a helicopter is operated by an individual. It can't stay up there forever, and you know if it's there because it's loud and big.

So, I thought of that as you were talking about your example of the middle school child who sings like Taylor Swift, compared to the AI, it sounds like Taylor Swift, and how there's no difference. The difference is that's a middle school child, and I'm not real worried about a middle school child taking over the music industry. So, I think we have to differentiate between the two.

I yield back.

Mr. ISSA. Would the gentleman yield?

Mr. GOODEN. This gentlemen?

Mr. ISSA. Yes, would you yield for a moment, please?

Mr. GOODEN. Oh, of course. Please.

Mr. ISSA. Thank you.

At some future time, I'd like you to talk about—and particularly, Mr. Navarro—about how you view the difference between a cover band and AI and where the guidelines are similar and when they would be different, and the same, obviously, from a legal stand-

point. I won't ask to have it answered at this time, where I'm out of time.

Mr. ISSA. I thank the gentleman for yielding.

We now go to the gentlelady from North Carolina, Ms. Ross.

Ms. ROSS. Thank you, Mr. Chair.

Thank you to all the witnesses for being here today.

We've heard multiple perspectives today on whether training AI on unlicensed copyrighted material should be permitted through the fair use doctrine. The Copyright Office considers several factors in evaluating this question of fair use.

One of those factors is the amount and how substantial the copyrighted work was that was used. The guidance notes that in some context using an even small amount of copyrighted work was determined to not be fair because the selection was an important part, or actually, the heart of the work. So, it could be just very small, but have such an imprint.

That strikes me as a key factor in the consideration of AI training as well. AI is built upon creative works, whether it's art or music or writing. AI would not exist at all if it couldn't learn from human beings. So, the work that AI learns from seems to me to constitute the heart of AI itself.

Now, the Copyright Office also considers whether the unlicensed use of copyrighted work would harm the existing or future market for the work, and we've heard about that. We've heard from the creators today about how AI is already doing that.

Mr. Altman has been with us this week, and I'm quoting him when he says,

When we're working on new models where, if an AI system is using your content, or if it's using your style, you get paid for that.

I hope that he's going to follow through on that because, as the Chair—or Ranking Member Nadler told us, Mr. Altman has a lot of money to pay you.

My first question is for Professor Callison-Burch. Is it a common industry practice to keep a careful record of how and whether copyrighted works, performances, and likenesses were used to develop or train an AI dataset?

Mr. CALLISON-BURCH. Thank you, Congresswoman.

Yes, it is. In fact, there was a *Washington Post* interactive feature published about two weeks ago sort of provocatively named “the secret lists of websites that make ChatGPT so smart.” That was produced in collaboration with researchers at AI2, where I'm currently taking my sabbatical, to exactly search for which websites were included in the training data of a very common training set, not necessarily the one that ChatGPT uses, but the one that many people in our industry do use.

Ms. ROSS. Just as a followup, do you think that would make it easy to devise a compensation system, since we have that trail of what's been used?

Mr. CALLISON-BURCH. That's a great question. So, I think the tricky part about creating a compensation scheme is, once again, there are a trillion words' worth of text in our training datasets. Each author represents a vanishingly small portion of that. We do not re-perform any of the songs or texts that are in our corpus. It's not a performance-based compensation scheme that would be pos-

sible. So, instead, it would have to be something to do with the volume of each person's contribution to that work. Again, I think if you do the math, it will end up being everyone gets a check for two cents, which doesn't make sense.

Ms. ROSS. We have similar things in the music industry, though. I want to get on to my—

Mr. CALLISON-BURCH. Those are orders of magnitude smaller.

Ms. ROSS. Yes, to my next question. This is for all three of our artists and creators.

Some AI developers have suggested that, if using copyrighted works is not deemed to be fair use, it would stop the development of AI. I imagine that most copyright owners would be willing to license their works to people who they would consent to license their works to, given the opportunity.

So, if a company wanted to license your work for AI, would you be willing to grant permission if you could agree on reasonable compensation? Go in whatever order you would like to go.

Mr. IRWIN. Absolutely, yes, of course. We're professional people. We get paid to do what we do. So, why wouldn't we want to license our work? The more work we get out there, the better it is for—it makes more incentive for everyone to create it, if there's compensation to create it.

Ms. ROSS. OK. Mr. Navarro and Mr.—

Mr. NAVARRO. The "if" is the big part of it. If we can agree, absolutely. If we can't agree, then I retain the right to withhold it.

Mr. SEDLIK. I agree, copyright protection is automatic. This can't be an opt-out and shouldn't require an opt-in. We are protected by copyright, and most visual artists are willing to license their works for various usages. I, personally, have offered my works for AI training licensing for years, and stock agencies increasingly do sell licenses for AI training.

Ms. ROSS. Thank you, Mr. Chair. I yield back.

Mr. ISSA. I thank the gentlelady.

We now go to my colleague from California, Mr. Kiley.

Mr. KILEY. Thank you, Mr. Chair, and thank you for putting together today's hearing, which is a very important and timely topic.

I have a very open mind on these matters. I'm extremely sympathetic to the concerns raised by the artists and the predicament that they're now in. At the same time, I understand the practical difficulties with some of the ideas that are being proposed, the potential impediments they might pose to innovation.

I'm also wondering whether copyright law is well-suited to the matter at hand in a lot of respects, as clearly the Copyright Clause and the body of law around it did not anticipate sort of a capacity for creation that is non-human in origin.

So, just to sort of start out, I want to try to analogize this to the process of human creation as best as possible. So, Mr. Callison-Burch has drawn a distinction with the training process and the output. You mentioned earlier how a child who's learning how to sing by studying Taylor Swift, there's no infringement there. It's the output that is going to be judged. All of the artists here, I'm sure, could cite people, artists who have been influences on their style, as they developed.

So, let's take another analogy. Let's say that I'm writing an essay on Abraham Lincoln and arguing how he ranks among U.S. Presidents. I go to the library. I check out every book I can possibly get on Abraham Lincoln. I learn all about him, and then, I reach an opinion and I write an essay about it. Of course, I'll cite to sources when I'm quoting them directly or for specific facts, but that learning process by which I formed my opinion, I'll—this is for you, Mr. Damle. Is there any copyright claim that is specific to that process?

Mr. DAMLE. I think that if, to just modify your hypothetical slightly, if you were to go to a library and make photocopies of books, right, that's a copy that you're making of those books. That would be copyright infringement if you didn't have a fair use defense. In your hypothetical, that would be very clearly within the scope of copyright infringement. You're making those copies not to—for the only purpose for which to learn the facts that are being conveyed in those copyrighted materials.

Then, yes, you're producing an output that might borrow those facts into a new work, but that work, as long as that work does not infringe the original work that you copied from, that's quintessential fair use.

Mr. KILEY. So, a court that would sort of study any copyright claim, would they look at my learning process or would they just sort of judge the output? Maybe they would if it's, like, about intentionality or damages claim or something. The core copyright claim, infringement claim itself, would they study my learning process, or would they just look at my work product?

Mr. DAMLE. I think it would be all the above. These copyright cases, having litigated a lot of these, really delve into a lot of the facts. So, the *Google Books* case, for example, looked at the way the manner, in which Google had acquired the copies, but, then, the outcome was driven mainly by the fact that all that Google was doing was providing a way for you to search within those books, and not providing you the whole book. It was providing just snippets of the book, and said, looking at that whole process, we're going to call the copying and retention of the copies for that end purpose to be fair use.

Mr. KILEY. Interesting. So, one of the things that Mr. Altman said is that he thinks we're moving toward a stage where, really, most of the training is based upon synthetic data that's generated by the AI itself. Maybe an analogy to this is sort of when DeepMind was training its goal-playing AI. At first, it studied the masters' games, then it, eventually just learned how to succeed by playing itself. So, does that impact the analysis?

This is for either of you, Mr. Damle or Mr. Callison-Burch. Do you have thoughts on that?

Mr. CALLISON-BURCH. This is a very interesting question. So, I've done a thought experiment regarding training AI systems on completely synthetic data. So, after having been a panelist on the U.S. Copyright Office listening tour, I thought about what about our training data do we need to learn the facts about the world that we care about, and learn the language patterns?

So, we need examples of language to learn the structure and grammar of language. We need some anchor into how people dis-

cuss the world to learn facts about it. There's no obligation that this be human-written text now.

I think that, in my little armchair experiment, we may have reached a copyright escape velocity, where, in theory, you could have a system like ChatGPT generate a trillion words' worth of text, which current copyright would be not copyrighted, and then, retrain a subsequent system, where you throw away the original one entirely. It's not derived from copyrighted works at all, but it still is likely just as performant as the original.

Mr. KILEY. My time is up. So, I'll just say, in closing I kind of worry that many of the issues we're talking about here might sort of be obsolete within a matter of years, especially as the capabilities of these systems advance.

I do think that there will always be a desire for works of purely human creation. I don't know. This is a bad analogy, but, we have at the grocery organic and the nonorganic sections. So, I think that this is an argument for making sure that we have transparency throughout the process, as to when you're dealing with something that was created by an AI versus a human. That may be well into the future we could have different markets, but I do believe there will always be a desire for the works that are produced by, the kind of works that are produced by the artists that are here today.

Thank you very much.

Mr. ISSA. I thank the gentleman.

I might note that, after we all saw the movie "WarGames," and found out that the outcome of tic-tac-toe, ultimately, is nobody wins, we still have generals, colonels, captains, soldiers, and some of them fighting in Ukraine as we speak.

With that, I go to the former U.S. Attorney and my colleague from California, Mr. Schiff.

Mr. SCHIFF. Thank you, Mr. Chair.

I should correct the record. I was an Assistant U.S. Attorney, so as not to think I was promoted more than I was. Thank you.

Mr. ISSA. Humility is also one of the traits I admire in you.

Mr. SCHIFF. Thank you, Chair.

The explosion of everyday AI use practically overnight has caused, caught the attention of many in the IP space, and for good reason. As many of you know, my constituents in Hollywood in the entertainment industry, and more broadly, in California, are particularly impacted by the rise of AI. We eagerly look to see how it will affect the creative industries.

A recent analysis of ChatGPT's training sources by *The Washington Post* found that 11 percent of the models' input data comes from arts and entertainment, including movies and television, art and design, and entertainment events.

Mr. Navarro, I want to echo a remark that you submitted in your testimony today. "Technology has long empowered human expression and AI will be no different." It's true that new technologies have the potential to complement or augment art and have certainly done so in the past. AI can be used to enrich the work of those in the creative industries.

Behind all that work are human artists, people. The copyright system is the foundation for the entire economic marketplace that

allows American artists and creatives to earn a living and American companies to create jobs for the sake of producing art.

Copyright law must continue to incentivize and protect this activity in the United States and beyond. Blanket AI exemptions do not exist in current law, nor should they, for copyright infringement.

I wanted to ask two questions:

- (1) Mr. Navarro, and that is, under what circumstances would you be interested, or do you think other artists would be interested, in licensing their work for AI? Or do you think there's a broad enough concern about moving away from human-generated music that you think artists shouldn't sell their works to be used for AI? Is one question.

Then, another question for the broader panel, and that is, disinformation is a grave concern with AI. It may very well affect people in the political world, but it also will affect people in the arts.

- (2) How do you see disinformation about artists or about their works being a danger, and how do you think that can be addressed?

Mr. NAVARRO. Certainly, the area of disinformation is underscored by the rise in the last few years of the deepfake phenomenon, which I was at your talk at SAG-AFTRA. I'm on the National Board of SAG-AFTRA.

This is a tremendous issue in terms of right of publicity, but, even just putting faces on bodies that are doing things those bodies wouldn't do—I don't think there's a single answer to whether it should or should not be allowed. I mean, what might be right for somebody to have their face put on a body that, suddenly, is a war hero might be OK. To have a face put on a body that's doing something pornographic would not be OK. I don't mean to be indelicate in that, but these are some of the concerns.

With regards to music, it's probably a little different. I think that every individual has their own line of demarcation as to what should or should not be done with their work. Certainly, when you're dealing with the physical countenance of somebody, that the standards and the strictures are going to be greater.

Mr. SCHIFF. Thank you.

We held an open hearing in the Intelligence Committee years ago on deepfakes, and among the most chilling observations was that, once you see a deepfake, even if you're later persuaded that what you saw was a fake, you never completely lose the lingering negative impression it left with you. So, the damage is done when you see it.

Would others like to comment on the danger of disinformation to artists?

Mr. SEDLIK. There's a very important initiative called the C2PA. It's the Content Authenticity Initiative, and it's led by Adobe and others. They've determined a way to track any changes made to images, whether by AI or otherwise.

So, that, especially when it comes to photojournalistic images, people could be confident in the providence of the images that they're looking at. That's C2PA.

Mr. SCHIFF. Thank you.

Mr. IRWIN. I was just going to say that, certainly, in my own experience—and I'm sure for the other people here—you put in a lot of hours from a very young age to try and do what you do and become really good at it. To dismiss that in any way, I think that's the problem that a lot of us are having, it was Malcolm Gladwell who wrote, "in 10,000 hours, you start to know what you're doing."

If the machines can do it that quickly, what is the incentive for us to keep going? What really is it? For years and years and years, for as long as you can remember, going back to the court composers, the arts people have been traded on. Oh, they'll do it because they love doing it. They love doing it. It's true, we do love doing it.

At some point, the love doesn't feed your family, and that's the real harm here, is there has to be a way for us to coexist. That's all we're looking for really.

Mr. CALLISON-BURCH. Mr. Chair, may I add a comment? Because I'd like to highlight agreements on this issue.

Mr. ISSA. Uh-hum, quickly.

Mr. CALLISON-BURCH. I am in absolute agreement with artists on the panel here that this is an important issue and people should not be allowed to imitate another person through deepfakes or through imitation. I think the right of publicity is something that's very worth considering, as you consider legislation on this.

Mr. ISSA. I thank the gentleman.

We now go to the gentleman from South Carolina, Mr. Fry.

Mr. FRY. Thank you, Mr. Chair. Thanks for having this hearing today.

This is a really interesting topic, right? We look at this. I have not ever in my practice done IP work. AI is the next wave of—or the wave of the future, if you will, what we're going to be dealing with in this country, really, on all spectrums.

Mr. Burch, I got a kick out of reading your bio, where your Ph.D. students joke that, whenever they ask you anything, your first response is, "Did you ask ChatGPT for that?" I think that really kind of sums up why we're here, right?

Mr. Damle, I want to ask you something. I've been bouncing around in committee hearings all day, so I may have missed this. Can you identify inadequacies of our existing laws to address copyright protection of AI?

Mr. DAMLE. I think that there's a lot of questions that still need to be answered about whether AI output is protectable by copyright and the circumstances in which it is. This is such a new issue. The Copyright Office has put out its guidance, but that guidance really addressed kind of one end of the spectrum of the question, which is, where there is virtually no or minimal human input into the creative process, does the output qualify for copyright protection? The office says no. That's just one end of the spectrum.

There's a whole area from there to somebody sitting with a paint brush and paint painting on a canvas, where you're using technology to assist you in the creative process, whether that's autofocus on a camera, whether that's Photoshop, or, indeed, whether that's generative AI.

I think generative AI, properly understood, is going to be a tool for human creativity. If you talk to artists that really have incorporated generative AI into their creative process, they don't see it as a substitute for their own creativity. They see it as a way to enhance their own creativity.

So, we're going to have a lot of hard questions in that space, where there is really an iterative process between the human author and the generative AI system to determine whether the output of that process is copyrightable.

Mr. FRY. Do you think we need a completely new set of rules? Does existing contract law maybe cover this? Or is it possible for this to fit in with existing law?

Mr. DAMLE. I think existing law is well-suited to deal with all the questions we've been talking about today. Congress had the wisdom in the 1976 act, and various amendments since then, to build a technology-neutral, flexible copyright regime. I think it's proven time and time again that, no matter the new technology that comes along, the laws are able to adapt to them.

There are instances where perhaps they're not, and Congress can step in and act in those circumstances. In general—and I think in this space—I think our existing copyright laws are well-suited to handle the questions that—as they arise in this.

Mr. FRY. When I was reading the guidance for this, and reading the CRS report that was issued kind of surrounding this, what struck me was it reminded me of a test in law on whether somebody was an independent contractor or an employee of a company, right? So, the test that they look at is the degree of control in which somebody exercises over that individual on whether they're an employee. You can call them an independent contractor all you want to, but if you're the one doing the schedule, you're the one putting in all these requirements in what they do, they're not actually an independent contract; they're actually an employee at that point.

So, to me, there are some similarities there that, when we're talking about the degree in which there's litigation on, hey, create a song that sounds like this, and there's no input, but where do we go? Should we be looking at it in those terms, as the degree in which we provide input to AI to the generated product as a test?

Mr. DAMLE. Thank you, Congressman.

I think that's a really useful analogy. In fact, there is part of copyright law, called the work for hire doctrine, that really asks that exact question—the extent to which you exercise, as the employer, control over the creative output of your employee, or to the extent they are able to do things on their own. That determines whether it's a work made for hire or not.

I think that we are people that are copyright lawyers are looking at that body of law already as a way to draw analogies in this space—considering the AI as an employee, or is it an independent contractor? Is it off running on its own or does the human author actually have some control over what the AI is outputting? So, I think it's a very, very useful analogy. You've sort of anticipated where, where copyright lawyers already are on this.

Mr. DAMLE. Thank you, Mr. Chair. With that, I yield back.

Mr. ISSA. I thank the gentleman.

We now go to the gentlady from California, Ms. Lofgren.

Ms. LOFGREN. Thank you, Mr. Chair.

This has been a very useful hearing, and I appreciate the testimony of each of these witnesses.

I believe that AI is going to upend a lot of careers. In fact, the House of Representatives has had three bipartisan, Congress-wide briefings on AI in the last two weeks. It's going to upend the practice of law. It's going to upend the practice of medicine. It will probably upend engineering.

The difference for creators, at least you have some protection in the law, which is copyright, which is absent other professions. The question is, how will that work to protect creators?

I was glad that Mr. Altman in his testimony indicated that creators should be compensated and do have rights and that recognition. The technology is complicated, and how that is going to work, we don't know yet.

I'll disclose that, a number of months ago, I put together some creators with the AI people to see if we couldn't have discussions. I was thinking, honestly, about the Music Modernization Act, which I think was very successful in reaching negotiations, so that creators could be compensated. I'll express some disappointment that progress that I expected to have been made by now has not yet been made. So, I'm hoping that those discussions will re-energize.

Here's a question, I guess, for Mr. Burch, Professor Burch, and maybe Mr. Sedlik, since you represent different ends of the knowledge base here. We've got, basically, two questions. You've got the input question, which is, basically, lines of code that have been assembled, and then, you've got the output question, which is how much of this is infringing.

Is it, in your judgment, even possible to reach an agreement like the Music Modernization Act did, so that creators can be fairly compensated?

Mr. CALLISON-BURCH. I think that there are a number of practical issues that make this very difficult to imagine, not least of which is the fact that we're not performing anything when we're outputting something from a system. It's not simply a collage where we're combining elements of existing work. It's genuinely been distilled into a form that's completely different than the original. That is going to be the crux of what makes this difficult.

I think that there might be a market for licensing images and songs, and things like that, that companies end up voluntarily entering into, but I don't think that the practical implementation of such a thing will be as easy as it was for the MMA, which I understand was already very complex.

Mr. SEDLIK. I don't think that most photographers are concerned or whining about the fact that AI is going to affect—it's going to compete with them in terms of it's easier to create images with AI than it is to create a photograph. It is.

What they're concerned with is the use of their works in that system, grinding it up and spitting it out as generated AI, based on their works, and the fact that they're not compensated, and it's done without their authorization.

Images are data, but they're not merely unprotected facts. An image, my image is my depiction of a fact and applying my creative expression to depict it. So, the copying of that, under 106, irrespective of anything else, is an infringement. It's copied into a system. That's infringement right there.

In terms of the output, that output may or may not resemble my work. It may or may not infringe on my work. The input, copying the work under 106 would be an infringement, unless, fact-specifically, it's fair use.

Ms. LOFGREN. I thank you for that. As I was listening. I was thinking back to my prior service as a staffer to my predecessor in office, Congressman Don Edwards, and his partnership with Bob Kastenmeier in the 1976 Act, and how the work that they produced has endured to this day—to protect the creative forces.

Obviously, we need to meld technology to help that protection. Adobe, which is located in my district, actually did the watermarking or their tagging. We actually used that in the January 6th Committee to prevent our material from being altered.

So, I think there's some real possibilities here. I think we've got a long road ahead, but I'm hoping that the discussions that are ongoing can ramp up a little bit, because I think that's probably the most productive way to reach a successful conclusion.

With that, Mr. Chair, I'd yield back.

Mr. ISSA. I thank the gentlelady.

We now go to the gentleman from Texas, Mr. Moran.

Mr. MORAN. Thank you, Mr. Chair.

I know that all of you have been sitting there for quite some time. So, before I ask my questions, I'd like to pause for a quick musical interlude.

[Music plays.]

I know that was just a short interlude, and all of you would, no doubt, like to hear the rest of that song.

[Laughter.]

As you heard, it is a rendition of something by Drake and the Weeknd. Many of you, I would suspect—maybe all of you—believe that this is likely their true voices and, in fact, a musical, a musical song written and produced by both the Drake and the Weeknd. In fact, it was computer-generated using only snippets of those artists' original voices.

Even the most ardent fans of Drake and the Weeknd—and I'll admit, they're not my first choice; I'm more of a TobyMac and Steven Curtis Chapman guy—but even the most ardent fans didn't realize that this was not their voices. In fact, in just a few short days, that song garnered over a half of million streams on Spotify before it was, ultimately, taken down.

I'll also admit that I am profoundly blown away by the artistic values sitting at this table, but then, also concerned about what's going to happen to the creativity if we don't get a hold of this artificial intelligence issue and protect the creativity of human development.

Just like what Drake and the Weeknd need protection for, all human creativity needs protection from what we're seeing from the artificial intelligence community here. I'm curious if any of you knew that story about that song. Were you all aware that this had

happened? Did you guys hear that song before you knew that it was not actually Drake and the Weeknd? Yes, it's amazing to me that it sounds exactly like them.

Mr. Navarro, I want to ask you a couple of questions. Mr. Irwin mentioned three primary issues to focus on—consent, credit, and competition—compensation. I appreciated the fact that you mentioned those. Mr. Navarro, as a generational singer and songwriter, why is not enough just to give credit to an artist if AI uses your voice or prior works as a basis to create something new? Why is credit—

Mr. NAVARRO. I am very familiar with this particular case. Royalties were generated. I serve on the Unclaimed Royalties Oversight Committee of the Mechanical Licensing Collective under the Music Modernization Act and the U.S. Copyright Office. Where do those royalties go? Who do they go to? They didn't authorize the use of their voices. AI could be used to put abhorrent words into their mouths. They didn't get their permission. They have contractual relationships with Universal Music. This operates in violation of it.

Many, many entities are harmed by using this without going through whatever the proper channels are. No means no. If they had said no, no means no.

Mr. MORAN. Yes, because you mentioned earlier, even if you have credit and compensation, without the consent it's really nothing more than compulsory licensing. Is that true?

Mr. NAVARRO. Yes.

Mr. MORAN. Do you consider, Mr. Navarro, derivative AI works from your original works that sound like you to be works in competition with you?

Mr. NAVARRO. I think they are works superseding me.

Mr. MORAN. Mr. Damle, I want to ask you a couple questions, because I wrote down some quotes that you had throughout the testimony today. I couldn't follow some of the consistency in them. I want to give you, read you back some of your quotes.

You said at one point, "existing law is well suited for everything we are dealing with today." Then at another point, you said, "people should not be allowed to imitate other works." Then you answered the question, you said earlier, "we need to look at other areas of the law other than copyright, for examples, like the Drake and the Weekend," the example that I just gave you, for protection in examples like that.

So, are we perfectly suited under existing law, or do we need additional laws to protect artists like the ones sitting at the table?

Mr. DAMLE. Thank you, Congressman. Just to clarify, I am not sure all those quotes were mine. I think only a couple were.

Mr. MORAN. All those quotes were—

Mr. DAMLE. So, the question of whether copyright law needs to be changed, I think the answer to that is no. I think that our copyright law is—I am a copyright lawyer, so I sort of focus my testimony on copyright law. I think copyright law is well suited, flexible enough to deal with the copyright questions that are being raised.

Now, I acknowledge that there are concerns like with this, with the Drake and Weekend track that are legitimate and need to be

considered. I think copyright law is really a blunt instrument, too blunt an instrument to deal with that.

So, it may be worth looking at other areas of law outside of copyright. Professor Callison-Burch mentioned right of publicity. There is trademark law. There are other areas of law that I am not expert in that may be better suited to deal with these situations than copyright law.

Mr. MORAN. OK. You may be right. Professor Burch may have said people should not be allowed to imitate other works. He may be the one I need to attribute that quote to. So, I stand corrected, not my first time in Congress, won't be my last time.

Thank you, gentlemen, for your time today. We appreciate it. It is a complicated issue. We want to get this right. Thank you.

Mr. ISSA. I thank the gentleman.

We now go to the gentlelady from Pennsylvania, Ms. Dean.

Ms. DEAN. I thank you, Chair Issa, Ranking Member Johnson, for holding this important and very interesting hearing. It is also very pleasant to be in the company of this talent and to have us really getting along very, very well because of the subject matter. So, thank you, thank you. It has been a real pleasure for me.

I know this won't be the last of our hearings. It is really among the first. There is so much to learn. There is so much for me to learn. We know as lawmakers we can't wait after widespread societal use to come back and say what did we do right, what should we have done in advance. We have learned that from other technologies.

I want to lay a couple of my biases on the table. I taught for 10 years at a different Philadelphia university. I taught writing, all different levels of writing, to students at LaSalle University. To very much what most of you have said, Mr. Navarro, in particular, I always taught my students know your craft, understand the mechanics, break the rules when it makes sense, but make sure you place your humanity in whatever you write. That is the genius of what you create, placing humanity in it.

My other bias is I am a copyright holder myself. My son, Harry Cunnane, and I wrote and published a memoir of his struggles with addiction, our family struggles with his disease of addiction, but much more importantly the power and the hope in recovery. We also wrote a children's version of that book. It was made into an audio book with the extraordinary generosity of Mr. Paul Williams writing and producing and performing the background music for our children's book.

My son, Pat Cunnane, is a television writer and movie screen writer. So, when some of you talk about how long some of this stuff comes, takes to come, I keep saying to Pat, where is the movie, it is years. Maybe we will see his name. He has done really well. I bring those biases to the fore.

When I consider what we have seen of generative AI, two questions seem glaringly obvious, so if you will help me with these. The first is the creation of these models being done in a way that respects the rights and interests of authors, musicians, artists, content creators, to your very point, consent, credit, and compensation? If I frame it a different way, is anybody doing this right?

Maybe I will start here with Mr. Sedlik. Anybody doing it right?

Mr. SEDLIK. I think that I am seeing a silver lining on the cloud. Some of that AI platforms are beginning to listen and to adapt their systems to respect rights of authors, or at least they are saying that this is in progress.

Ms. DEAN. OK.

Mr. SEDLIK. There is really two gateways here. There is the gateway to allow works into the data bases that are used for ingestion of images into these systems. Then there is also a gateway in terms of the prompts that are entered and uploading of images as image prompts, copies of our works uploaded by others without our permission as image prompts. Those two gateways are of concern.

Ms. DEAN. Mr. Navarro and Mr. Irwin, could you offer me your thoughts? Anybody doing this right?

Mr. NAVARRO. I am not aware of who is doing it right. I don't believe that everyone is doing it wrong. I think it is so brand new, as it is the Wild West out there.

I think of the earliest days of sampling, when people sampled stuff for records routinely and didn't give credit, compensation, nothing. Now, it is routine. Can't Touch This by MC Hammer credits Rick James for Super Freak, and his estate gets compensated. So, we are in the process of trying to get it right.

Ms. DEAN. OK. Mr. Irwin, briefly, if you don't mind, I have one more question to ask.

Mr. IRWIN. Yes, I have no knowledge of anyone doing it right. I think there is starting to be some overtures as to, and inquiries as to let's get together and talk about it. At the moment, no, not to my knowledge, no.

Ms. DEAN. Dr. Callison-Burch, in the time I have remaining, I was particularly taken in your testimony by something you said about what is, the impact on labor, on workers. You used the expression and you suggested will paralegals go the way of the lamplighter. What are the implications for the labor market?

Mr. CALLISON-BURCH. I think this is the large issue that everyone needs to deal with and where Congress can have the most effect. I feel that at the moment we may be on a precipice of mass unemployment. I think the probability is very small. It is such a dire outcome that you really need to consider some sort of legislation like in case of emergency, start a new WPA.

I think that dealing with this as a copyright issue almost entirely misses the point that these systems are coming. We have in America experienced unemployment as a result of automation before. It has largely affected blue collar work but now has the potential to also affect white collar work as well.

Ms. DEAN. Fascinating.

Mr. Chair, again, I thank you. I yield back.

Mr. ISSA. I thank the gentlelady.

We now go to the very patient gentlelady, Ms. Lee of Florida.

Ms. LEE. Thank you, Mr. Chair. Thank you to all our witnesses who are here today. We so appreciate your time and your testimony helping inform us about how we might embrace the emerging technology of artificial intelligence, but also recognize the immeasurable value of our artists and the need to protect and balance intellectual property and copyright protections.

I would like to return, Mr. Damle, to your testimony. Specifically, during the questions by Congressman Moran and Fry, you talked a bit about your perception that existing copyright law was adequate to take on this new emerging landscape and continue to resolve that balancing between the property rights of artists and our new technology that we see.

One thing that concerns me is that when Congress fails to be sufficiently clear we leave to judges the task of figuring it out. Of course, we want judges applying the law not creating it. We don't want to put judges in the role of being policymakers.

So, I would like for you to elaborate a little bit more on your conclusion that our existing statutory framework is sufficient to take on this new challenge.

Mr. DAMLE. Thank you, Congresswoman. It is an excellent question.

I think if you just look at history here fair use has existed in the copyright law for about almost 200 years. Over that time, it has dealt with lots of massive shifts in technology.

In 1984, the Supreme Court applied fair use to the then-new technology of VCRs and said, applying again 180-year-old law that did not have any understanding of that kind of recording technology, and made a really fact-bound, cautious, and careful decision about whether home recording was fair use or not.

Fast forward through the era, recently the Supreme Court handled a case involving software APIs, again, something no one could have ever dreamed of as being a copyright issue 180 years ago. Yet, the court, again looking at the very specific facts of that case, determined that the reuse of software APIs was fair use.

There have been other cases going the other way, looking at new technologies like Napster and saying that is not fair use, that is an exploitive use of the technology.

So, my perspective comes from looking at that long history and how copyright has been able to manage shifts, even major shifts in technology.

Ms. LEE. Professor Burch, what is your take on that same question and the conundrum of not overregulating, but at the same time not leaving it to courts to try to create policy?

Mr. CALLISON-BURCH. I think that it is worth considering all these issues and deciding whether or not copyright needs to be extended. I think there is many things that my fellow panelists are talking about that are 100 percent valid and need to be incorporated into the ethical guidelines that AI system developers create, including right of publicity and copyright and characters and things like that.

I don't, and those are not currently covered sufficiently by copyright law. They may be covered by other laws sufficiently. I think that is where the target should be, like what is the output of these systems and what is correct use and incorrect use of the output of these systems.

Ms. LEE. OK. Mr. Navarro, one of the advantages about getting to ask questions near the end is that I can bring to you this one. Is there anything that you wanted to share with the Committee today that you have not to this point in the hearing been asked?

Mr. NAVARRO. That is a good question. Yes. First, when we look at the ingestion of trillions of pieces of data, that is one way of looking at it. When we look at the impersonation of voices, that is really one piece of data. That is very specific. It is easy to get lost in the shift between the macro and the micro as we figure out where to go on this. So, I think we need to take a look at both.

I think the other is that I appreciate the technology. I appreciate the technical use of terms. My music isn't data.

Ms. LEE. Mr. Irwin, the same question to you, anything you wish to share with us today that you haven't already testified to.

Mr. IRWIN. I am going to follow on what Dan said. I am a little distressed that we are calling this training, to be honest, because in my mind we train athletes or animals. We don't train machines. We equip machines with data, as the technology people like to call it, but as we like to call it music.

It is very Orwellian how the tech industry manages to change terminology on us. It is not data or content to us. It is music. It is photographs. It is not file sharing. It is stealing. It is very simple. They are the sort of things I wanted to get into the record that I didn't get to say.

Ms. LEE. Thank you.

Mr. Chair, I yield back.

Mr. ISSA. A brilliant move for such a new member.

With that, we go to the gentleman from Maryland for five minutes, Mr. Ivey.

Mr. IVEY. Thank you, Mr. Chair and Mr. Ranking Member.

Ms. Lee, you stole my question there. I did want to go back to the deepfake issue. Mr. Sedlik, I think you talked a little bit about technology that exists that allows, that gives the ability to separate out what is real and what is fake. You said C2 something. I wanted to get more details on that.

Mr. SEDLIK. Sure. So, a consortium came together managed I think by Adobe or founded by Adobe and with various industry players. They looked at, they explored methods of creating a technology that would allow you to use software to determine whether or not an image has been altered and what the provenance of that image is.

That information is stored in the image so that when you are using software to view an image you can tell whether or not that image has been altered, for example, swapping out a head, changing, removing something, adding something, any sort of revision to the image, to provide that information to the public and to people who might rely on those images. That is C2PA.

Mr. IVEY. OK. Here is where I am going with this. So, I guess what you are saying is that, a movie or some visual depiction that has the code embedded in it so that distinction could be made, then you could, you have the software to identify it. I guess the question I would have is: What if it is not software-generated or if it is just a visual?

Here is sort of the big question. So, for example, body cameras worn by police, and I know this isn't a creative content question. Body cameras have become super relevant in court with respect to police cases and in many instances videos. Former President

Trump's video in his most recent trial was pretty significant as well.

I was sort of curious if it is AI-generated entirely, so you would take the subcomponents of the training piece or whatever, and I went to law school instead of any kind of hard science classes, so I am struggling with the technology terms. Would you be able to tell if an image or a video is generated entirely by an AI approach as opposed to some of it being real or some of it not? Is there a way that you could tell if it is authentic or not?

Mr. SEDLIK. There is, the AI platforms are doing some work on this and coming forward with technology to be able to allow the public to discern what is AI and what is not. With respect to other types of creations or recordings, like you mentioned the police cameras and such, those have, those embed time code and other information in the recording so that you can tell whether or not there has been anything removed.

In terms of finding something or determining whether an image is AI-generated or not, that is right now at the experimental phase with scientists looking at it, image scientists. I think we are going to see something in the very near future.

Mr. IVEY. OK. Mr. Navarro, did you have some thought on that as well or—

Mr. NAVARRO. From a musical standpoint, we sometimes use our taste. I did a touring test at South by Southwest in March where pieces were being played. I was able to guess about $\frac{3}{4}$ of them, but not all of them. It happened to be with very simple forms of music that were style-based. Sometimes the vocals were gibberish. Other times it was just, were moves made musically that no great musician or a great producer would allow.

It gets more difficult. This particular piece that we just heard from the gentleman from Texas, that would have been really, really difficult to tell. So, from that standpoint, I think that is where one of the great dangers is, is in not knowing the difference.

Mr. IVEY. Professor Callison-Burch, I wanted to sort of touch base with you on this kind of issue as well. I know we have been talking about copyright law, but from the standpoint of protecting the community, the world, whatever from these sorts of fabrications, which I don't know if they are illegal at this point or not. Are there steps we need to take to address this?

Mr. CALLISON-BURCH. Absolutely. My Ph.D. students have done the largest-ever study on human detectability of machine-generated text. We have found that over time it is getting harder and harder to detect, but that people can be trained to detect, so I think similarly to how Dan is saying that the latest clip of the fake Drake song is increasingly passable as human, whereas five years ago it wouldn't have been. Like that is a trend that we are on. I think any sort of basic media literacy that we teach to our children should include this as an element.

Mr. IVEY. OK. With respect to visuals? This is from the court standpoint on some of these images. Sometimes people's—whether you go to jail or not is determined by a visual or a video clip that a jury might see. Are there ways to make sure that whatever is presented or a judge making a determination about what could be viewed is actually authenticate?

Mr. CALLISON-BURCH. There is expert testimony through examination of the images. I think you could likely still detect artifacts of a machine-generated image if it is done carelessly. I think it is going to get harder. So, it is certainly something to keep under consideration as time progresses.

Mr. IVEY. All right. Well, I want to thank all of you for coming in today. I apologize for a question that is a little off the topic, but I think important for us.

Mr. Chair, I really wanted to thank you and the Ranking Member. I agree with some of the previous comments about this being a refreshing change from some of the types of hearings we have had, previously. I thank you for this.

Mr. ISSA. Well, I thank the gentleman. I will welcome you to the Subcommittee, which has historically been extremely bipartisan. So, we are going to keep that tradition going when it comes to, we can't pass intellectual property reform except on a completely bipartisan basis.

Now, last and probably least, you get me. I am going to start with this. Thank you, Mr. Ranking Member. This is an actual portrait. You will recognize some of the characters. They are pretty much mostly all dead. Some of them, in fact, all of them were produced by an artist who took a number of photographs, including Lincoln photographs, to produce this product. I am confident that Andy Thomas did not pay for most of the photographs.

Do you see a problem in, and I will start with Mr. Navarro, in that being fair use, the collection of photographs, since for the most part these gentlemen were not available to sit or stand for this portrait?

Mr. NAVARRO. I think if it were me, I probably would have made an attempt to compensate the photographers. I don't know enough about the derivative work clauses in the Copyright Act that can allow something like that. I happen to think that is a wonderful painting. I am a Democrat. I still think it is a—

Mr. ISSA. I wanted to bring both, because I have another one with all the Democrats. They both hang in my office. You are welcome to come afterwards.

Mr. NAVARRO. As such, my personal opinion is that creativity has been enhanced, and communication has been enhanced. However, were it me, I would have made an attempt to contact the photographers.

Mr. ISSA. I believe that he would have bought those pictures if they were commercially available, whether they were under copyright or not.

So, and I point that out because it is one of the challenges that I face in trying to steer this Committee now and in the future is I try to look back, as I did there, on existing copyright implementation, existing art. Andy Thomas clearly does his own art. He works off of, even when he is doing a living person, he generally will come in, take pictures, and take those pictures back to do his work. It is a style. It works very well for him obviously. I ask that more for all of us to opine on afterwards.

I do have one similar situation. If we believe Wikipedia as always correct, The Beatles derived their inspiration from among others Elvis and Chuck Berry. Elvis gave credit to his inspiration,

including none other than Dean Martin. Dean Martin credited The Mills Brothers and Perry Como. Perry Como gave significant credit to his inspiration to Bing Crosby.

Oddly enough, Bing and each of those people were all alive at one point, all entitled to their copyrights at the same time. Yet, we were able to work our way through who, what, and where. Every one of these artists undoubtedly took sheet music and practiced with those inspirations that they had. Every one of them listened to those people's recordings. Undoubtedly, every one of them performed and tried to do it at least in training in that style.

So, the question I ask, and I will just go right down the line starting with Sy, we have a conundrum here at a minimum. If I am taking a class, high school, college, whatever, the material used for that is typically paid for as copyrighted material once to educate me. After that, my education, assuming I am not using the script itself, my education goes forward without further copyright. If I take a significant amount from one of those college books that I purchased on which the copyright was paid, I have an obligation to disclose it, and if it is beyond a snippet to pay.

Is that really what you believe is at the root of how computers, and I apologize to Mr. Irwin, learn or collect data?

Mr. DAMLE. I think the analogy is a very close one, both as a legal matter and, as Professor Callison-Burch can talk about, the technology. At least as I understand the technology, that is also true as well, that the copyrighted works are being used not to create a collage or record the copyrightable content within them, but to simply learn statistical facts about the works themselves.

It is a very similar process to the way humans learn. Now, the problem with machines is that they learn much more inefficiently than humans do. I can read three or four books about a topic and then be conversant about that topic and maybe even write my own article about that topic. Machines are not that skilled. They haven't quite caught up to us yet. So, they require, to do a similar kind of learning, they require billions and billions of pieces of work to reach those same learnings.

Mr. ISSA. Just because they are a bad student doesn't mean that the copyrighted material shouldn't be paid for, does it?

Mr. DAMLE. Now, the question of whether it should be paid for is a different question.

Mr. ISSA. You are the lawyer. The gentlemen on this end want their work paid for even if it is used in the classroom. That is the final part of this question, bringing it all back together, is you two have been very good at calling this education, training, et cetera. If the analogy of the classroom, of teaching, of learning is there, there is also the analogy that this copyrighted material is paid for. You didn't go to the public library, so to speak. You went and got this material and ingested it just as I ingested, for the most part, my college years.

So, I'll go to Mr. Burch. I really want to get through this for everyone because I think it is part of what brings a close to this hearing.

Mr. CALLISON-BURCH. Thank you. This is a great analogy. So, again, I think it highlights the fact that the systems are learning.

They are learning facts about the world. Those facts are not copy-rightable.

I think where the material is acquired from and whether that is fairly, fair use or not fair use is exactly the right question. Many of my students learn by going out to the Internet and retrieving facts from the Internet that they do not pay for. That is what is happening here. So, I think—

Mr. ISSA. You mean you haven't published a book that you make them buy like my professors? No—

Mr. CALLISON-BURCH. I have not published a book that I have forced my students to buy.

Mr. ISSA. Mr. Irwin.

Mr. IRWIN. I think what was really interesting in your little Wikipedia piece you read was the word inspiration. I think inspiration is what we need to keep hold of here, because there is a big difference between getting a book and you want to emulate your idol. You want to emulate The Beatles, or you want to emulate The Rolling Stones.

In my case growing up, my parents bought me sheet music. The sheet music was Mozart or Beethoven. It was public domain. They still had to buy the sheet music for me to learn. I learned how Beethoven writes music, how Mozart writes music, learned The Beatles, and I learned all these things. I learned them.

Every time, in your case, the inspiration was because they heard it on the radio. If they heard it on the radio, there were royalties being paid. If they bought the sheet music, there were royalties being paid. If they bought the record and took it home and played it, there were royalties being paid. That is really where we are now. Yes, they were learning. That inspiration was paid for at every time along the way.

Mr. ISSA. Mr. Navarro.

Mr. NAVARRO. Mr. Irwin has hit it on the head with the notion of inspiration. I don't think we should ever litigate against inspiration. That is how we evolve as a culture. I think the beauty of it is that no matter how inspired you are you are never going to get it perfect. Therein lies the individuality of the subsequent artist is they bring their own stamp to whatever they learned on. I learned on many singers to develop myself as a singer.

Mr. ISSA. You are not giving credit to one the way Wikipedia did for these?

Mr. NAVARRO. Well, no. Well, there are a few. It branches out. We still change things a little bit.

Also, with regards to stuff like that, I spent some years in advertising before I became a professional songwriter and musician. I date back from the era where if it got too close on the radio you said celebrity voice impersonated. You had to reveal it, that it was an impersonation lest somebody—yes, well, Tom Waits and Bette Midler with that.

All this is to say is that inspiration should never be stifled. I am not interested in stifling technology. We benefit from it, guardrails, credit, consent, credit, and compensation. Thank you.

Mr. ISSA. Mr. Sedlik.

Mr. SEDLIK. I agree with my colleagues. Your example of the classroom was very insightful. One of the largest areas for copy-

right licensing is educational use, the textbook industry and in turn photography and illustrations being licensed for educational use.

If you want to teach in the classroom about B.B. King, you can't use my photograph of B.B. King without licensing that photograph. If you want to teach about me or my photograph of B.B. King, that is going to be fair use.

It is true that all artists have been influenced by others. There is a difference between influence and inspiration or copying or theft. All our works are made based on our lived experience as humans and taking those works and being inspired by multiple of them to create, for example, the portrait that you showed.

If he relied on multiple portraits to create each individual portrait and was inspired by them rather than copying them, then there might not be an infringement there. If he traced it on a light box, if he projected it on the canvas, if he copied all the expression or the heart of the expression, there might be infringement there. That is a case-by-case, fact-dependent analysis.

Mr. ISSA. I think one of the interesting things about Andy Thomas is he manages to capture each of these men, both Republicans and Democrats, better than they were, which is something that a computer probably will take a long time to learn to do.

In closing, there are two things that were not discussed fully. One is the fact that even if copyright use is not attributable to a single artist, in other words the trillion into the works may not be effectively able to be done, it doesn't mean that the ingestion of that material should not somehow go to the benefit of the copyright industry as a whole. That will be one of the things that is not a statutory remedy today but could be.

To a certain extent it is like orphan works. You can't necessarily get them back to somebody, but you don't get to use them completely for free just because we can't assign them.

The last one, which was brought up here today, and I would like all of you to opine on it, because I think it was a real threat, but there wasn't time in this hearing to pick it up. In patents, in copyrights, and in trademarks currently based on some bad actors, computers can generate an infinite amount of combinations. If they generate an infinite amount of combinations of copyright material, they could, in fact, create a body of copyright that could eclipse future innovation, simply push out tens or hundreds of trillions of songs, of variations of art and, in fact, then make a claim, a troll-like claim that everybody else who comes up with an original piece, there I find enough to say that you took it from me even though you may never have seen the trillion different outputs.

It doesn't sound, it sounds far-fetched until those of you who look at AI and look at the petabytes per minute now that are being ingested on the Internet, and you realize that, in fact, infinity is closer to us than we ever thought it was.

So, I would like you to opine on that, because one of the challenges that I face is to limit copyright or patent applications or trademark applications that are computer-generated if, in fact, they serve only to limit human's ability to do individual creation. I would like you all to opine on that within your own expertise.

We will leave the record open for the next, I think five days is the Committee rule. I will shove it in if I get it later, as long as the Chair lets me.

So, I want to thank you. I expect that we have your numbers. We will be calling you. You have our contact information. I would hope that you continue to help us.

Mr. Ranking Member, do you have any closing statements?

Mr. JOHNSON of Georgia. I do not, other than to thank the witnesses for your testimony today.

Mr. ISSA. With that, we stand adjourned.

[Whereupon, at 12:38 p.m., the Subcommittee was adjourned.]

All materials submitted for the record by Members of the Subcommittee on Courts, Intellectual Property, and the Internet can be found at: <https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=115951>.

