S. HRG. 119–202

TOO BIG TO PROSECUTE?: EXAMINING THE AI INDUSTRY'S MASS INGESTION OF COPYRIGHTED WORKS FOR AI TRAINING

HEARING

BEFORE THE

SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM

OF THE

COMMITTEE ON THE JUDICIARY UNITED STATES SENATE

ONE HUNDRED NINETEENTH CONGRESS

FIRST SESSION

JULY 16, 2025

Serial No. J–119–30

Printed for the use of the Committee on the Judiciary

www.judiciary.senate.gov
www.govinfo.gov

U.S. GOVERNMENT PUBLISHING OFFICE
61–891 WASHINGTON : 2026

COMMITTEE ON THE JUDICIARY

CHARLES E. GRASSLEY, Iowa, *Chairman*

LINDSEY O. GRAHAM, South Carolina	RICHARD J. DURBIN, Illinois, *Ranking Member*
JOHN CORNYN, Texas	SHELDON WHITEHOUSE, Rhode Island
MICHAEL S. LEE, Utah	AMY KLOBUCHAR, Minnesota
TED CRUZ, Texas	CHRISTOPHER A. COONS, Delaware
JOSH HAWLEY, Missouri	RICHARD BLUMENTHAL, Connecticut
THOM TILLIS, North Carolina	MAZIE K. HIRONO, Hawaii
JOHN KENNEDY, Louisiana	CORY A. BOOKER, New Jersey
MARSHA BLACKBURN, Tennessee	ALEX PADILLA, California
ERIC SCHMITT, Missouri	PETER WELCH, Vermont
KATIE BOYD BRITT, Alabama	ADAM B. SCHIFF, California
ASHLEY MOODY, Florida	

KOLAN DAVIS, *Chief Counsel and Staff Director*
JOE ZOGBY, *Democratic Chief Counsel and Staff Director*

SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM

JOSH HAWLEY, Missouri, *Chair*

LINDSEY O. GRAHAM, South Carolina	RICHARD J. DURBIN, Illinois, *Ranking Member*
JOHN CORNYN, Texas	AMY KLOBUCHAR, Minnesota
TED CRUZ, Texas	CHRISTOPHER A. COONS, Delaware
MARSHA BLACKBURN, Tennessee	RICHARD BLUMENTHAL, Connecticut
KATIE BOYD BRITT, Alabama	CORY A. BOOKER, New Jersey

STEPHEN ANDREWS, *Republican Chief Counsel*
SAURABH SANGHVI, *Democratic Chief Counsel*

CONTENTS

OPENING STATEMENTS

WITNESSES

APPENDIX

TOO BIG TO PROSECUTE?: EXAMINING THE AI INDUSTRY'S MASS INGESTION OF COPYRIGHTED WORKS FOR AI TRAINING

WEDNESDAY, JULY 16, 2025

UNITED STATES SENATE,
SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM,
COMMITTEE ON THE JUDICIARY,
Washington, DC.

The Subcommittee met, pursuant to notice, at 12:03 p.m., in Room 226, Dirksen Senate Office Building, Hon. Josh Hawley, Chair of the Subcommittee, presiding.

Present: Senators Hawley [presiding], Durbin and Welch.

OPENING STATEMENT OF HON. JOSH HAWLEY, A U.S. SENATOR FROM THE STATE OF MISSOURI

Chair HAWLEY. Welcome, everyone, to the hearing today, which is entitled "Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training." This is the third hearing of the Senate Judiciary Committee's Subcommittee on Crime and Counterterrorism, which I am delighted to work on with my colleague, Ranking Member Durbin.

I want to say a special thank you to the witnesses for being here. Many of you, I think all of you, traveled in order to be here today. Thanks to everybody for accommodating our change in time. The Senate floor is going to be tied up here later today, and thus, no Committee business is happening, so thanks, all of you, for being here and for accommodating us.

I am going to make just a few opening remarks. Senator Durbin will do the same. Then we will swear in the witnesses and be off to the races.

Let me just start by saying that today's hearing is about the largest intellectual property theft in American history. For all of the talk about artificial intelligence and innovation and the future that comes out of Silicon Valley, here is the truth that nobody wants to admit. AI companies are training their models on stolen material, period. That is just the fact of the matter. And we are not talking about these companies simply scouring the internet for what is publicly available. We are talking about piracy. We are talking about theft. For years, AI companies have stolen massive amounts of copyrighted material from illegal online repositories.

Now, the FBI and the Department of Homeland Security regularly prosecute individuals who engage in exactly the same kind of behavior using platforms like LimeWire or Napster in the old days,

using a process called torrenting. But have these Big Tech companies been prosecuted? No, of course not. They are getting off scot-free. And this hearing will show us that Meta and Anthropic and other AI companies are willfully using these illegal networks, these torrenting networks as they are called, to steal vast swaths of copyrighted materials.

The amount of material that we are talking about is absolutely mind-boggling. We are talking about every book and every academic article ever written. Let me say that again, every book and every article ever written, billions of pages of copyrighted works, enough to fill 22 libraries the size of the Library of Congress. Think about that, 22 libraries of Congresses full of works. That is how much has been stolen.

And this theft was not some innocent mistake. They knew exactly what they were doing. They pirated these materials willfully. As the idea of pirating copyrighted works percolated through Meta, to take one example, employee after employee warned management that what they were doing was illegal. One Meta employee told management that, and I quote now, "This is not trivial." And he shared an article asking, "What is the probability of getting arrested for using torrents"—illegal downloads—"in the United States?

Another Meta employee shared a different article saying that downloading from illegal repositories would "open Meta up to legal ramifications." That is a nice way of saying that what they were doing was exactly, totally, 100 percent barred by copyright law.

Did Meta management listen? No. They bulldozed straight ahead. We will see evidence today that Mark Zuckerberg himself approved the decision to use these pirated materials. And then the best part, Meta management tried to hide it. They tried to hide the fact that they were engaged in the illegal download of pirated works, and not just the illegal download, but the illegal distribution of these same works. They tried to hide it by using non-company servers. They went so far as to train their AI model—get this. Meta trained its AI model to lie to users about what data it had been trained on. I mean, you talk about an inception-level-worthy deception, training the AI model to lie about what its own sources were. This isn't just aggressive business tactics. This is criminal conduct.

And I just want to point out, Meta's conduct is not an exception. This is the rule when it comes to what is happening right now in the AI space among these mega companies. Big Tech operates on the model of do whatever you want and count on the lobbyists and the lawyers to fix it later. They don't care about the rule of law. They don't care about America. They don't care about freedom. They certainly don't care about working people. They care about power and they care about money. And every time they say things like, we can't let China beat us, let me just translate that for you. Every time they say that, oh, we can't let China beat us, what they are really saying is, give us truckloads of cash and let us steal everything from you and make billions of dollars on it. That is the translation. We are going to see that in the testimony and the evidence today.

Here is the bottom line. We have got to do something to protect the people of this country. I am all for innovation, but not at the

price of illegality. I am all for innovation, but not at the price of destroying the intellectual property of the average man and woman in this country. We have laws for a reason. Those laws ought to be enforced, and Big Tech should not be above the law. Enough is enough. It is time to enforce the law, and that is what this hearing today is about.

Now, I will turn it over to Ranking Member Durbin.

OPENING STATEMENT OF HON. RICHARD J. DURBIN, A U.S. SENATOR FROM THE STATE OF ILLINOIS

Senator DURBIN. Thanks, Mr. Chairman.

The way AI interacts with intellectual property rights, particularly copyrights, is a critical topic we can't overlook. America's creative industries, including software, music, movies, literature, collectively contribute over $1 trillion to our economy each year, employing millions of people. While AI can be an incredible tool that unlocks further creativity, writers, artists, musicians, and others are rightfully concerned about what technology would mean to them personally. Should AI companies be able to use their materials freely as "fair use" or should they receive compensation when their works are used to train AI models?

I want to tell you, chapter one, how I discovered intellectual property. I was an attorney in Springfield, Illinois, and in a rash moment decided to buy a restaurant. So I joined a few friends and bought a restaurant, and we had live music. And I got a phone call one day from a fellow who said, I just was out at your restaurant. I said, great, did you have a good time? Couldn't have been better. Saturday night, the music was terrific. And I said, well, I am glad you had a good time.

And he said, you played 10 BMI tunes and six ASCAP tunes. I said, no, I didn't, I didn't play any tunes. He said, well, the way the law is written, you are responsible for the fact that copyright material was used by you to make a profit at your restaurant. I said, tell it to the judge. He said, no, before you say that, call your friend over in Jacksonville, Illinois, a few miles away and ask him about a similar experience. And his reaction was the same as yours. I called my friend who said, ask him how much money he needs each month for ASCAP and BMI, and we started paying it. That was my first course in intellectual property. I hold onto it to this day.

So how can creators compete with AI products that generate content at the push of a button, especially when the content might mimic or even produce their own work? These are just a few of the questions that we are going to consider in this hearing as we try to find the right balance between promoting technological innovation, protecting the work of our Nation's creators, and continuing to incentivize creativity in years to come.

We must recognize that AI innovation and protection of intellectual property rights are not mutually exclusive. That is why it is troubling, as I listened carefully to the Chairman, to hear stories about steps Big Tech companies are taking to train their AI models on copyright materials without compensation to the creators of these works. For example, rather than license authors' works, companies like Meta and Anthropic have obtained copyright materials

from sites that host pirated copies of the authors' books and writings. Anthropic pirated over 7 million books from shadow libraries. As Anthropic's CEO put it, Anthropic had many places from which it could have purchased, but it preferred to steal them to avoid "legal practice business slug," whatever that means. While Anthropic later became not so gung-ho about training their LLM on pirated books for legal reasons, it kept the pirated copies that it had already downloaded anyway. I don't get that.

As a judge in the *Meta* case recently put it, "Companies have been unable to resist the temptation to feed copyright-protected materials into their models without getting permission from the copyright holders or paying them for the right to use their works for this purpose."

This hearing is going to be interesting. Thanks, Mr. Chairman.

Chair HAWLEY. Thank you very much to the Ranking Member.

It is the practice of the Judiciary Committee and all of its Subcommittees to swear in witnesses before they testify, so could I ask you to stand up, raise your right hand, and repeat after me.

[Witnesses are sworn in.]

Chair HAWLEY. Very good. We will now proceed to opening statements. We will give 5 minutes to each witness. I will just say a brief word of introduction before each witness. We will just go straight down the table here down the dais. We will start with Mr. Max Pritt. Mr. Pritt is a partner at Boies Schiller, and he represents authors in a civil copyright infringement suit against Meta, among other matters.

Mr. Pritt, the floor is yours.

STATEMENT OF MAXWELL PRITT, PARTNER, BOIES SCHILLER FLEXNER LLP, SAN FRANCISCO, CALIFORNIA

Mr. PRITT. Chairman Hawley, Ranking Member Durbin, thank you for the invitation and opportunity to testify today. The Art of the Deal by Donald Trump, Hillbilly Elegy by J.D. Vance, Theodore Roosevelt: Preacher of Righteousness by Josh Hawley, these are just a handful of the many, many millions of copyrighted books and publications that some of the world's largest and wealthiest corporations—Meta, OpenAI, Anthropic, and others—knowingly and intentionally pirated from illicit online marketplaces for financial gain and to seek a competitive advantage in AI.

Today, this Committee begins to investigate and shine a light on what is likely the largest infringement of American intellectual property by U.S. companies in our Nation's history. As tech companies scrambled to release generative AI models and to catch up with OpenAI's ChatGPT, many of them turned to illicit online repositories to take tens of millions, if not hundreds of millions, of books and scholarly publications and articles for free instead of buying them or licensing them from copyright owners. By pirating these works, AI companies have built a multibillion-dollar industry that is projected to be a trillion-dollar industry in the next few years without paying a single cent to the authors whose works power their products or the publishers responsible for introducing those works to the public here and abroad.

Take Meta, for example. From the early days of its generative AI program, Meta concluded that training its models using books and

articles would help their performance. But instead of buying or licensing these works from copyright owners, Meta decided to take them from notorious online marketplaces of stolen copyrighted works, including some of the same ones targeted by the Department of Justice and the FBI for criminal copyright infringement. And Meta didn't just download books from these illegal repositories. It used the same kind of peer-to-peer file-sharing networks that powered Napster. In other words, Meta also made copies and sent them to other pirates.

In total, Meta pirated well over 200 terabytes, terabytes of pirated books and articles, a size comparable to the entire printed collection of the Library of Congress 20 times over, or the equivalent of a stack of many billions of pages of text. Meta's piracy included many millions of works, including at least 12 books authored by Members of this very Subcommittee and every U.S. President and Vice President in the 21st century. Meta also made and sent copies of over 40 terabytes of pirated works to others.

In doing so, Meta has helped to revive online piracy by propping up the foreign criminal syndicates that run these illicit marketplaces to violate U.S. copyrights around the globe. As Anna's Archive, the largest illicit online marketplace of stolen literature in the world today, says on its own website, "Shadow libraries were dying. Then came AI."

Meta is not alone, and it was not the first U.S. company to engage in rampant domestic piracy for its own commercial purposes. Pending lawsuits against OpenAI and Anthropic revealed that both companies also pirated millions of copyrighted works. And the decisions to engage in this mass domestic piracy were made at the highest levels. Company documents that are now public show, for example, the decision to pirate instead of license was approved by Meta's co-founder and CEO, Mark Zuckerberg, himself.

This decision to engage in mass piracy was made, even though key employees knew that doing so was both illegal and unethical. One Meta researcher argued that using pirated material should be beyond our ethical threshold. Another called Meta an accomplice to piracy. Yet another warned that if the media got wind of the company's use of pirated data, it could undermine Meta's negotiating position with regulators, the very people in this room and across the hall, in the White House, and in State houses across the country. And when asked if he cared whether Meta protects human creativity rather than exploits it, Meta's head of AI partnerships testified, he does not care.

AI companies now seek a pass for this unprecedented piracy by invoking a limited exception to copyright infringement called fair use, which Congress codified in the Copyright Act of 1976. They also argue they can't compete with China if they can't infringe every American's copyright. Nonsense. Our tech companies employ the best and brightest minds in the world, and they are the wealthiest corporations in the world. It is not credible for these companies to argue they can invest hundreds of billions of dollars into hiring talent and building data centers to power their commercial AI products and models, but they can't pay a single cent to copyright owners. There is no carveout in the Copyright Act for AI companies to engage in mass digital piracy.

I am grateful to Chairman Hawley, Ranking Member Durbin, and this Subcommittee for your attention to the issue. I look forward to your questions.

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

Chair HAWLEY. Thank you very much.

Next up is Professor Mike Smith. Professor Smith is professor of information technology and marketing at Carnegie Mellon University. He has written extensively on piracy and its effects on innovation. Professor Smith.

STATEMENT OF MICHAEL SMITH, PROFESSOR OF INFORMATION TECHNOLOGY AND MARKETING, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PENNSYLVANIA

Professor SMITH. Chairman Hawley, Ranking Member Durbin, I am very honored and thankful for the opportunity to testify today on this important issue. My testimony today is informed by 25 years of empirical research into the impact of new technologies on the markets—on the creative markets and my experience serving on a roundtable of 10 economists convened by the U.S. Copyright Office to study the implications of generative AI on copyright policy.

My research into piracy started in the early 2000's when digital piracy was a relatively new problem for the creative industries. During that period, many in the tech community argued that piracy was fair use because it would not harm legal sales, was unlikely to harm creativity, and any legislative efforts to curtail piracy would not only be ineffective, but would also stifle innovation.

My empirical research over the past 25 years has studied these questions. In 2020, my colleagues and I surveyed over 40 papers published in peer-reviewed academic journals as part of a piracy landscape study we wrote for the U.S. Patent and Trademark Office. Our report drew three broad conclusions.

First, the peer-reviewed academic literature shows that digital piracy does harm creators by reducing their ability to make money from their creative efforts.

Second, the peer-reviewed academic literature shows that digital piracy does harm society by reducing the economic incentives for investment in creative output.

Third, the peer-reviewed academic literature shows that copyright enforcement has been effective in reversing these harms while also allowing businesses and legal online distribution platforms to thrive and innovate.

Today, we're hearing many of the same arguments we heard in the early days of the internet. Allowing generative AI companies to use pirated content to train their models is fair use because it won't harm legal sales, won't harm creativity, and any enforcement efforts to curtail the use of pirated material for training will not only be ineffective, but will also stifle innovation.

My response to those arguments is that while the time has changed, the underlying economic principles are the same today as they were in 2000. And by applying those principles, I think we can draw many of the same conclusions.

First, the use of pirated content to train generative AI models will harm sales for creators. Allowing generative AI companies to train their models with pirated content is likely to harm markets for creators by damaging the original markets for their work, by damaging licensing markets for those works, and by creating perverse incentives for bad actors to add new copyrighted content to pirate networks, in essence, allowing generative AI companies to launder licensable content through piracy.

Second, the use of pirated content to train generative AI models will harm society by reducing economic incentives for creators. This conclusion is similar to the early piracy research: Economic incentives drive creative output. But there is a new and unique indignity to our current situation. When piracy is used to train generative AI models, we're not only stealing from creators, we're then using the theft of their content to create tools that can flood the market with machine-generated output, which in turn will replace many of those creators, particularly emerging artists.

And third, as in the early days of piracy, I believe that enforcing copyright law in the context of generative AI training can be effective at reversing these harms and can create a world where both the creative industries and the technology industries are able to thrive. If the Napster and Grokster decisions had gone the other way in the early 2000's, it is hard to imagine that Spotify and Netflix would exist today, and that would be to the detriment of consumers, the creative community, and the technology community.

I think today we have a similar opportunity to create a win-win-win for society, creators, and tech firms by making it clear that piracy is wrong and that a vibrant technology economy depends on a vibrant creative economy. We found a way to make licensed streaming and sales channels work for consumers, copyright owners, and platforms in the early 2000's. We must do the same for generative AI today.

Generative AI has the potential to benefit industry and society in many ways, but achieving that potential will require a more robust and transparent partnership between technology firms and the creative industries. On our current path, we risk killing the goose—or in this case, the authors, musicians, coders, and filmmakers—who laid the golden eggs that are key to the present and future value of generative AI output.

I thank you and look forward to your questions.

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

Chair HAWLEY. Thank you very much.

Next up is Professor Bhamati Viswanathan. Did I get that right, Professor? Am I close?

Professor VISWANATHAN. Perfect. Thank you.

Chair HAWLEY. Okay.

Professor VISWANATHAN. Perfect.

Chair HAWLEY. Professor Viswanathan is a professor of law at New England Law School, and she is an expert in AI and copyright. Thank you for being here. The floor is yours, Professor.

STATEMENT OF BHAMATI VISWANATHAN, PROFESSOR OF LAW, NEW ENGLAND LAW SCHOOL, BOSTON, MASSACHUSETTS

Professor VISWANATHAN. Chairman Hawley, senior Ranking Member Durbin, and Members of this Subcommittee, thank you so much. I am honored to testify today on a subject that I feel passionate about.

I feel that Senator Hawley did an excellent job of laying the table for us. I would like to drill down on what he's presented us with so far and help us walk through this.

So first, it's an interesting moment that we're at. Generative AI is a promising set of technologies, and I think we can all agree that they're beneficial. However, the training that they're engaging in is deeply problematic and troubling, and courts don't know what to do about this yet. They haven't reached a consensus on what should obviously be done about the training of AI on pirated works. So I would like to give us a call for action and a solution as I talk us through this.

First, we know that what pirate websites are doing is illegal. How do we know that? Multiple actions have been brought against pirate sites, and in every case, the pirate websites or repositories have lost. The FBI, the Department of Homeland Security have gone after pirate websites and tried to shut them down. Now, of course, we all know this can be like whack-a-mole, right? They shut down, they come back up again. But the point is it's well-established that what they're doing is illegal, and that makes sense.

If you and I stole books from the library or from a bookstore and said, I need to train, I need to learn, I need to develop my mind, we wouldn't argue that this is fair use. We'd say you can't steal the materials even for a good cause. That's not even what's happening here. The AI generative companies are going to pirate websites, stealing the materials that have already been stolen. It is a crime compounding a crime. How is this fair?

Say you want to go drag racing, an illegal activity. I tell you, hey, there's a shop down the street that sells stolen cars for cheap. Go buy a car and you can drag race. You go, great, that helps me be able to afford what I want to do. You buy a stolen car, you drag race, you win. Do you now get to say, hey, it's okay that I stole that—I bought that stolen car. It's okay that I engaged in illegal activity. Neither activity is legal, and one is compounding the other, and that's what's happening here. It's simple. It's a crime compounding a crime.

And it's not a victimless crime. As Professor Smith showed us, there are real victims here, the loss of author's livelihoods. Mr. Baldacci will be eloquent on this topic, but as an author myself, I feel the same. The loss of my livelihood not only hurts, but it affects what I have spent my life training to do.

It contravenes copyright laws, basic incentive structure. I don't just teach copyright, I teach constitutional law as well. This is enshrined in the United States Constitution. The Intellectual Property Clause is one of the things that makes this country not just great, but robust, powerful, economically hugely successful. Over $1 trillion in revenues from the creative content industries, this is truly at risk right now, this entire incentive structure that was brilliantly thought of by our Founders.

It has negative incentives. If you know that you can go to a pirate website and steal things, why would you ever pay for anything again? The generative AI companies have shown us the way to massive theft, not just by themselves, but by others as well. It depreciates the quality and the quantity of works out there. The tradeoff of copyright law is you, the copyright author, take the risk, and the market rewards you with rewards if the marketplace likes what you've done. There is no incentive structure anymore. That's been undermined by what's happening now.

And there's a solution. The solution is licensing. It already exists, the licensing of works, the fair compensation of creators. These are all things that actually exist now. We don't even need new legislation in some ways. We might want that as well someday, but right now we have a solution. Enforce good, standard, accepted, acknowledged licensing practices.

None of this is to say that we're against innovation. We all believe in innovation. We believe that generative AI has potential. But you cannot compromise the livelihood of creators. You cannot compromise our trove of creative activity and our entire world of art and culture and the things that we have done that make us most human and that enrich us the most—you cannot compromise those simply by saying we need new technologies to flourish. What we need is for new technologies to flourish fairly, sustainably, in ways that make sense to us and that have already been provided for by our Constitution, by the U.S. copyright law, by intellectual property law itself.

It is critical that Congress recognize that this is the tradeoff that matters for the livelihoods of everyone whose lives right now and well-being are at risk.

Thank you so much.

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

Chair HAWLEY. Thank you very much, Professor.

Next is Mr. David Baldacci. Mr. Baldacci is one of the best-selling authors in America. I don't know how many books he has had as the number one New York Times bestseller. I bet he knows. Maybe he will tell us. I have read his books. I am delighted to have him here today. He is going to tell us about AI's impact on authors. Welcome, Mr. Baldacci.

STATEMENT OF DAVID BALDACCI,
BESTSELLING AUTHOR, RICHMOND, VIRGINIA

Mr. BALDACCI. Thank you. It's a lot, number one, best-selling.

[Laughter.]

Mr. BALDACCI. I'll leave it at that.

Chairman Hawley, Ranking Member Durbin, Members of the Subcommittee, 119 years ago, Mark Twain traveled to D.C. and appeared before a Congressional Committee to advocate on behalf of copyright—stronger copyright laws. He was the most pirated author of his day. I'm pirated all over the world as well. I get why that upset him. He thought creative arts was the lifeblood of this country, and I agree with him. That was the first time at that hearing that he wore his signature white suit publicly, and he did so because he thought it represented purity of thought and spirit.

I don't own a white suit, and even if I did, I don't think my wife would have let me wear it today, so you just get blue.

Twain once said that "Travel is fatal to prejudice," meaning if you meet people where they live, you find out they're just like you. I had no chance to leave the segregated world of Richmond, Virginia, when I was growing up, but I visited the library every week, and I liked to think through books. I traveled the world without a plane ticket or a passport. And born from my love of reading came my desire to be a writer.

I worked away for decades and getting rejected over and over, but I kept going, honing my craft, remaining disciplined, taking the rejections head on, and using them as motivation, and finally I was successful. And after 60 novels under my belt, I work just as hard as I ever have. It's the American way. You work hard, you play fair, you stay the course, and you'll make it.

I truly believed that until my son asked ChatGPT to write a plot that read like a David Baldacci novel. In about 5 seconds, 3 pages came out that had elements of pretty much every book I'd ever written, including plot lines, twists, character names, narrative, the works. That's when I found out that the AI community had taken most of my novels without permission and fed them into their machine learning system. I truly felt like someone had backed up a truck to my imagination and stolen everything I'd ever created.

I'm aware of the argument that what AI did to me and other writers is no different than an aspiring writer reading other books and learning how to use them in original ways. I can tell you from personal experience that is flatly wrong.

I was once such an aspiring writer. My favorite novelist in college was John Irving. I read everything that Irving wrote. None of my novels read remotely like a John Irving novel. Why? Well, unlike AI, I can't remember every line that Irving wrote, every detail about his characters and his plots. The fact is, also unlike AI, I read other writers not to copy them or steal from them but because I love their stories. I appreciate their talent. It's motivated me to up my game.

What AI does is take what writers produce as an incredibly valuable shortcut. It's like super fuel to teach software programs what they need to know. And I have learned that these trillion-dollar companies didn't even buy my books. They got them off a website that has pirated works. They complained that it would be far too difficult to license the works from individual creators, so apparently, it was more efficient to steal it. Trillion-dollar companies with battalions of lawyers did not have the resources to do things lawfully.

I was once a trial lawyer. If I had made that argument in court, I would either have been laughed out of the courtroom or held in contempt by the judge and rightly so. If AI companies only needed words, they could have fed every dictionary in the world into their machine learning, but that was not nearly good enough because it would mean decades of additional work and hundreds of billions of dollars of additional investment. What they needed was complete, well-crafted, living, breathing stories with characters that seemed real, plots that made sense, dialog that appeared genuine, human-

ity on the page. In sum, they needed us and our craft that we learned with the sweat of our brows and the flexing of our imaginations.

And these companies have swooped in, stolen that labor in order to make enormous profits. But we, the writers, the true source of all of this, will receive nothing. AI will allow anyone, with no effort at all, to order up a novel written in the vein of an established writer. And that book can be sold saying that it reads just like a David Baldacci novel. Yes, it does read like my novels because it is my novel. It is my imagination.

People complain about cheap imported goods hurting American workers. Well, we have cheap books being created by American technology flooding the market. That will mean lower profits for publishers and less money to spend on new emerging writers. Trust me, that hurts all of us.

Online vendors now require the author to disclose if a book was not human-created. It's getting to the point where they will have to limit the number of books that someone can publish on a weekly or even daily basis. This is insane.

Source code and elements of algorithms are also protected by copyright. I would hazard to bet that if I stole any of the AI community's source codes or algorithms and then tried to profit off them, they would unleash a tsunami of lawsuits against me. However, if, as AI contends, fair use is actually my entire body of work, there is no more copyright protection for anyone. I'm sure AI believes that their IP should be fully protected against interlopers, and I agree with them. Thus, I am deeply disappointed they don't feel the same about people like me.

The AI community apparently is there entitled to steal our work product despite it being copyrighted because what they're doing is so transformational. Well, let me tell you, billions of people have been transformed by books. Many significant events in human history and in this country had seminal authors in their works that wrote at the head of the pack. We didn't truly emerge from the dark ages until the invention of the printing press when books became widely available. Books also teach empathy, making the world a kinder, gentler, more meaningful place.

I'm only one man, but books transformed my life, propelling me to a far better existence. I am sure there are aspects of AI that will also transform the world, but if you want to bet on which side is more transformational for all of us, I will bet on books every single time.

Thank you.

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

Chair HAWLEY. Thank you very much, Mr. Baldacci, very well said.

Next up and finally is Professor Edward Lee. Professor Lee is professor of law at Santa Clara University School of Law, where he has written extensively about the intersection of AI and copyright law.

Thank you for being here, Professor Lee.

STATEMENT OF EDWARD LEE, PROFESSOR OF LAW, SANTA CLARA UNIVERSITY SCHOOL OF LAW, SANTA CLARA, CALIFORNIA

Professor LEE. Chair Hawley, Ranking Member Durbin, and other Members of the Subcommittee, thank you for this opportunity to testify. I am a professor of law at Santa Clara University School of Law. I'm also a book author and a photographer, and my personal experience informs my scholarship and understanding of the importance of copyright to authors and artists across the country.

In my testimony, I will discuss whether using copyrighted works to train AI models is a fair use, giving particular attention to the two recent decisions by Judges Alsup and Chhabria in cases filed by book authors against Anthropic and Meta. This novel question of law, which has important implications for U.S. national interest, has sparked sharp disagreements among parties, stakeholders, and now Federal judges. As Judge Bibas noted in an earlier non-generative AI case, this question of law is difficult.

In my opening remarks, I would like to stress three points. First, I believe Judges Alsup and Chhabria correctly concluded that the use of copies to—the use of copies of works to train an AI model serves a highly transformative purpose in developing a new technology under factor one of fair use. During training, an AI model is exposed to vast training materials, typically many millions of works. Through a process called deep learning, the model identifies the statistical relationships among words and within subparts of words, thereby enabling the model to conduct numerous functions, including research, translation, delivery of medical advice, generation of content, and so forth.

As Judge Chhabria concluded in his opinion, "The purpose of Meta's copying was to train its large language models, which are innovative tools that can be used to generate diverse texts and perform a wide range of functions." And as Judge Alsup recognized, "The technology at issue was among the most transformative many of us will see in our lifetimes."

Now, the history of AI development strongly supports this conclusion. It is important to understand why AI researchers at universities began training AI models on large datasets. This practice originated not at AI companies, not at Big Tech, but at universities where AI researchers discovered a key insight. Scaling or using larger and more diverse datasets actually worked in developing and improving AI models, an achievement that escaped researchers for many years. This seminal breakthrough, which took decades to figure out, has propelled the advances of AI that we are witnessing today.

Second, while I agree with the ultimate findings of fair use in both cases, it's important to remember that fair use is fact-specific and decided on a case-by-case basis. In some cases, a transformative purpose in AI training might be outweighed by the other factors. For example, an AI model that routinely produces outputs that are infringing, such as regurgitations, might not be a fair use even in the training of the model due to insufficient guardrails on the model.

Critically, in the cases against Anthropic and Meta, the judges concluded the plaintiffs did not show the models had produced any infringing outputs of the plaintiff's works. And that can be appealed, but that is the findings of both judges.

My final point is the need for caution, caution by the courts, caution by Congress, and the States. I believe it's important to weigh the United States' interest in AI innovation. President Trump has issued an executive order making U.S. development and global leadership in AI a national priority. China has its own priority and a plan of surpassing the United States and becoming the world leader in AI by 2030. The United States' national priority in AI counsels caution.

Indeed, in *Google v. Oracle*, another technology fair use case of national importance, the U.S. Supreme Court itself cautioned, "Given the rapidly changing technological, economic, and business-related circumstances, we believe we should not answer more than is necessary to resolve the parties' dispute." Judges Alsup and Chhabria departed from this approach in some controversial parts of their opinions that were just dicta. I disagree with Judge Alsup's suggestion on pirated books and Judge Chhabria's suggestion on copyright dilution, as more fully elaborated in my written statement.

At this juncture, I think the best approach is for Congress to wait and see how other district courts, the courts of appeals, and potentially the U.S. Supreme Court resolves these difficult issues, including access to pirated shadow libraries in the many pending copyright lawsuits across the country.

Thank you, Senator.

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

Chair HAWLEY. Thank you very much, Professor. Thanks for being here. Thanks again to all of our witnesses.

We are going to now have 7-minute rounds of questioning, and we will see if we can fit in maybe a couple of rounds, just depending on the time that we have. I will start, and then we will go to the Ranking Member and any other Members who arrive in that time.

Professor Viswanathan, let me just start with you, if I could, and let's see if we can just drill down on some of the specifics here. Mr. Baldacci mentioned in his opening statement that AI could just feed dictionaries into their platforms in order to train them. They don't do that. They prefer published works, fully formed works. Why is that? Can you give us an insight into that?

Professor VISWANATHAN. That's absolutely right. They learn syntax, structure. They learn how we learn language, right? When you learn language, you just don't learn words. You don't memorize words. You don't memorize notes when you learn music. You learn structure and syntax. And the point that Professor Lee is making is correct. They need large datasets. More is better to learn predictive language models. However, more is not everything. It's not pirated works.

Chair HAWLEY. So let me just ask this. You said that they are not buying the books. They are not buying Mr. Baldacci's book or anybody's book who is sitting up here, anybody in the audience.

They are getting them. They are stealing them. They are pirating them from somewhere. If they are not buying the books, they are not stealing them out of libraries, where are they getting them?

Professor VISWANATHAN. These large repositories of materials that are available online, there are many. Some are licit, some are not licit. The pirate websites in particular are not licit. So if you need a lot of material, you go out and you scoop up all that material that you can find, but you don't go to pirate websites to get that material if what you want to do is legal. None of these works are licensed. None of these works are licensed. No author has been compensated to date.

Chair HAWLEY. They go to these—let's call them shadow libraries—to get the works illegally. By the time they go to the shadow library, the works there are already stolen, right? They have already stolen Mr. Baldacci's book, Professor Lee's book, everybody's, your books. They have stolen them. When they go to the shadow library, how do they get them? I mean, how does the AI company then take possession of the particular work?

Professor VISWANATHAN. There's a process called torrenting, and I will not trouble you all with the details of torrenting, but essentially huge amounts of data streamed to you and you get them. At the same time, you can send them out. That's called seeding. You can send them out at the same time. Uploading and downloading exists at the same time. This is a peer-to-peer process. So not only are you taking in these pirated materials, you are also distributing them. The violation of copyright law exists at the reproduction of these works, at the making available of them by the pirate libraries, the dissemination of them, and your dissemination gen AI company of them as well.

Chair HAWLEY. So they are both taking the works and distributing them as well in this thing called, kind of like Napster, this thing that you call torrenting. Let me ask you this. I mean, is torrenting legal? That is not legal, is it?

Professor VISWANATHAN. Torrenting can be illegal, but in this case, it is not. And in this particular case, this is benefiting the—now I agree with Judge Alsup who said, if you're taking it from pirate libraries, no way. That is not acceptable, right? Part of what we're seeing here, Judge Chhabria said, well, it's not helping the pirate websites. Well, yes, it is. The pirate websites, there's one in particular called Anna's Archive. They actually put on their website, hey, gen AI companies, come train on us. We'll do some data swaps. Or, you know what, you can make us a donation too. This is directly helping the pirate websites thrive, flourish, proliferate.

Chair HAWLEY. Let me ask you this. Have there been, to your knowledge, any criminal enforcements against these torrenting platforms?

Professor VISWANATHAN. Yes, there have been attempts to. Again, it's like a game of whack-a-mole. You get one, you knock it down, it pops up again in some jurisdiction that you don't have control over.

Chair HAWLEY. What is the key to criminal enforcement? You know, civil versus criminal in this context, when do we have a criminal case against torrenting? What is the key to that?

Professor VISWANATHAN. Okay. This is a really important point. What's criminal here? Criminal copyright liability has two prongs to it. Prong one is you have to do it willfully, and prong two is you have to do it for commercial advantage or gain. We clearly know that prong two is met. This is for commercial advantage or gain. I don't think Meta is doing this out of the goodness of its heart. Prong one, willful means you need to know that what you are doing is illegal. There's lots and lots of evidence now, particularly from the *Kadrey v. Meta* case, that shows that they knew this was illegal. They even had to ask all the way up the chain of command to Mark Zuckerberg and say, hey, is this okay? And he said, yes, it's okay.

So not only did he do it knowing it was illegal, he did it knowingly, he did it willfully, intentionally, and whether or not he knew what statute it was legal doesn't matter. For this to be willful, you have to know that what you're doing is wrong, and this meets that prong. So this is, in fact, amounting to what you might call criminal copyright liability.

Chair HAWLEY. Mr. Pritt, let me just ask you about this, about the willful aspect, and let's talk about Meta in particular, since Professor Viswanathan just mentioned Meta. They are one of the biggest monopolists in the world and one of the biggest AI companies now in the world, if not the biggest. So let's just talk about them for a second. Meta uses torrents to acquire pirated data for its Llama model, is that right?

Mr. PRITT. Correct.

Chair HAWLEY. How much data would you estimate that Meta has torrented? It is illegally downloaded and also then shared in this peer-to-peer scheme.

Mr. PRITT. It has pirated well over 200 terabytes of copyrighted material from multiple—I don't call them shadow libraries because they're not libraries—but illicit criminal enterprises.

Chair HAWLEY. And how much has it paid the copyright holders for these works that it has used, to your knowledge?

Mr. PRITT. Nothing.

Chair HAWLEY. Nothing, zero. So billions of works, billions of books like Mr. Baldacci's, zero payment. If Meta were to pay, do you have any idea what the cost might be? I mean, to your knowledge and your discovery, did they ever explore paying? I mean, is there any sense of how much this might have cost them?

Mr. PRITT. Early on, they explored licensing. They assigned two individuals part-time to attempt to license, and they decided it would take too long, for example, and that's when they turned to piracy. At the time, they had public documents show that certainly tens of millions, if not hundreds of millions, had been contemplated for licensing at that time.

Chair HAWLEY. Okay. So let's just think about this. Hundreds of millions of dollars, that is the value, maybe sort of the base, the bare value of the works that they have used, like the works that you all have written on this panel, hundreds of millions, and they paid zero of that.

So let's just drill down a little further. Did Meta know what they were doing was wrong? Do you, Mr. Pritt, believe in the evidence

you have seen that there is any evidence to suggest that Meta's employees knew what they were doing is illegal?

Mr. PRITT. I think the documents that have become public clearly show that.

Chair HAWLEY. Let's just look at a few of these documents. I am going to show you a few things, and I will ask you to help me interpret them to make sure that we get them right. Let's start here with a Meta employee, a Meta engineer working on their AI project, Eleonora Presani. She says, "I don't think we should use pirated material." This is in a chat with other Meta employees. "I don't think we should use pirated material. I really need to draw a line there." She goes on, "I feel that using pirated material should be beyond our ethical threshold. Sci-Hub, ResearchGate, LibGen are basically like Pirate Bay or something like that. They are distributing content that is protected by copyright, and they are infringing it." How do you read this, Mr. Pritt? Does this look like knowledge to you?

[Poster is displayed.]

Mr. PRITT. That's certainly what we've argued in the case.

Chair HAWLEY. Let's look at another Meta employee. Here is Nisha Deo in the same chat. She replies and said, "It's the piracy (and us knowing and being accomplices) that's the issue." This is a Meta engineer working on their AI project. "It's the piracy (and us knowing and being accomplices) that's the issue."

[Poster is displayed.]

Let's look at another one. Here is the response that another Meta engineer in the same chat gave. "Well, we want to buy books and be nice, open people here. But, however, to make it happen and not letting the bad guys win"—that's the beat-China argument—"we need to make a case—fast—and cut some corners here and there." "We need to cut some corners here and there." Mr. Pritt, what are we looking at here? I mean, is this knowledge of illegal activity?

[Poster is displayed.]

Mr. PRITT. When they refer to bad guys, I think they're actually referring to OpenAI and other AI competitors.

[Laughter.]

Mr. PRITT. But yes, this is certainly one of the many documents that show that they knew these were pirated websites that contained copyrighted materials, and they were taking them for free.

Chair HAWLEY. So here we have it in black and white. Don't believe me. Read the evidence. These are Meta's own engineers, Meta's own employees saying, they know what they are doing is ethically wrong, illegal, likely to subject them to legal liability, and they are doing it anyway because they need the money.

There is a lot more here. We will come back to this. I want to give Senator Durbin a chance to ask questions. Senator Durbin.

Senator DURBIN. Thanks, Mr. Chairman.

I want to ask startup questions with Mr. Baldacci. A number of authors have shared with the public the process they go through to write a book. I believe John Irving in The Imaginary Girlfriend did that. I think John McPhee has done that in the past. Stephen King has done that. Give us a kind of an insight, now that you have published successfully in volume, what the process is in writing a novel.

Mr. BALDACCI. Well, you know, one, you have to sort of be in love with words and storytelling because that is sort of the essence of what you're trying to create. You draw upon personal experiences, your own curiosities, people you've met along the way, things that have happened to you, places you've traveled to, humanistic experiences that a software platform really can't replicate. And if it ever manages to do it, I would like another planet to live on, quite frankly.

And for me, it was 20 years of hard work learning the craft before I ever was published at all. I started writing short stories and wrote them for 15 years when I was in college and law school and tried to get them published and was not successful. But it's a craft that you build over time. And you have a lot of frustration, a lot of dips and valleys. Good times happen, bad times happen, rejections happen. You learn from them, you keep going. And at the end of the day, hopefully, you get good enough to where someone who has the ability to make your career happen will read your material and respond to it, and you can then maybe hopefully write for a living. And that's what happened to me after a long period of incubation.

You never really see a lot of young writers—you know, you're not going to see a lot of teenage writers making it big because writing is about life, and you have to have something to sort of write about. And it takes a long time. And that is why I felt when my son brought this up where every single one of my books was presented to me in an outline in like 3 seconds, it really felt like I had been robbed of everything my entire adult life that I had worked on now was in the possession of someone else that someone else I didn't even know could then use to write their own books that are actually my books. I mean, that's not supposed to happen in this country.

And that's what was so enraging to me that I—I license my work all over the world. I license it for different foreign publishers. I license my work for television and movies and all types of endeavors. And I am open to any offer. If someone comes to me and wants to license my work, I will listen to them. If we can negotiate something that's agreeable to both parties, I will do it, and they can use my work for the parameters that are in the licensing agreement, and life can go on and people can be happy.

But the uncertainty of like stealing stuff from pirated sites operated in Russia just so you can gain an advantage and you don't really care about what happens to the likes of me and other writers coming up—I make a lot of money from my publisher, and my publisher has used that money to take risks on new writers coming up they ordinarily would not have been able to take a risk on. So when you hurt established writers like me, you hurt all the other writers coming behind us.

Senator DURBIN. So when you are in the creative process of writing novels and other things, are you policing against plagiarism?

Mr. BALDACCI. I get—I am pirated a lot, but I never worry about that because my ideas are my ideas. And I—nobody has the sort of mindset and the experiences that I have, nor do I have the mindset and experiences of other people. It is very individualized.

I never worry about that I'm going to inadvertently take something away from another writer because my stories are my own.

And that's why a software platform, the only thing they can do is take from what has already been created. They can't create anything really on their own. They take my mishmash and put it all together and throw it out the other end, but it still looks like my stuff because it is my stuff.

Senator DURBIN. Professor Lee, if I understand part of your argument here, you were suggesting that this is the age of innovation. Deep learning deserves special treatment. We've been through this argument in Congress before. Section 230 is a good illustration of that. We decided this fledgling industry called the internet just may not have a future, better be careful, so we exempted them from liability. Is that what you are suggesting?

Professor LEE. Not at all, Senator. My position is that we should pay heed to the existing Supreme Court precedence on fair use, which repeatedly states that fair use is a flexible doctrine decided on a case-by-case manner. And there is a way for authors to prove market harm based on a taking or the copying of protected elements of their works.

Judge Alsup said, if the authors show that there is market harm based on an output of this model, you could bring another case. And that's exactly, I think, the approach to strike the correct—as you mentioned earlier at the opening remarks—to strike the right balance between protecting copyrighted works and authors and protecting innovation. Even just a story in *Emerson v. Davies* recognized that not everything in a book is protected by copyright. Authors build on the past books to write new books, and that fuels creation.

And here, the line that Judge Chhabria and Alsup drew in terms of non-infringing output—or excuse me, just Judge Alsup—there is no copyright claim in the production of non-infringing works.

Senator DURBIN. I am sorry to interrupt you, but I only have a minute left. It looks to me like you are shifting the burden to the author of the creative work when there is an assertion of fair use here. So Meta or others can virtually steal this creative product of Mr. Baldacci and others, and then he has the responsibility of proving that there has been an economic loss to him as a result of it?

Professor LEE. Not at all, Senator. The judges explained in their opinions that the—yes, the initial burden for fair use is on the defendant, but the defendants in both cases provided evidence that there was no output of infringing works. And the question then becomes, will the plaintiffs present contrary evidence? And neither judge found evidence of outputs that had substantially similar copies of the plaintiff's works. So the entire——

Senator DURBIN. So, ultimately, the thievery, if you want to use that word, of the creative work is for the economic benefit of those who are creating the AI, is it not?

Professor LEE. Not necessarily. I think if the plaintiffs are able to prove cognizable market harm from the copying of their copyrighted expression, then the fair use argument is likely to fail for their training.

Senator DURBIN. I am coming at it from a different angle. I am talking to you about why do we have AI? Why are we interested in AI? Clearly, it is a commercial purpose, is it not?

Professor LEE. Oh, entirely. For the AI companies, yes.

Senator DURBIN. For the companies. So that they are ultimately the winners in this approach that you are taking. We assume we are in the world of new innovation here, and there is a use of someone else's creative work. The burden is on them to prove that they have lost money because of that piracy. But the ultimate winner in this is going to be the AI because if they escape this responsibility, they can use Mr. Baldacci's product and make money off of it.

Professor LEE. Yes, if the training is considered a fair use, the direct benefit would be to the AI companies. I grant that. But in terms of the larger national interest, it redounds to the benefit of the United States. If we have a priority in AI development, and if we are in a competition or arms race with China, winning the AI race by United States companies benefits the United States, in my view.

Senator DURBIN. And Mr. Baldacci should be prepared to pay the price for that, right?

Professor LEE. Well, I would suggest that if it is so easy to generate copies of Mr. Baldacci's novels or any other authors, that should go in the complaint in these lawsuits. And some of the lawsuits do allege infringing outputs. So those are yet to be resolved. But my ultimate position is that we should not throw out the window the established Supreme Court precedence on how to apply fair use. It is case-by-case, flexible, and it balances the interests of both sides in terms of copyright, as well as innovation.

Senator DURBIN. Thank you.

Chair HAWLEY. I just want to followup on this line of questioning, Professor Lee. When you say that it would be to the benefit of the United States, isn't Mr. Baldacci a citizen of the United States?

Professor LEE. Entirely. I'm not saying that Mr. Baldacci does not benefit from the copyright. There is another——

Chair HAWLEY. But let's take a different author, Professor Viswanathan. She is a citizen of the United States?

Professor LEE. Yes.

Chair HAWLEY. So I am just struggling to understand, when you say that the mass theft of their works will benefit the United States ultimately, you are saying that the mass theft and potential impoverishment of American citizens ultimately redounds to the good of America?

Professor LEE. Not at all, Senator.

Chair HAWLEY. I think you are being a little too imprecise, right? What you mean to say is it may benefit American corporations. It may impoverish American citizens, but it will benefit American corporations.

Professor LEE. Well, Senator, there is a balance to be struck and the courts——

Chair HAWLEY. Well, indeed, but you are waving the magic wand that this will benefit the United States, said we are in an arms race with China. I am just trying to drill down on your assertion.

I think what you are really saying is is that the enrichment of certain multinational corporations that are incidentally based in the United States taking the works and personal property of American citizens is a good thing. That is a little bit less clear to me.

Professor LEE. Well, the way that I view the national interest, as stated by President Trump's executive order, is that there is a national priority in maintaining the United States' dominance and leadership globally in AI. And I would defer to the view of the AI czar, David Sacks, who said if there is no pathway to fair use in AI training, we will lose the race with China.

Chair HAWLEY. Well, you think that we should allow an unelected AI czar to decide what the rights of American citizens are?

Professor LEE. No, not at all. This is going through the courts. I would let the courts decide all of these disputes. And there are presently 44 lawsuits around the country, so this is not a time for Congress to intervene in terms of deciding these very difficult questions.

Chair HAWLEY. It just sounds strange to me to say that the United States, as a nation, is going to benefit from the mass violations of its citizens' rights. I thought what made us a nation was our common citizenship, the things that we agree on together, the rights that we hold in common. And your argument seems to be it is fine to violate those rights en masse if it redounds to the benefit of the Nation. I think what you are really saying is to the benefit of certain people in the Nation and their immediate interests.

Let me ask you about something else you said, fair use.

Professor LEE. Can I respond?

Chair HAWLEY. Well, just a second. I have limited time here. Fair use, you said, is a flexible doctrine. It is an equitable doctrine. And these companies aren't exactly coming to this with clean hands, are they? They are coming to claiming fair use after they have stolen Mr. Baldacci's work. They didn't take it from the library. They didn't license it. They didn't buy it. They went to a pirated illegal site and took it. And now they are coming and claiming the cover of equity. That seems kind of strange, doesn't it? Is that how equitable law works?

Professor LEE. That is the very question, the initial acquisition, whether that was justified as fair use. And the two judges disagreed on how to treat that initial acquisition from the shadow libraries. So I think it would be incorrect for us to assume that it is necessarily a violation. And the Supreme Court in *Google v. Oracle* had an opportunity to discuss or require considerations of bad faith in the fair use analysis, and it rejected that opportunity and even cited Judge Leval's very influential fair use article saying that fair use is not limited to the well-behaved.

Chair HAWLEY. Okay.

Professor LEE. Now——

Chair HAWLEY. We appreciate you being here, and thank you. You are making these arguments very gamely. That is helpful, I think, to have this debate. But I just want to point out that there is a lot of hand-waving going on here. Every time we get down to the nub of the question, can these giant corporations take the copyrighted work of individual citizens, we get distracted with, well, it

is for the good of the country, maybe it is not so bad, we have an arms race on, there is an AI czar. Actually, I don't think it is that complicated. I think it is pretty simple. I think in America, we have rights. Those rights are what protect us. These rights are being violated. And if we are going to succeed as a nation and uphold our principles as a nation, we better darn well enforce the individual rights on which the nation is founded. I mean, it is just a thought.

Senator Welch, am I catching you off guard?

Senator WELCH. I was kind of enjoying it.

[Laughter.]

Chair HAWLEY. Well, you are welcome to ask questions if you would like.

Senator WELCH. I would hate to step on anyone, but especially a colleague Senator and the Chair of the Committee, you know, mid-expression of righteous outrage and indignation with which I am aligned, so thank you very much. Thank you. And I appreciate you calling this hearing because this is incredibly important.

You know, Senator Blackburn and I have a bill which is called the TRAIN Act, and it is trying to address this question of artistic content being used. And, you know, we have got a celebrated author here, and it would protect you. But what I appreciate about you being here, Mr. Baldacci, is there is a lot of folks who are aspiring to be David Baldacci. There are a lot of artists aspiring to be a Taylor Swift. And it is the folks who have made it that are in a position to advocate. And it is not, I don't think, going to benefit you, but it is going to benefit artists who have so much to contribute even though they are not yet discovered.

And, you know, this is the reality, and this is where I think the Chairman is really right. The AI companies need content, so they don't care where it comes from. It is just a voracious, insatiable appetite. And they are going to go into copyrighted material. We just know that. And to suggest they won't I think is naive. And the question and the burden here is that is going into copyrighted material. And the artist has the right to have that copyright respected.

The burden is that how do you know they used it? That is the whole point of the TRAIN Act where if there is copyright infringement, a reasonable assertion of that and suspicion of it is going to require disclosure on the part of the AI platform.

So I wanted to ask a little bit about that. And I will start with you, Mr. Baldacci. Do you have any suspicion that some of your works have been used to train AI systems?

Mr. BALDACCI. I have been told and I have been shown a data base, and it's part of the—part of a class-action lawsuit against the AI community. And I think they've conceded that they've taken at least 44 of my novels and fed them into their large language models.

Senator WELCH. I mean, that is astonishing. Literally, you have got 44——

Mr. BALDACCI. Well, at least they didn't take them all, so that was nice.

[Laughter.]

Senator WELCH. Just wait.

[Laughter.]

Mr. BALDACCI. I know.

Senator WELCH. And so you don't know for sure, and the only way you are going to find out is hopefully through this class-action litigation that you are part of.

Mr. BALDACCI. Well, I certainly learned that when my son put in ChatGPT that ChatGPT was intimately familiar with my entire body of work because it was able to throw out, you know, plotlines that took from many of my novels, so someone had to feed my novels into ChatGPT. Otherwise, it could not have created that response.

Senator WELCH. And we just can't allow that. You know, that is just really wrong. Thank you. So we are in agreement here that we need some reforms here to protect the artist.

Mr. Smith, you know, music, it is the same situation. And, you know, our music industry is so important. Using the word industry is wrong. Music is so important. It really helps people get a sense of who they are, it helps people connect, and it is across political divisions. That is what is one of the inspiring things about the incredible contributions that musicians provide to our society. And can you just explain what the dangers are of allowing AI models to freely train off copyrighted works?

Professor SMITH. Sure. There are multiple dangers. What we have seen in the early piracy research is that Article I, Section 8, Clause 8 is actually a really good idea. Giving artists incentives to create actually yields more creation. And when artists' incomes are lowered through piracy, they have lower incentives to create. I think we see the same thing here, both directly by participating in these pirated networks, the generative AI companies are making it easier for other people to steal. But then indirectly, they're also making it harder for licenses to be signed. Mr. Baldacci talks about signing licenses, but when you sign a license with a generative AI company, you're signing with a gun held to your head because they can say, either sign what I'm offering or I'm going to go steal it instead.

Senator WELCH. Well, that is the adhesion contract that good lawyers like Senator Hawley still remember from law school days. No, but explain that a little bit more because, you know, this is where I think all of us have some real appreciation for young artists. They have a vision that there is something inside them that they can express and that it will make a difference to people who hear it or people who read it. And they start out against their parents' will most of the time, right, because it is not an income-producing, promising career, and a lot of them don't succeed, commercial success. But they actually are contributing in a local community to a sense that helps develop our culture and helps create respect for the creative process and helps create respect that there are other things than the career path that some of us up here have followed where you can make a real contribution and a meaningful contribution.

So this is the concern I have about how this AI and the grabbing is going to make it tougher for those folks against great odds to keep at it. So maybe you could just, from your experience, talk a little bit about how it would adversely impact any chance they have

of being able to pay their bills at the end of the month while they are trying to create inspirational music for the benefit of all of us.

Professor SMITH. Yes, I deeply share that concern, Senator, and it's based on peer-reviewed academic research showing that creative output goes down when piracy is allowed to flourish. I worry that the future David Baldaccis of the world won't get through that hump, and we won't get to appreciate their creative output if we allow piracy to continue to be used to train these generative AI models.

Senator WELCH. Well, thank you. My time is just about up, but I just want to express my gratitude to each of the witnesses. I didn't have a chance to speak with you, but I think this is an extraordinarily important issue.

I yield back.

Chair HAWLEY. Thank you, Senator Welch. Senator Durbin.

Senator DURBIN. Mr. Pritt, you represent plaintiffs in a lawsuit against Meta that alleges copyright infringement of the plaintiffs' authors' works. Do you have any idea how much Meta as a company is valued?

Mr. PRITT. That's a good question. Many trillions, I believe.

Senator DURBIN. Did Meta compensate any of the copyright owners in your case for the use of their works?

Mr. PRITT. No, but Meta did spend money on contributing its processing power to pirate from illicit websites and also to pay Amazon to host pirated data.

Senator DURBIN. Which, of course, did not inure to the benefit of your plaintiffs.

Mr. PRITT. Certainly not.

Senator DURBIN. How does the downloading and uploading of pirated copyright material impact the analysis of whether a copyright infringement could meet the mens rea requirement or willfulness necessary for criminal infringement?

Mr. PRITT. I would let the professors answer that question. Certainly as to willfulness in the civil copyright context, as the documents Senator Hawley showed, I think the answer is clear, that the piracy committed by Meta was knowing and intentional.

Senator DURBIN. Anyone else want to comment on that? Mr. Lee, Dr. Lee?

Professor LEE. Yes, thank you, Senator. The standard of willfulness for criminal copyright infringement requires knowledge that it is illegal to engage in that particular copying. Now, I don't want to relitigate what Judge Chhabria has already ruled on, but he was given all of this evidence that was submitted by Mr. Pritt and his colleagues. He saw the comments by engineers, but he also saw comments and analysis by lawyers of Meta advising them on whether this was permitted or not under fair use law. And Judge Chhabria made a determination. The crime fraud exception simply didn't apply.

And I'm not privy to all of the analysis that Judge Chhabria made, but I'm assuming it was based on the question not being resolved, the legal question of whether accessing or copying from a pirated website to serve a highly transformative purpose is the very question raised in the lawsuit. There was no prior precedent that has so held that it is piracy or illegal, let alone criminal in-

fringement, to do that. And that is the very question that Judge Chhabria ruled on. And to assume that it is piracy is begging the question—with all due respect, it is begging the question that the courts are the appropriate determiners of.

And that can be appealed, you know, and I am sure it will be appealed, but here the question of whether acquiring for a putative fair use purpose is unlawful, Judge Chhabria ruled it was not. It was for the fair use purpose of developing the AI model. I believe that is supported by the text of Section 107.

Senator DURBIN. So Professor Viswanathan, would you like to comment on that?

Professor VISWANATHAN. I would, thank you so much. The very fact that we're talking about this kind of behavior as to whether or not it's criminal, right, the very fact that we're here talking about willful, knowing, intentional, massive scale training on pirated materials. Let's just step back for a moment from the question of whether it comes under criminal copyright infringement. Does it come under fair use at all? Is this what fair use was developed to be? Fair use, for those of you who don't take my copyright class, sorry about that, fair use is an affirmative defense. Yes, I infringed, but I did it for a good reason, a societally beneficial reason.

All right. Maybe creating a world's repository of generative AI companies is that, but it doesn't seem to me that it squares with the other things that we think of as fair use. What's well-established fair use? Education, criticism, commentary, First Amendment purposes that we consider valuable and necessary and that are done in good faith. I educate in good faith. I don't want to have to clear all those copyrights to educate. Okay, great, we allow you to do that.

That is not what's going on here. I don't want to relitigate the cases, Professor Lee, but Judge Chhabria was clearly distressed by this. And when he raised the possibility, as you rightly say, in dicta, that market dilution might be what's happening, he's saying, look, exactly what the Senator was talking about, flooding—what Mr. Baldacci was talking about, flooding the market with subpar works that substitute for the original works. This is not what fair use was intended to achieve or to facilitate.

And the very fact that these companies are arguing we're in good faith, we're doing fair use purposes, to me, this shouldn't even be a defense that they're allowed to raise. But okay, they will raise it, and it will be litigated. But boy, it just does not seem consonant with what fair use was ever meant to do.

Senator DURBIN. Thank you. Thank you, Mr. Chairman.

Chair HAWLEY. Mr. Pritt, if I could just ask you another question or two about some of the evidence. We talked about Meta engineers saying that they realized what they were doing was crossing an ethical line, that they felt they shouldn't be doing it, but they had to cut some corners. Let me just ask you, did Meta ever try to hide what it was doing? Did it try to hide the fact that it was pirating these works?

Mr. PRITT. What the documents show is that in 2024, when Meta began to use Anna's Archive, it decided intentionally to not use its own servers and instead to go through Amazon Web Services in

order to ensure that the seeding, the sharing of pirated works would not be traced back to Meta's own IP.

Chair HAWLEY. It doesn't sound to me like a company and executives that think what they are doing is above board. It sounds like a company that thinks that what they are doing is probably illegal in some manner, but they want to go on doing it anyway.

Let me just show you a couple of documents, help us understand what we are seeing here. These are more Meta engineers now, again, working on AI. We have got the first one, Nikolay, who says, "not sure we can use Meta's IPs to load through torrents pirate content, haha."

[Poster is displayed.]

[Laughter.]

Chair HAWLEY. I emphasize, these are their documents. I mean, for all of Professor Lee's—and again, I appreciate Professor Lee making these arguments, but for all of Professor Lee's comments that we are not sure if it is really pirated or not, they thought so. This is Meta. Meta thought so. The next employee, "I'm curious to start looking at some samples, but I feel like we should get some clarity on what's allowed and how," smiling emoji. Nikolay again, "haha, yes, I think torrenting from a corporate laptop doesn't feel right."

[Poster is displayed.]

I mean, what are we looking at here, Mr. Pritt? I mean, is this an attempt to be above board and forthcoming, and, you know, they think everything's fine?

Mr. PRITT. I think that is a very difficult conclusion to draw from these documents. And with all due respect to Professor Lee, as I am still litigating the case against Meta on behalf of a group of authors, Judge Chhabria in that case specifically declined to decide whether Meta's piracy, what it has engaged in, in terms of the downloading, the making available, the making additional copies, and then sending those copies, over 40 terabytes of data, to other individuals, is in fact fair use. And no court, including the Supreme Court, has ever held that rank piracy is somehow fair use. And instead, the Supreme Court case law, still the law of the land, says that fair use presupposes good faith and fair dealing. I will leave it to you whether or not you think any of these documents shows good faith and fair dealing.

Chair HAWLEY. Well, let's just look at one other document and ask ourselves if this looks like good faith and fair dealing. More Meta employees, more AI engineers. "Frank, can you clarify why we can't use Facebook infra"—internal—"for this again?" Frank Zhang replies, "avoiding risk of tracing back the seeder from a Facebook server." And he clarifies, "avoiding risk of tracing back the seeder/downloader are from Facebook servers." So here we have Meta employees saying they know they are pirating, they think it is ethically wrong, they think it is illegal, and they are actively avoiding trying to create a paper trail. They are trying to hide it. I mean, that doesn't sound like fair use to me. Does it sound like fair use to you, Professor Lee? I mean, do you think this is fair use?

[Poster is displayed.]

Professor LEE. I would just say I agree with Judge Chhabria's approach. The distribution claim is still alive in the case, and this aspect of the torrenting may well be infringement and not fair use.

Chair HAWLEY. I will just say this. If this isn't infringement, Congress needs to do something. I mean, if the answer is that the biggest corporation in the world worth trillions of dollars can come take an individual author's work like Mr. Baldacci, lie about it, hide it, profit off of it, and there is nothing our law does about that, we need to change the law. And if nothing else comes out of this hearing today, I hope that is it. And I hope that this is motivation to this body that we need to be paying attention to what is going on here.

Mr. Baldacci, you said you would rather live on a different planet if there was AI that could write your books. I am sure that that will never happen. They will never write your books. I want to live on a different planet if this can go on and it is perfectly legal. We have got to do something about this.

[Applause.]

Chair HAWLEY. Let me just ask you, Mr. Pritt, finally, what about Mark Zuckerberg in all of this? I mean, do we think that Zuckerberg knew about this, approved this? I mean, what does the evidence suggest?

Mr. PRITT. Certainly, the documents that have become public in the case explain that the decision whether or not to use Library Genesis, which is a notorious illicit marketplace, for example, for actual training as opposed to exploration was escalated to Mark Zuckerberg.

Chair HAWLEY. I think the judge said something to this effect—let's just look here if we have got it—that in fact, Zuckerberg was asked about it. There it is. In the spring of 2023, after failing to acquire licenses and following escalation up to Zuckerberg, Meta decided to just use the works acquired from a torrenting platform as training data. So they just did it anyway. They just, yes, you know, do it anyway. Forget it. Don't pay Mr. Baldacci. Don't pay anybody. It costs too much. A lot cheaper to take it for free and then make billions of dollars off of it.

[Poster is displayed.]

Listen, I will just conclude with this. I want to thank all the witnesses for their testimony. And Senator Welch, if you have more questions, or Senator Durbin, I am happy to let you ask those.

For my part, I just want to say, I think that this is a moral issue as much as anything else. I think this is an issue about who are we going to be as a country? Are we going to be a country, as it is written into our Constitution, where we protect the rights of our citizens? It is part of what makes us Americans. And we welcome the creative genius of people like Mr. Baldacci and the marvelous diversity of imagination and viewpoints and perspectives that has come to characterize our country. Are we going to protect that? Are we going to allow a few mega corporations to vacuum it all up, digest it, and make billions of dollars in profits, maybe trillions, and pay nobody for it? That is not America. That is not our country. It never has been.

Listen, I am all for the free market. I am glad Mark Zuckerberg can make his billions. That is fine. But not by running over people

like Mr. Baldacci or anybody else or any young author who is trying to get a start or any other person, creative, noncreative, or just a working guy who puts something on Facebook. Why should all his stuff get taken? I just think that is wrong. I think it is morally wrong. I think, frankly, it is not consonant with our principles as Americans, and I think we can and should do better than that.

Senator Welch, Senator Durbin?

[No response.]

Chair HAWLEY. I want to thank again the witnesses for being here. Thanks to each of you. I know you had to travel far for this. And thank you again for accommodating our schedule. Thanks to everyone who has been here today.

And with that, we will stand adjourned.

[Whereupon, at 1:25 p.m., the hearing was adjourned.]

[Additional material submitted for the record follows.]

United States Senate
Committee on the Judiciary
Subcommittee on Crime and Counterterrorism

Statement of David Baldacci
July 16, 2025

Mark Twain once said that travel is fatal to prejudice, meaning if you meet people where they live, you find out they're just like you. However, I had no chance to leave the segregated world of Richmond, Virginia when I was growing up. But I visited the library every week and I like to think through books I traveled the world without a plane ticket or a passport. And born from my love of reading came my desire to be a writer. I worked away for decades, getting rejected over and over. But I kept going, honing my craft, remaining disciplined, taking the rejections head on and using them as motivation. And finally, I was successful. And after sixty novels under my belt, I work just as hard as I ever have. It's the American way. Work hard, play fair, stay the course and you'll make it.

I truly believed that until my son asked ChatGPT to write a plot that read like a David Baldacci novel. In about five seconds three pages came up that had elements of pretty much every book I'd ever written, including plot lines, character names, narrative, the works.

That's when I found out the AI community had taken most of my novels without permission and fed them into their machine learning system.

I truly felt like someone had backed up a truck to my imagination and stolen everything I'd ever created.

I'm aware of the argument that what AI did to me and other writers is no different than an aspiring writer reading other books and learning how to use them in original ways.

I can tell you from personal experience that is flatly wrong.

I was once such an aspiring writer. My favorite novelist in college was John Irving. I read everything that Irving wrote. None of my novels read remotely like an Irving novel. Why? Well, unlike AI, I can't remember every line that Irving wrote, every detail about his characters, and his plots. The fact is, also unlike AI, I read other writers not to copy them but because I loved their stories, I appreciate their talent, it motivated me to up my game. What AI does is take what writers produce as an incredibly valuable shortcut to teach software programs what they need to know.

And I have learned that these trillion-dollar companies didn't even buy my books. They got them off websites that have pirated works. They complain that it would be far too difficult to license the works from individual creators. So apparently it was more efficient to steal it. Trillion-dollar companies with battalions of lawyers did not have the resources to do things lawfully? I was once a trial lawyer. If I had made that argument in court I would either have been laughed out of the courtroom or held in contempt by the judge. And rightly so.

Keep in mind that copyrighted books don't simply end up in AI training datasets as part of some indiscriminate sweep of the internet. Complete books are not posted online by their copyright owners like website content, blogs, news articles, or other text. Books are unique in that they are sold as digital files that include technical protection measures against copying and downloading through online retailers like Amazon, Barnes and Noble, Kobo, and others, to be read on digital devices. So the only way for the AI companies to access free books online was through pirate websites, virtually all of them based abroad, in Russia, Ukraine, and other countries outside the reach of U.S. law enforcement. And it was not an isolated instance of one bad actor—every major large language model in commercial use today was trained on pirated books, apparently with the full knowledge and authorization of the companies' highest decision-makers. This is the largest criminal-

level[1] copyright infringement ever perpetrated in this country and it was committed by some of the wealthiest companies in the country. The pirate sites they used are among the most notorious on the web that authors and publishers have been trying to get shut down for nearly a decade without success. Why? For one, lawsuits against these sites are often pyrrhic victories since civil judgments can be extremely difficult to enforce in foreign jurisdictions. For example, one of the most notorious sites, Library Genesis, has 7.5 million books, each one representing years of labor by its author. It was slapped with a $15 million judgment in 2017 and another $30 million in 2024 in a separate lawsuit, but continues to operate. Another notorious site, Z-Library, was indicted in federal court in 2022. The FBI seized 240 domains used by the site and arrested two of its principals—Russian nationals—in Argentina. But the site soon came back online under new domains and its pirate book repositories were reuploaded by others.

And it wasn't just the *quantity* of books that made these sites attractive to the AI companies; equally important was the quality. If AI companies only needed words, they could have fed every dictionary in the world into their machine learning. But that was not nearly good enough because it would mean decades of additional work and hundreds of billions of dollars of additional investment. What they needed was complete, well-crafted, living, breathing stories, with characters that seemed real, plots that made sense, dialogue that appeared genuine. Humanity on the page. In sum, they needed us and our craft that we earned with the sweat of our brows and the flexing of our imaginations. And these companies have swooped in, stolen that labor in order to make enormous profits. But we, the writers, the source of all this, will receive nothing.

AI will also allow anyone, with no effort at all, to order up a novel written in the vein of an established writer. And that book can be sold saying that it reads just like a David Baldacci novel. Yes, it does read like my novels. Because it is my novel. It is my imagination.

---

[1] See 17 U.S.C. § 506.

People complain about cheap imported goods hurting American workers. Well, we have cheap books being created by American technology flooding the market. The Authors Guild receives reports that for many if not most new anticipated top selling books (they usually have significant advance sales), an AI-generated book that is intended to directly compete against the real book and divert sales is posted the day of or even before the release date of the real book. As AI becomes more widespread, the number of such books will only increase, forcing authors into an endless game of whack-a-mole. That will mean lower profits for publishers, and less money to spend on new, emerging writers. That hurts all of us. Online vendors now require the "author" to disclose if a book was not human created. It's getting to the point where they will have to limit the number of books that someone can publish on a weekly or even daily basis. This is insane.

All this comes at a time when writers are already facing unprecedented hurdles in earning a living. Between 2009 and 2018, authors' median incomes dropped 42%.[2] The Authors Guild's most recent authors' earnings survey found that the median writing-related income for full-time authors in 2022 was just over $20,000, with only half of that from books.[3] Looking at all authors, including those who reported writing part-time, the median book income was $2,000 in 2022, and the median income from books plus other writing-related work was $5,000.

Stemming this tide of piracy and unfair competition requires congressional action. I urge Congress to pass legislation that gives copyright owners the ability to obtain judicial orders blocking access to foreign piracy sites or online services. In addition, Congress should adopt commonsense transparency legislation requiring AI companies to disclose any unlicensed copyrighted works used in training. And to help ensure that consumers aren't

---

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

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

deceived into purchasing machine-generated outputs aimed at capitalizing on the work of human authors, Congress should require that AI-generated content be labeled as such.

More broadly, I urge Congress to consider what the AI companies' position means for copyright as a whole and the future of the creative professions in this country. Source code and elements of algorithms are also protected by copyright. I would hazard to bet that if I stole any of the AI communities' source codes or algorithms and then tried to profit off them, they would unleash a tsunami of lawsuits against me. However, if, as AI companies contend, fair use is actually my entire body of work, there is no more copyright protection for anyone. I'm sure the AI community believes that their IP should be fully protected against interlopers, and I agree with them. Thus I am deeply disappointed that they don't feel that people such as myself should enjoy the same rights and protections.

The AI community apparently believes that they are entitled to steal our work product despite it being copyrighted because what they are doing is so transformational.
Well, billions of people have been transformed by books. Many significant events in human history had seminal authors and their works that rode at the head of the pack. We didn't truly emerge from the dark ages until the invention of the printing press when books became widely available. Books also teach us empathy, making the world a kinder, gentler, more meaningful place.

I'm only one man but books transformed my life, propelling me to a far better existence.

I'm sure there are aspects of AI that will also transform the world.

But if you want to bet on which side is more transformational, for all of us, I will bet on books every single time.

Thank you.

Testimony before the U.S. Senate Committee on the Judiciary
Subcommittee on Crime and Counterterrorism

Hearing on Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training

July 16, 2025

Edward Lee
Professor of Law
Santa Clara University School of Law

Chair Hawley, Ranking Member Durbin, and other Members of the Subcommittee: Thank you for the opportunity to testify. I am a Professor of Law at Santa Clara University School of Law.[1] I am also a book author and a photographer,[2] and my personal experience informs my scholarship and understanding of the importance of copyright to authors and artists across the country.

In my testimony, I will discuss whether using copyrighted works to train AI models is a fair use, giving particular attention to the two recent decisions by Judges Alsup and Chhabria in cases filed by book authors against Anthropic and Meta. This novel question of law, which has important implications for U.S. national interest, has sparked sharp disagreements among parties, stakeholders, and now federal judges. As Judge Bibas noted in an earlier non-generative AI case, this question of law is difficult.[3] In my opening remarks, I would like to stress three points.

Transformative purpose of AI training. First, I believe Judges Alsup and Chhabria correctly concluded that the use of copies of works to train an AI model serves a highly transformative purpose in developing a new technology under Factor 1 of fair use.[4] During training, an AI model is exposed to vast training materials, typically many millions of works. Through a process called deep learning, the model identifies the "statistical relationships among words," thereby enabling the model to conduct numerous functions, including research, translation of foreign languages, delivery of medical advice, generation of content, and so forth.[5] As Judge Chhabria concluded, "The purpose of Meta's copying was to train its LLMs [large language models], which are innovative tools that can be used to generate diverse text and perform a wide range of

---

[1] My law review article on "Fair Use and the Origin of AI Training" will be published by Houston Law Review. *See* Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5253011 [hereinafter *Origin of AI Training*]. I published other articles on copyright issues raised by generative AI. *See* Edward Lee, *AI and the Sound of Music*, 134 YALE L. J. FORUM 187 (2024); Edward Lee, *Prompting Progress: Authorship in the Age of AI*, 76 FLA. L. REV. 1445 (2024). On my website CHATGPT IS EATING THE WORLD, I track and analyze all the U.S. copyright lawsuits—currently 44 pending lawsuits—against AI companies. I have included a map of the United States listing these cases attached at the end of this statement as Appendix E.
[2] *See* Edward Lee, CREATORS TAKE CONTROL (2023); Edward Lee, PICERRIFIC PHOTOGRAPHY.
[3] Thomson Reuters Enter. Centre GMBH v. ROSS Intell., Inc., 2025 WL 1488015, at *1 (D. Del. May 23, 2025) (granting petition to file interlocutory appeal on fair use and copyrightability of headnotes while noting "[o]ur circuit has not yet spoken on this 'novel and difficult question[ ] of first impression.'") (internal citation omitted); *id.* ("these questions are hard").
[4] *See* Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *7 (N.D. Cal. Jun. 23, 2025); Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *9 (N.D. Cal. June 25, 2025). I summarize the decisions in attached Appendix A.
[4] *Kadrey*, 2025 WL 1741691, at *9-*10. The four factors of fair use in Section 107 are quoted in Appendix A.
[5] *Id.* at *5, *9.

functions."[6] And, as Judge Alsup recognized, "The technology at issue was among the most transformative many of us will see in our lifetimes."[7]

The history of AI development strongly supports this conclusion. It is important to understand *why* AI researchers at universities began training AI models on large datasets. The practice originated, not at AI companies, but at universities where AI researchers discovered a key insight: *scaling*, or using larger and more diverse datasets actually worked in developing and improving AI models—an achievement that escaped researchers for many years.[8] This seminal breakthrough, which took decades to figure out, propelled the advances in AI witnessed today.

Some uses might not be fair. Second, while I agree with the ultimate findings of fair use in both cases, it's important to remember that fair use is a fact-specific doctrine decided on a case-by-case basis. In some situations, a transformative purpose in AI training might be outweighed by other fair use factors. For example, an AI model that routinely produces outputs that are infringing, such as regurgitations, might not be a fair use—even in the training—due to insufficient guardrails. Critically, in the cases against Anthropic and Meta, the plaintiffs did *not* show the models produced infringing outputs of their works.[9]

National priority in AI innovation. My final point is the need for caution—caution by the courts, Congress, and the states. I believe it's important to weigh the United States' interest in AI innovation. President Trump issued an executive order making U.S. development and global leadership in AI a national priority.[10] China has its own priority and a plan—of surpassing the United States and becoming the world leader in AI by 2030.[11] The United States' national priority in AI counsels caution.

Indeed, in *Google v. Oracle*, another technology fair use case of national importance, the Supreme Court itself cautioned: "Given the rapidly changing technological, economic, and business-related circumstances, we believe we should not answer more than is necessary to resolve the parties' dispute."[12] Judges Alsup and Chhabria departed from this approach in some controversial parts of their opinions that were just dicta. I disagree with Judge Alsup's suggestion on pirated books and Judge Chhabria's suggestion on copyright dilution, as more fully elaborated in my written statement. At this juncture, I think the best approach is for Congress to wait and see how other district courts, the courts of appeals, and potentially the Supreme Court resolve these difficult issues in the many pending copyright lawsuits.

---

[6] *Id.*; *see id.* at *10 ("First, an LLM's consumption of a book is different than a person's. An LLM ingests text to learn 'statistical patterns' of how words are used together in different contexts. It does so by taking a piece of text from its training data, removing a word from that text, predicting what that word will be, and updating its general understanding of language based on whether it was right or wrong—and then repeating this exercise billions or trillions of times with different text. This is not how a human reads a book. Second, unlike the hypothetical professor, Meta did not just give the plaintiffs' books to one person. Meta copied the plaintiffs' books as part of an effort to create a tool that can generate a wide range of text.").
[7] *Bartz*, 2025 WL 1741691, at *18.
[8] Lee, *Origin of AI Training*, at 149, 152, 156, 170-76, & nn. 229-51, 326-3 (tracing history of AI research and discovery of scaling by researchers, including citations of AI research articles).
[9] Bartz, 2025 WL 1741691, at *7; *Kadrey*, 2025 WL 1752484, at *15 ("Llama does not allow users to generate any meaningful portion of the plaintiffs' books. Neither party's expert opined that Llama was able to regurgitate more than 50 words from any of the plaintiffs' books, even in response to 'adversarial' prompting designed specifically to make LLMs regurgitate.").
[10] Executive Order, *Removing Barriers to American Leadership in Artificial Intelligence*, WHITE HOUSE (Jan. 23, 2025).
[11] New Generation Artificial Intelligence Development Plan (2017) (issued by State Council on Jul. 20, 2017).
[12] Google LLC v. Oracle Am., Inc., 593 U.S. 1, 20 (2021).

EXECUTIVE SUMMARY

- **Highly transformative purpose of AI training.** Judges Alsup and Chhabria correctly concluded, in *Bartz v. Anthropic* and *Kadrey v. Meta*, respectively, that the use of copies to train an AI model serves a *highly transformative* purpose in developing a technology under Factor 1 of fair use.

- **Origin of scaling by university researchers:** The history of university researchers training AI models on larger and more diverse datasets—a process called *scaling*, which proved to be a seminal breakthrough that led to the advances in AI today—supports this finding of a transformative purpose.

- **U.S. national interest in AI development:** President Trump's Executive Order declaring AI development a U.S. national priority and the Supreme Court's precedents recognizing that fair use is fact-specific and that copyright law must balance copyright and innovation both counsel caution and the avoidance of overbroad rulings or amendments that might jeopardize the U.S. national interest in AI development.

- **Need for caution:** Accordingly, I disagree with Judge Alsup's suggestion, in dicta, that pirated books are "irredeemably" infringing no matter the transformative purpose. And I disagree with Judge Chhabria's suggestion that most AI training is illegal under a new theory of market dilution. Neither categorical approach finds support in the text of the Copyright Act or case law.

- **No legislation needed at this time:** At this early stage of the copyright litigation involving AI companies, the best course for Congress is to wait and see how the cases are resolved by other district courts, the courts of appeals, and potentially the Supreme Court. *Bartz* and *Kadrey* are just two of more than forty AI copyright lawsuits.

I. THE ORIGIN OF AI TRAINING AND ITS TRANSFORMATIVE PURPOSE

More than forty copyright lawsuits against AI companies are now pending in the United States.[13] A central question in these copyright lawsuits is whether an AI company's use of copyrighted works to train and develop its AI models is a fair use—or not. In two different cases, federal judges, Judge Alsup in *Bartz v. Anthropic* and Judge Chhabria in *Kadrey v. Meta*, recently held that an AI company's use of copyrighted works to train AI models served a highly transformative purpose in developing the AI technology and, after balancing the four factors of fair use, the use was a fair use.[14] (Both judges were critical of the AI companies in non-precedential parts of their opinions related to acquiring pirated books and market dilution, respectively. I discuss these issues in Parts II and III.)

I believe Judges Alsup and Chhabria correctly concluded that the use of copyrighted works to train AI models serves a highly transformative purpose under Factor 1 of fair use, which

[13] *See Master List of Lawsuits v. AI*, CHATGPT IS EATING THE WORLD (updated Jun. 30, 2025).
[14] *See Bartz*, 2025 WL 1741691, at *7, *18; *Kadrey*, 2025 WL 1752484, at *9, *23.

examines the purpose and character of the defendant's use of copyrighted works.[15] Here, the purpose was to train an AI model and develop an innovative new technology. This "further purpose" in AI training is different from the authors' purpose in creating their books "for entertainment or education."[16] During training, the AI model is exposed to vast training materials, typically many millions of works, to identify the "statistical relationships among words." From this deep learning, the model develops the ability to conduct numerous functions, including research, translation of foreign languages, delivery of medical advice, generation of content, and so forth.[17] Critically, the plaintiffs in both cases did *not* show that the AI models had produced any infringing outputs of their books or any substantially similar copies.[18] Treating as fair use the creation of a new technology that does not redistribute significant portions of any works used in its development is amply supported by past fair use decisions, including *Google v. Oracle* and *Authors Guild v. Google*, as summarized in Appendices B and C and distinguished from cases in Appendix D involving technologies that merely redistributed infringing copies.[19]

As Judge Chhabria concluded, "The purpose of Meta's copying was to train its LLMs, which are innovative tools that can be used to generate diverse text and perform a wide range of functions."[20] And, as Judge Alsup recognized, "The technology at issue was among the most transformative many of us will see in our lifetimes."[21]

The history of AI development strongly supports this conclusion. It is important to understand *why* AI researchers at universities began training AI models on large datasets, much of which contained numerous copyrighted works used without permission. The practice originated, not at AI companies, but at universities where AI researchers discovered a key insight: *scaling*, or using larger and more diverse datasets actually worked in developing and improving AI models.[22] This seminal breakthrough, which took decades to figure out, propelled the advances in AI witnessed today.

Some uses might not be fair, however. A highly transformative purpose in AI training does not guarantee such use is a fair use, however. I agree with Judge Alsup's and Judge Chhabria's respective findings of fair use in the particular facts of the cases *Bartz* and *Kadrey*. But these decisions do not mean that AI training is fair use in every case. Courts must balance all four

---

[15] 17 U.S.C. § 107. *See Bartz*, 2025 WL 1741691, at *7; *Kadrey*, 2025 WL 1752484, at *9.
[16] *Kadrey*, 2025 WL 1741691, at *9-*10.
[17] *See id.* at *5, *10.
[18] *See Bartz*, 2025 WL 1741691, at *7; *Kadrey*, 2025 WL 1752484, at *9, *15.
[19] *See* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 33 (2021) ("Here Google's use of the Sun Java API seeks to create new products. It seeks to expand the use and usefulness of Android-based smartphones. Its new product offers programmers a highly creative and innovative tool for a smartphone environment. To the extent that Google used parts of the Sun Java API to create a new platform that could be readily used by programmers, its use was consistent with that creative 'progress' that is the basic constitutional objective of copyright itself."); Authors Guild v. Google, Inc., 804 F.3d 202, 216, 224 (2d Cir. 2015) ("Google's making of a digital copy of Plaintiffs' books for the purpose of enabling a search for identification of books containing a term of interest to the searcher involves a highly transformative purpose, in the sense intended by *Campbell*."; Google Book Search's snippet copy of small parts of the books in search results did not produce "meaningful or significant effect" of cognizable market harm, even though it could result in "some loss of sales" of books due to a user's ability to find an unprotected historical fact contained in the book).
[20] *Kadrey*, 2025 WL 1741691, at *9.
[21] *Bartz*, 2025 WL 1741691, at *18.
[22] Lee, *Origin of AI Training*, at 149, 152, 156, 170-76, & nn. 229-51, 326-3 (tracing history of AI research and discovery of scaling by researchers, including citations of AI research articles).

factors of fair use under the facts of each case, including the potential market harm to the copyright holder's original work and derivative works.

As explained in my scholarship, an AI model that routinely produces outputs that are infringing, such as regurgitations, might not be a fair use due to insufficient guardrails (and using more of the works than reasonably necessary for the purpose of developing an AI model).[23] Moreover, the outputs of an AI model are separate uses and can constitute separate acts of infringement if they are substantially similar copies of works in the training datasets.[24] And, in some cases, AI outputs that copy a specific artist's style—generated "in the style of" an artist—may well include copyrightable elements to support an infringement claim.[25]

II. HOW TO WEIGH THE USE OF PIRATED BOOKS FROM SHADOW LIBRARIES

My second recommendation relates to the controversial issue related to the use of pirated books from shadow libraries online. These shadow libraries were created by unnamed people who have escaped legal efforts to shut the sites down, even in the face of court orders.[26] Even when a website is blocked, a shadow library can easily resurface at a different site hosted by foreign locations.[27] The most contentious issue in the book author lawsuits is whether the AI company's acquiring and copying pirated books from shadow libraries online was (i) a separate infringing use or (ii) a use for the further purpose to train the defendant's AI models.[28]

Judges Alsup and Chhabria disagreed on this issue. Judge Alsup treated Anthropic's acquisition of copies from shadow library as a separate use for Anthropic's own library building—and held that such library building was not fair use.[29] In dicta, Judge Alsup suggested an even more categorical approach—that *any* acquiring of pirated copies was "inherently, irredeemably infringing" no matter what the transformative purpose and even if the copies were "immediately discarded" after a transformative use.[30] By contrast, Judge Chhabria took a flexible approach viewing the acquisition of the copies in relation to the defendant's further, transformative purpose in acquiring them, namely, AI training.[31] But Judge Chhabria recognized that the use of

---

[23] *Id.* at 213-15.
[24] *Id.* at 113 ("Under *Warhol*, courts can find that uses in AI training serve a fair purpose, but uses in AI outputs that are 'regurgitated' or substantially similar copies do not.").
[25] *See id.* at 202 ("Granted, some AI generators may copy copyrightable elements when generating a work in response to a person's prompt to create in 'the style of' a specific artist. But the proper remedy is a copyright infringement lawsuit, not concocting a mutant species of copyright dilution that penalizes non-infringing works."); *see also* 2 PATRY ON COPYRIGHT § 4:14 (2025) (distinguishing between unprotectable communal or generalized styles versus a style distinctive to an individual based on copyrightable elements of specific works).
[26] *See* Ashley Belanger, *"Most notorious" illegal shadow library sued by textbook publishers*, ARS TECHNICA (Sep. 15, 2023).
[27] *Id.*
[28] AI researchers determined that using books provided high-quality data to train LLMs that yielded better LLMs. For example, in internal emails disclosed as part of Kadrey's summary judgment motion, Meta developers concluded that the use of the controversial Library Genesis dataset is "essential to meet SOTA [state of the art] numbers across all categories." *Evidence of Meta's use of LibGen dataset and seeding torrents to share files. Wanted to compete with OpenAI and Mistral.*, CHATGPT IS EATING THE WORLD (Feb. 6, 2025); *see* Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *5 (N.D. Cal. June 25, 2025) ("To be able to generate a wide range of text—in different languages or styles, or regarding different subject matter—an LLM's training dataset must be large and diverse…. But while a variety of text is necessary for training, books make for especially valuable training data. This is because they provide very high-quality data for training an LLM's 'memory' and allowing it to work with larger amounts of text at once.").
[29] *See* Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *11-*14 (N.D. Cal. Jun. 23, 2025).
[30] *Id.* at *11.
[31] *Kadrey*, 2025 WL 1752484, at *12.

pirated books can weigh *against* fair use as market harm if the evidence showed the "use of shadow libraries benefited those libraries or their other users."[32] In *Kadrey*, Judge Chhabria concluded the plaintiffs failed to present sufficient evidence of such market harm.[33]

Judge Chhabria's flexible approach to pirated copies offers the better way for courts to address the issue of pirated books. First, it is faithful to the text of the Copyright Act. Unlike the first-sale doctrine and other copyright exceptions, the text of Section 107, the fair use provision, contains no requirement that the defendant use a "lawfully made copy" to qualify for fair use.[34] In enacting the Copyright Act of 1976, Congress knew how to draft a per se requirement of a "lawfully made copy" for a copyright exception, but did not do so for fair use.[35] The text of Section 107 forecloses the adoption of any per se requirement that people must obtain a lawfully made copy to assert a valid fair use defense.

Second, when it had a chance to recognize such a per se requirement in *Harper & Row*, which involved a purloined manuscript of a book, the Supreme Court did not do so—instead weighing the purloined character in the overall balance of fair use factors.[36] In *Google v. Oracle*, the Court also declined to recognize "bad faith" of the defendant as a relevant factor, while quoting from Judge Leval's seminal fair use article that "[c]opyright is not a privilege reserved for the well-behaved."[37] The *Warhol* Court recognized that "[m]ost copying has some further purpose."[38] A defendant should not be precluded from asserting a further purpose to justify making an unauthorized copy as a part of a fair use defense, but the defendant faces potential liability if the defense fails. Likewise, the Copyright Office's pre-publication report on AI training rejected a categorical approach but instead recommended that the use of pirated copies "should weigh against fair use without being determinative."[39] Third, Judge Chhabria's approach to pirated books leaves open the possibility that an AI company's use of them will *not* be fair use based on the submission of evidence that such use materially supported a shadow library.[40] Far from

---

[32] *Id.*

[33] *See id.* at *21. ("But although the plaintiffs discussed Meta's use of shadow libraries at length, they did not argue that it had these effects or was relevant to the fourth factor beyond allowing Meta to get the books without paying….[T]he plaintiffs' counsel did suggest that, by using shadow libraries, Meta (and other companies like it) would reduce the stigma associated with shadow libraries and encourage more people to use them. It's not clear whether this would matter in the overall analysis. But in any event, counsel conceded that the record contains no evidence of this dynamic playing out.") (internal citation omitted).

[34] *Compare* 17 U.S.C. § 109(a) ("the owner of a particular *copy … lawfully made* under this title") (emphasis added) *with id.* § 107 ("fair use of a copyrighted work"); Kirtsaeng v. John Wiley & Sons, Inc., 568 U.S. 519, 537 (2013) (discussing "lawfully made" copy requirement in §§ 109(c) (exception to public display), 109(e) (exception for video games in coin-operated equipment), and 110(1) (in-classroom teaching exception to public display and performance but not if copy "not lawfully made"); *see also* 17 U.S.C. § 108(c)(2) ("lawful possession of such copy" by library or archives).

[35] *Kirtsaeng*, 568 U.S. at 537 (the prior 1909 Copyright Act's language for the first-sale doctrine was even more explicit in requiring for the first-sale exception: "'[N]othing in this Act shall be deemed to forbid, prevent, or restrict the transfer of any copy of a copyrighted work *the possession of which has been lawfully obtained.*'" Copyright Act of 1909, § 41, 35 Stat. 1084 (emphasis added)).'"

[36] Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985).

[37] Google LLC v. Oracle Am., Inc., 593 U.S. 1, 32-33 (2021) (quoting Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1126 (1990)).

[38] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 528-29 (2023).

[39] U.S. Copyright Office, *Copyright and Artificial Intelligence Part 3: Generative AI Training* (Pre-Publication Report) 52 (May 2025).

[40] *Kadrey*, 2025 WL 1752484, at *12, *21.

giving a green light to the use of pirated books datasets, Judge Chhabria's approach signals great legal risk for any AI company that does so.[41]

This fact-specific approach allows courts to carefully balance the four factors of fair use. Granted, the file sharing of unauthorized copies of works is infringement in many cases. But Section 107 and Supreme Court precedent do not support a rigid, categorical approach to treat every unauthorized copy as "inherently, irredeemably infringing" no matter what the transformative purpose the defendant had. Even copying for the purpose of library-building, the Supreme Court in *Grokster* said was not "necessarily infringing."[42] Even in the illegal music-file sharing cases, the courts initially evaluated the defendant's asserted purpose, such as the practice of sampling to decide whether to purchase a copy or the convenience of space-shifting in digital format, but ultimately held it was not transformative.[43] By contrast, in both *Bartz* and *Kadrey*, both judges found the defendant's use in AI training was highly transformative.[44] And the notion that property that was initially illicit is "irredeemably" so and can never be repurposed for legitimate public ends is not recognized in other areas of federal law.[45]

Assume for the sake of argument the courts in a decision or Congress in an amendment adopts a per se requirement that a defendant must initially acquire a "lawfully made copy" of a work to be able to assert a fair use defense. This categorical approach would dramatically shrink the scope of fair use. *Every* fair use necessarily involves an unauthorized copy—indeed, that is very question whether the unauthorized copy is a fair use. Judge Alsup's opinion repeatedly referred to "pirated book" without explaining the term, much less whether it is different from an

---

[41] Moreover, evolving norms in AI training might coalesce around some best practices. *See The EU Code of Practice and future of AI in Europe*, OPENAI (Jul. 11, 2025) (intent to sign EU's Code of Practice for General Purpose of AI); EU Code of Practice for General-Purpose AI Models, Copyright Chapter, Measure 1.2 (measure to copy "only lawfully accessible copyright-protected content when crawling the World Wide Web").

[42] MGM Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 931 (2005) (discussing video library building in Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 424, 454-55 (1984) as not necessarily infringing).

[43] *See, e.g.*, BMG Music v. Gonzalez, 430 F.3d 888, 890 (7th Cir. 2005) (sampling purpose failed because "[i]nstead of erasing songs that she decided not to buy, she retained them."); A&M Records, Inc. Napster, Inc., 239 F.3d 1004, 1018-19 (9th Cir. 2001) (space shifting purpose failed because file-sharing also involved distributing music files to others); *see also* UMG Recordings, Inc. v. MP3.com, 92 F. Supp. 2d 349, 351 (S.D.N.Y. 2000) (space shifting service was "simply another way of saying that the unauthorized copies are being retransmitted in another medium—an insufficient basis for any legitimate claim of transformation"). Based on these and other music-file sharing decisions, any defendant engaged in music file sharing of copyrighted works is likely engaging in infringement, with no fair use defense. *See In re DMCA § 512(h) Subpoena to Twitter, Inc.*, 608 F. Supp. 3d 868, 879 (N.D. Cal. 2022) (Chhabria, J.) ("In some cases, no analysis is required; it is obvious, for example, that downloading and distributing copyrighted music via peer-to-peer systems does not constitute fair use."); *see also* U.S. v. Slater, 348 F.3d 666, 668-69 (7th Cir. 2003) (upholding denial of jury instruction of fair use in criminal case against defendant who was participant in website to distribute illegally pirated software). These cases involve mere redistribution of copies of works, which, as shown in the table in Appendix D attached to this statement, is typically not fair use.

[44] Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *7 (N.D. Cal. Jun. 23, 2025); Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *9 (N.D. Cal. June 25, 2025).

[45] *See* Kristina Rae Montanaro, Note, *"Shelter Chic": Can the U.S. Government Make It Work*, 42 VAND. J. TRANSNAT'L L. 1663, 1664-65 (2009) (discussing U.S. Customs Service's donation of seized counterfeit goods for humanitarian relief after Hurricane Katrina, exceeding $20 million in value); *The US Government Sold Nearly 10,000 Silk Road Bitcoin*, NASDAQ (Mar. 31, 2023); *NCDA & Partners Mark End of Project Donating Nearly 100K Seized Counterfeit Jackets to NY Charities*, NASSAU CO. DISTRICT ATT'Y (Apr. 28, 2022) ("Nassau County District Attorney Anne T. Donnelly today announced the completion of a six-year long effort to donate nearly 100,000 counterfeit jackets – seized during multiple investigations – to more than 160 charities across Long Island and the greater New York area."); *Real Property Auctions*, U.S. TREASURY (listing auctions of real property seized by federal government); *Disposition of Seized, Forfeited, Voluntarily Abandoned, and Unclaimed Personal Property*, U.S. DEP'T OF INTERIOR (allowing donation of seized drug paraphernalia for law enforcement or educational purposes).

unauthorized initial copy.[46] If it is the same, then every dataset of copyrighted works collected for AI training is pirated—and every AI company's and every researcher's acquisition of the unauthorized dataset is infringement. Such an extreme result would greatly hinder AI development in the United States.

Even limited to the pirated books datasets online, Judge Alsup's suggested categorical approach to pirated books disproportionately favors Big Tech and other well-financed companies that have the resources to spend many millions of dollars to buy physical books and manually scan digital copies of them for AI training as Anthropic eventually did.[47] And, among Big Tech, Google might have a big advantage given its Google Book search database. Small tech companies and independent researchers would have little chance in contributing to innovation in AI models. Such a rule favoring Big Tech companies is bad for innovation in the United States.[48]

III. THE U.S. NATIONAL INTEREST IN AI DEVELOPMENT AND INNOVATION

The U.S. national interest in AI development and innovation counsels caution by the courts, Congress, and the states. Technological progress is just as important to the United States as artistic progress.[49] In three technology-related copyright cases, the Supreme Court recognized the important need to balance the competing interests in copyright and technological innovation.[50] As the Federal Circuit explained citing the legislative history of the Copyright Act of 1976, the fair use doctrine provides a way for courts to address "technological innovations."[51] The fair use doctrine is an American doctrine, which originated in the United States and has accommodated innovation from the VCR to programs that enhance the Internet and smartphones, all technologies of great national significance. In *Grokster*, the Court also described the *Sony* safe harbor—for technologies capable of substantial non-infringing uses—as a doctrine that "leaves breathing room for innovation and a vigorous commerce."[52]

In an executive order, President Trump declared: "It is the policy of the United States to sustain and enhance America's global AI dominance . . . to promote human flourishing, economic competitiveness, and national security."[53] Then-President Biden had recognized the same priority in AI in an earlier executive order.[54] And, in 2018, Congress established the independent National Security Commission on Artificial Intelligence, which warned in 2021: "For the first time since World War II, America's technological predominance—the backbone of its economic

---

[46] Edward Lee, *Judge Alsup's Solomonic judgment on fair use in AI training & acquiring pirated books: is it the blueprint for the future of AI training? Part I: Pirated copies*, CHATGPT IS EATING THE WORLD (June 25, 2025).
[47] *See Bartz*, 2025 WL 1741691, at *2.
[48] *See generally* Mark Lemley & Watt Wansley, *How Big Tech Is Killing Innovation*, N.Y. TIMES (Jun. 13, 2024); Mark A. Lemley & Matthew T. Wansley, *Coopting Disruption*, 105 BOSTON UNIV. L. REV. 458 (2025).
[49] *See* Lee, *Origin of AI Training*, at 145-47.
[50] *See* Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 933 (2005); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 22 (2021); Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 442-43 (1984); Lee, *Origin of AI Training*, at 126-29, 143-47.
[51] Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992); H.R. REP. NO. 94-1476, at 66 (1976) ("The bill endorses the purpose and general scope of the judicial doctrine of fair use, but there is no disposition to freeze the doctrine in the statute, especially during a period of rapid technological change."); S. REP. NO. 94-473, at 62 (1975) (same).
[52] *Grokster*, 545 U.S. at 933.
[53] President Donald J. Trump, Executive Order, *Removing Barriers to American Leadership in Artificial Intelligence*, WHITE HOUSE (Jan. 23, 2025).
[54] *FACT SHEET: Biden-Harris Administration Announces New AI Actions and Receives Additional Major Voluntary Commitment on AI*, Internet Archive (Jul. 26, 2024).

and military power—is under threat. China possesses the might, talent, and ambition to surpass the United States as the world's leader in AI in the next decade if current trends do not change."[55]

Although courts decide private disputes between parties, the consideration of fair use allows courts to consider the *public benefit* of the defendant's use, as well as how it may serve the overall constitutional goal of "promoting Progress" in the United States.[56] As the Supreme Court explained in another technology case, *Google v. Oracle*, the transformative purpose "to create a new platform" that enables others to create new applications was "consistent with that creative 'progress' that is the basic constitutional objective of copyright itself."[57] Indeed, Judges Alsup and Chhabria cited the Copyright Clause or the *Google* Court's analysis of the transformative purpose in developing a new computing platform.[58]

Given the U.S. national priority in AI innovation, both the courts and Congress should proceed cautiously before adopting a categorical or inflexible rule that might greatly hamper AI innovation. For example, in an extensive section of dicta, Judge Chhabria concluded that, in "most cases," AI training on copyrighted works is likely "generally illegal" and that AI "companies, to avoid liability for copyright infringement, will generally need to pay copyright holders for the right to use their materials."[59] To reach that sweeping conclusion, Judge Chhabria speculated for many pages on a new theory of copyright market dilution, even though he held that the plaintiffs had failed to present sufficient evidence of their own putative market harm to survive summary judgment.[60]

The new theory of copyright market dilution should be rejected. It impermissibly expands the scope of copyright to non-infringing works of others and treats those non-infringing works as cognizable market harm that a copyright holder can claim under fair use.[61] Thus, in Judge Chhabria's view, even if people using AI do not produce any infringing outputs, the fair use defense in training the respective AI model would still fail *simply because the model can produce non-infringing outputs in the same genre or type of work* in the training datasets.[62] For example, if an AI model was trained on romance novels, every non-infringing romance novel someone creates using that model can constitute market dilution under Factor 4—even though the non-infringing romance novel contains no copied protected expression from the works in the training datasets.[63] Judge Chhabria conceded that no court has ever recognized such an expansive view of market harm to include non-infringing works of others.[64]

---

[55] National Security Comm'n on Artificial Intelligence, *Final Report* 7 (2021).
[56] Google LLC v. Oracle Am., Inc., 593 U.S. 1, 35-36 (2021) ("Further, we must take into account the public benefits the copying will likely produce.").
[57] *Id.* at 30.
[58] *See* Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *8 (N.D. Cal. Jun. 23, 2025) (citing Copyright Clause, U.S. CONST. art I, § 8, cl. 8); Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *9 (N.D. Cal. June 25, 2025) (quoting and citing *Oracle*, 593 U.S. at 30).
[59] *Kadrey*, 2025 WL 1752484, at *1-*2.
[60] *Id.* at 1-*2, *15-*23.
[61] *Id.*
[62] *Id.* at *15.
[63] *Id.* at *18.
[64] *Id.* ("it's never made a difference in a case before").

As I explain at greater length in my law review article, copyright market dilution is overbroad and likely unconstitutional.[65] Dilution is a trademark concept. It did not become recognized under federal trademark law until Congress amended the Lanham Act in 1995, expanding the scope of trademarks for famous marks to prohibit dilution.[66] To apply dilution to expand the scope of copyrights in the fair use analysis is problematic. To borrow Justice Scalia's apt phrase in an analogous case, the new theory of copyright dilution is "a species of mutant copyright law": it misuses a trademark concept to protect copyrighted works.[67]

Market dilution extends copyright beyond the constitutional limitations in the Copyright Clause, which gives Congress the power to grant exclusive rights to authors only in "*their respective writings*."[68] A genre, type, or kind of work is not a "writing"; it is just an idea or method for identifying a stock or common way to organize expression, such as an article, essay, novel, poem, or play.[69] As Judge Alsup recognized in rejecting market dilution in *Bartz*, "Copyright does not extend to 'method[s] of operation, concept[s], [or] principle[s]' 'illustrated[ ] or embodied in [a] work.'"[70] Indeed, as Justice Story explained, "Every book in literature, science and art, borrows, and *must necessarily borrow*, and use much which was well known and used before."[71] This fundamental limitation on copyright explains why no novelist can own the genre for novels, romance or otherwise, let alone claim cognizable market harm from other novels that do not infringe. Yet the theory of mutant copyright dilution proposes to do just that—expand copyright to genres and uncopyrightable ideas.[72] Such an expansion turns the Copyright Clause on its head. Instead of promoting progress, the goal is to protect copyrights—and perversely to reduce the creation of new, non-infringing works.

---

[65] See Lee, *Origin of AI Training*, at 188-208.
[66] Erin J. Roth & Robert B. Bennett, Jr., *The Federal Trademark Dilution Act: Potent Weapon or Uphill Battle*, 16 MIDWEST L. REV. 1, 7-11 (1999) (history of Congress's enactment of dilution protection for famous marks in Lanham Act in 1995).
[67] Dastar Corp. v. Twentieth Century Fox Film Corp., 539 U.S. 23, 34 (2003).
[68] U.S. CONST. art. I, § 8, cl. 8.
[69] *See* Hassett v. Hasselbeck, 757 F. Supp. 2d 73, 89-90 (D. Mass. 2010) ("While the books address some of the same topics, the order of presentation is not identical or nearly so. To the extent there is any general similarity related to the selection and ordering of the topics, the defendants' exhibits demonstrate that the general sequence and topic selection of these works are customary to the genre, and thus unprotected under the doctrine of *scenes a faire*. *See Coquico,* 562 F.3d at 68. Moreover, courts have held that the general thematic ordering and arrangement of a work is not usually copyrightable. *See LaPine,* 2009 WL 2902584, at *9; *see also Dunn,* 517 F.Supp.2d at 544 (holding that the claim that two works have substantial thematic and structural similarity 'has little or no support in the law as a basis for a copyright claim'). To the extent there is any similarity between the structures of the two works, that similarity relates to unprotected elements of the works and does not support a finding of substantial similarity."); *see also* Nichols v. Universal Pictures Corp., 45 F.2d 119, 121-22 (2d Cir. 1930) (discussing how copyright does not protect unprotected ideas and elements in a play); Peters v. West, 692 F.3d 629, 636 (7th Cir. 2012) ("Copyright protects actual expression, not methods of expression. 17 U.S.C. § 102(b); Baker v. Selden, 101 U.S. 99 (1879). Just as a photographer cannot claim copyright in the use of a particular aperture and exposure setting on a given lens, no poet can claim copyright protection in the form of a sonnet or a limerick.").
[70] Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *8 (N.D. Cal. Jun. 23, 2025) (quoting 17 U.S.C. § 102(b)).
[71] Emerson v. Davies, 8 F. Cas. 615, 619 (No. 4,436 (CCD Mass. 1845) (Story, J.) (emphasis added); *see* Zechariah Chafee, *Reflections on the Law of Copyright*, 45 COLUM. L. REV. 503, 511 (1945) ("Progress would be stifled if the author had a complete monopoly of everything in his book for fifty-six years or any other long period. Some use of its contents must be permitted in connection with the independent creation of other authors. The very policy which leads the law to encourage his creativeness also justifies it in facilitating the creativeness of others.").
[72] *Cf.* Design Basics, LLC v. Signature Construction, Inc., 994 F.3d 879, 889 (7th Cir. 2021) ("Standard elements in a genre—called *scènes à faire* in copyright law—get no copyright protection. *Scènes à faire* are 'so rudimentary, commonplace, standard, or unavoidable that they do not serve to distinguish one work within a class of works from another.'" Bucklew v. Hawkins, Ash, Baptie & Co., 329 F.3d 923, 929 (7th Cir. 2003). If standard elements received copyright protection, then the creation of a single work in a genre would prevent others from contributing to that genre because the copyright owner would have exclusive rights in all of the genre's basic elements.").

That result also violates the First Amendment under which "*more speech*, not less, is the governing rule."[73] Copyright market dilution seeks to protect copyrights and *reduce* new expression embodied in non-infringing works people created using AI.[74] It penalizes, under fair use, *non-infringing expression* of others—with the likely consequence of making illegal the very AI technology that people used to create the non-infringing expression. As Judge Chhabria openly stated, most AI training will be *generally illegal* under market dilution.[75] If so, then the First Amendment rights of many people who use AI will be impaired. Such a radical change to the traditional contours of copyright, fair use, and the idea-expression dichotomy requires strict scrutiny under the First Amendment.[76]

Courts can no more treat as dilution the *non-infringing* expression of others under copyright law simply because they used AI tools than courts can treat as defamation people's *truthful* expression simply because they used AI tools in creating the expression. The First Amendment protects non-infringing expression and truthful expression alike.[77]

In applying fair use, courts must balance competing interests, including the larger public interest and benefits.[78] Heeding the Supreme Court's fair use precedents counsels caution. As the *Google* Court admonished in another technology case, "Given the rapidly changing technological, economic, and business-related circumstances, we believe we should not answer more than is necessary to resolve the parties' dispute."[79]

The extensive dicta in *Kadrey* on a new, expansive theory of copyright dilution based solely on non-infringing expression—which is protected by the First Amendment—failed to follow the Supreme Court's admonition. Instead, it opined that, in most cases, AI training is outright illegal. Such a radical categorical approach flouts the Supreme Court's repeated avoidance of bright-line rules in applying fair use.[80] And, if adopted, it jeopardizes the United States' national interest in AI. As the White House AI Czar David Sacks advised, "There must be a fair use concept for training data or models would be crippled. China is going to train on all the data regardless, so

---

[73] Citizens United v. FEC, 558 U.S. 310, 361 (2010) (emphasis added).
[74] See Lee, *Origin of AI Training*, at 202, 204-05.
[75] *See Kadrey*, 2025 WL 1752484, at *1; *id.* at *2 ("The upshot is that in many circumstances it will be illegal to copy copyright-protected works to train generative AI models without permission. Which means that the companies, to avoid liability for copyright infringement, will generally need to pay copyright holders for the right to use their materials.")
[76] *See* Eldred v. Ashcroft, 537 U.S. 186, 191 (2003) ("When, as in this case, Congress has not altered the traditional contours of copyright protection, further First Amendment scrutiny is unnecessary."); Lee, *Origin of AI Training*, at 208-09.
[77] *See* Pan Am Sys., Inc. v. Atlantic Northeast Rails & Ports, Inc., 804 F.3d 59, 65-66 (1st Cir. 2015) (First Amendment protects truthful information, which is complete defense to defamation); Neil Weinstock Netanel, *First Amendment Constraints on Copyright after* Golan v. Holder, 60 UCLA L. REV. 1082, 1128 2013) ("Courts must, accordingly, interpret and apply the idea/expression dichotomy and fair use privilege in a manner consistent with their vital First Amendment role. Further, following *Golan*, statutory provisions that disturb copyright's built-in First Amendment accommodations, or that otherwise abridge noninfringing speech, lie vulnerable to First Amendment challenge.").
[78] *See* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 35-36 (2021); *see also* Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 928 (2005) ("The more artistic protection is favored, the more technological innovation may be discouraged; the administration of copyright law is an exercise in managing the tradeoff."); Goldstein v. California, 412 U.S. 546, 559 (1973) ("Where the need for free and unrestricted distribution of a writing is thought to be required by the national interest, the Copyright Clause and the Commerce Clause would allow Congress to eschew all protection.").
[79] *Google*, 593 U.S. at 21.
[80] *See* Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 528 (2023); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 18-19 (2021); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).

without fair use, the U.S. would lose the AI race."[81] Indeed, already one court in China, stressing the need for "*encourag[ing] technological progress*," indicated that the use of copyrighted works to train an AI model is permissible under Chinese copyright law provided the AI model does not produce infringing outputs, in a decision upheld on appeal.[82]

CONCLUSION

Any categorical approach that would make all AI training illegal and not fair use—such as the dicta in Judge Chhabria's opinion in *Kadrey v. Meta* concluding that most AI training is illegal under the speculative new theory of copyright market dilution—should be rejected. Such a ruling is contrary to the Copyright Clause's limitation of copyright to only authors' "respective writings," the Copyright Act's exclusion of ideas, methods, and genres from protection, and the Supreme Court's repeated admonition on the fact-specific nature of fair use. If adopted, such a ruling would not only hamper AI innovation in the United States, but it also may prompt U.S. companies to relocate their AI training offshore to countries with copyright exceptions for text-data-mining (TDM) or fair use that would allow AI training, including Israel, Japan, and Singapore.[83] The fair use provision Congress codified in the Copyright Act as a flexible doctrine to accommodate "rapid technological change" does not support, much less require, such a drastic result.[84]

---

[81] @DavidSacks, X (Jun. 24, 2025, 10:10 AM), https://x.com/davidsacks/status/1937558998166954092; *see* National Security Comm'n on Artificial Intelligence, *Final Report* 2, 4 (2021) ("But we must win the AI competition that is intensifying strategic competition with China. China's plans, resources, and progress should concern all Americans. It is an AI peer in many areas and an AI leader in some applications. We take seriously China's ambition to surpass the United States as the world's AI leader within a decade.…The federal government must partner with U.S. companies to preserve American leadership and to support development of diverse AI applications that advance the national interest in the broadest sense."). In 2021, China amended its copyright act to include, in clause 13 to Article 24, a general exception for "[o]ther circumstances provided by laws and administrative regulations." Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L. REV. 1163, 1194 (2025). Although the provision has not yet been applied to this circumstance, some legal commentators suggest it could support AI training in China. *See id.*

[82] *See* Shanghai XX Cultural Dev. Co. v. Hangzhou XX Intelligent Tech. Co., (2024) Zhe 0192 Min Chu 1587 (Hangzhou Internet Ct. Sept. 25, 2024), *aff'd*, (2024) Zhe 01 Min Zhong 10332 (Hangzhou Interm. People's Ct. Dec. 30, 2024) ("Therefore, this court believes that It can be considered [permissible] use when there is no evidence that generative artificial intelligence is for the purpose of using the original expression of the right work, has affected the normal use of the right work, or unreasonably harms the legitimate interests of the relevant copyright holders.") (bracketed translation of "permissible use" inserted); *id.* ("Finally, there should be a prudent and inclusive approach to generative AI that *encourages technological progress and business development*. The creation and development of generative artificial intelligence requires the introduction of huge amounts of training data at the input end, which is unavoidable [to avoid] using other people's works. In view of the purpose of generative AI to use other people's works in the data training stage, it should in principle be used to learn and analyze the thoughts, feelings, language features, characteristic styles, etc. expressed in previous works, and extract corresponding rules, structures, patterns, and trends from them to facilitate subsequent transformational creation of new works. This kind of 'use behavior' to aggregate a large number of works as analysis sample data for training to improve the creative ability of the work is not for the purpose of reproducing the original expression of the work, and generally the data training is only to temporarily retain the previous work when analyzing the structural characteristics of the corpus data, the data training and generation process did not display the previous works to the public. Therefore, this court believes that it can be considered [permissible] use when there is no evidence that generative.") (bracketed translation of "permissible use" inserted and emphasis added). An AI platform was held secondarily liable for allowing infringing outputs of the character Ultraman. *See* King & Wood Mallesons, *Chinese AIGC Platform Found Secondarily Liable for Copyright Infringement*, LEXOLOGY (Feb. 28, 2025).

[83] *See* Sag and Yu, *supra*, at 1179-80, 1185-92; Jonathan Band, *Israel Ministry of Justice Issues Opinion Supporting Use of Copyrighted Works for Machine Learning*, DISRUPTIVE COMPETITION PROJECT (Jan. 19, 2023).

[84] *See* H.R. REP. NO. 94-1476, at 66 (1976).

APPENDIX A: Comparison of Decisions by Judge Alsup & Chhabria

FAIR USE FACTOR	JUDGE ALSUP Bartz v. Anthropic PBC	JUDGE CHHABRIA Kadrey v. Meta Platforms
Is downloading "pirated" books datasets separate use from training model?	***Separate use for library:*** Anthropic downloading / building a permanent library of pirated books was *not fair use*. Infringing. Trial on damages.	***Same use to train*:** Meta downloading was for the further purpose of training AI model. Fair use (but would not be had plaintiffs proven market dilution).
(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes	***Favors fair use (+).*** Train LLMs is **exceedingly "transformative — spectacularly so"** because it maps statistical relationships to produce technology that produces new, noninfringing outputs. This technology is "among the most transformative many of us will see in our lifetimes." No allegation of output infringing authors' works. Cites *Google Books* decision.	***Favors fair use (+).*** Meta's use of the plaintiffs' books had a **"further purpose" and "different character" than the books—that it was highly transformative.** "The purpose of Meta's copying was to train its LLMs, which are innovative tools that can be used to generate diverse text and perform a wide range of functions….The purpose of the plaintiffs' books, by contrast, is to be read for entertainment or education." Commercial use tends to be less important when the secondary use is highly transformative. Cites *Oracle* decision.
(2) the nature of the copyrighted work	***Disfavors fair use (-).*** Creative expressive books.	***Disfavors fair use (-).*** But second factor weighs less if transformative purpose.
(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole	***Favors fair use (+).*** "Compelling benefits of training the LLMs on strong examples were not offset by revelations to the public of any portion of the works themselves. What was copied was therefore especially reasonable and compelling."	***Favors fair use (+).*** "The amount that Meta copied was reasonable given its relationship to Meta's transformative purpose. *See Oracle*, 593 U.S. at 34. Everyone agrees that LLMs work better if trained on more high-quality material."
(4) the effect of the use upon the potential market for or value of the copyrighted work.	***Favors fair use (+).*** No cognizable market harm. Authors concede that training LLMs did not result in any exact copies nor even infringing knockoffs of their works being provided to the public. ***Rejects new theory of *copyright market dilution*.** The Copyright Act seeks to advance original works of authorship, **not to protect authors against competition.** Rejects lost licensing as a market Copyright Act entitles authors to exploit.	***Favors fair use (+).**** *Slight public benefit.* Likely help Llama create new expression. ***But accepts new theory of *copyright market dilution*:** harm by helping to enable the rapid generation of countless works that compete with the originals, **even if those works aren't themselves infringing.** But finds Plaintiffs failed to present sufficient evidence to create genuine issue. Rejects lost licensing as a market authors entitled to exploit. Circularity problem.
JUDGE'S CONCLUSION	**FAIR USE** (but not for library)	**FAIR USE** (but not if market dilution)

APPENDIX B: Technology Fair Use Decisions Decided or Favorably Cited by Supreme Court

Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025) (table 2).

CASES	TECHNOLOGY DEVELOPMENT: Use of Copyrighted Works to Create New Technology?	TECHNOLOGY USAGE: Use of Copyrighted Works in Public Use of Technology?	FACTOR 1: Use of Copyrighted Works Had Further Purpose or Different Character?
Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984).	No	Yes, by users of VCR for personal time-shift recordings.	No, time-shifted copies of free TV shows.
Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021).	Yes, use of Java declaring code for Android operating system to facilitate computer programmers' ability to write apps for Android.	Yes, declaring code was part of Android OS and can be used by programmers writing Android apps.	Yes, "use of the Sun Java API seeks to create new products," i.e., "a highly creative and innovative tool for a smartphone environment."
Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).	Yes, in reverse engineering of OS to find uncopyrightable element necessary for interoperability of new game.	No	Yes, "intermediate copying of computer code as an initial step *in the development of a competing product*."
Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).	Yes, in reverse engineering of OS to find uncopyrightable element necessary for game emulator to make games run on PC.	No	Yes, "*creates a new platform*, the personal computer, on which consumers can play games designed for the Sony PlayStation."
Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).	Yes, to create a database to enable within-book search of published books and to enable text data mining analysis of frequency of use of words in entire corpus of books.	Yes, copies stored in database and snippets of books shown.	Yes, copies in database serve further purpose of searching within-text of all books in database to find relevant sources.

APPENDIX C: Lower Courts' Finding Fair Use in Other Technology Cases

Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025) (table 3).

CASES	TECHNOLOGY DEVELOPMENT: Use of Copyrighted Works to Create New Technology?	TECHNOLOGY USAGE: Use of Copyrighted Works in Public Use of Technology?	FACTOR 1: Use of Copyrighted Works Had Further Purpose or Different Character?
Kelly v. Arriba Soft Corp., 336 F.3d 811, (9th Cir. 2003).	Yes, to create a searchable database of online images	Yes, copies stored in database and outputs show thumbnail images	Yes, copies in database serve further purpose of searching online images to find relevant ones.
Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).	Yes, to create a searchable database of online images	Yes, copies stored in database and outputs show thumbnail images of reduced resolution	Yes, copies in database serve further purpose of searching online images to find relevant ones.
A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009).	Yes, to create a searchable database of student papers	Yes, copies stored in database but no direct quotations	Yes, copies in database serve further purpose of finding potential plagiarism in student papers
Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, (2d Cir. 2014).	Yes, to create a database to enable within-book search of published books. Secondary use to store digital copies for preservation.	Yes, copies stored in database. No snippets of books shown, but (i) pages numbers in books term is found; (ii) access to full books to people with print-disability.	Yes, copies in database serve further purpose of searching within-text of all books in database to find relevant sources.
Field v. Google, Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006).	Yes, to create a searchable database of cached copies of Internet websites	Yes, copies stored in database and "cached" static copy of website publicly accessible.	Yes, copies in database serve further purpose of allowing static view of snapshot of website, useful when website is down
White v. West Pub. Corp., 29 F. Supp. 3d 396 (S.D.N.Y. 2014).	Yes, to create a searchable database of copies of legal briefs to "creat[e] an interactive legal research tool."	Yes, copies stored in databases of West and Lexis.	Yes, copies in database serve further purpose of "creating interactive legal research tool."

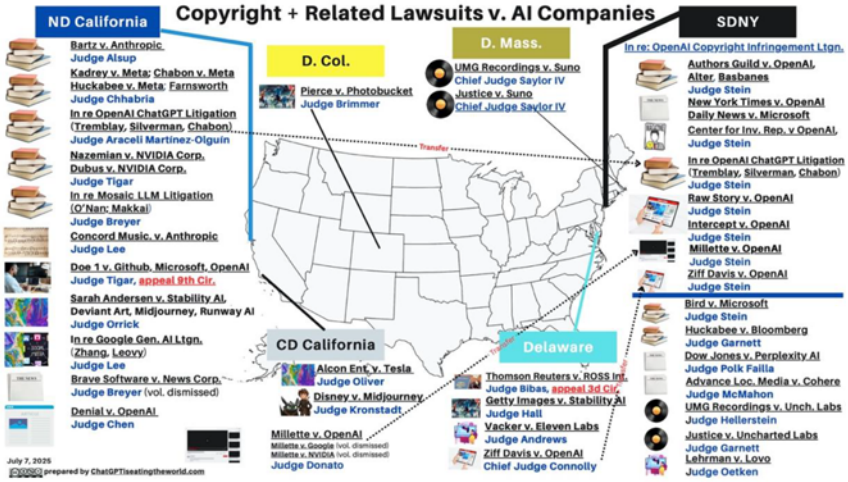
APPENDIX D: Courts' Rejection of Fair Use in Technology Cases

Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025) (table 4).

CASES	TECHNOLOGY DEVELOPMENT: Use of Copyrighted Works to Create New Technology?	TECHNOLOGY USAGE: Use of Copyrighted Works in Public Use of Technology?	FACTOR 1: Use of Copyrighted Works Had Further Purpose or Different Character?
American Broad. Co. v. Aereo, Inc., 573 U.S. 431 (2014).	No.	Yes, service enabled users to record TV shows using remote personal antennas and recording offered by online service.	[No court decision on fair use.]
Infinity Broad. v. Kirkwood, 150 F.3d 104 (2d Cir. 1998).	No.	Yes, retransmission of radio broadcasts over telephone.	No, service just "sell[s] access to unaltered radio broadcasts."
A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).	No.	Yes, file-sharing copies online.	No, file sharing of music files "does not transform the copyrighted work."
Video Pipeline v. Buena Vista Home Ent., 342 F.3d 191 (3d Cir. 2003).	No.	Yes, service made "clip previews" of Disney movies sold to retail websites selling home videos.	No, clip previews substituted for Disney's movie trailers.
Capitol Records, LLC v. ReDigi, Inc., 910 F.3d 649 (2d Cir. 2018).	No.	Yes, service makes copies of music files to facilitate resales of them online, while attempting to ensure deletion of seller's copy.	No, service enables "resale of digital music files, which resales compete with sales of the same recorded music by the rights holder."
U.S. v. ASCAP, 599 F. Supp. 2d 415 (S.D.N.Y. 2009).	No.	Yes, wireless service provider planned to offer previews of ringtones of copyrighted music without license.	No, wireless carrier's preview of ringtones served same purpose.
Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018)	Yes, to create a searchable database of TV and radio broadcasts	Yes, user views up to 10 minutes of recordings relevant to search topic.	Yes, but only modestly in allowing clients to time shift and to view what "they want at a time and place that is convenient.
Hachette Book Group v. Internet Archive, 115 F.4th 163 (2d Cir. 2024).	Yes, to create a searchable database of online library of books and other works, some of which are copyrighted	Yes, user given access to entirety of work.	No, service made "digital copies of the Works and distributes those copies to its users in full, for free."
UMG Recordings v. MP3.com, 92 F. Supp.2d 349 (S.D.N.Y. 2000).	No.	Yes, service copied "thousands of popular CDs in which plaintiffs held the copyrights, and, without authorization, copied their recordings onto its computer servers so as to be able to replay	No, service "simply repackages those recordings to facilitate their transmission through another medium."

		the recordings for its subscribers."	
Associated Press v. Meltwater U.S. Holdings, 931 F. Supp. 2d 537 (S.D.N.Y. 2013).	Yes, to create a searchable database for a news clipping service to provide clients with excerpts of news article	Yes, user receives 300-word excerpts of news articles relevant to searches, including email feeds	No, service "copies AP content in order to make money directly from the undiluted use of the copyrighted material" as a substitute of original works.
American Broad. Co. v. Aereo, Inc., 573 U.S. 431 (2014).	No.	Yes, service enabled users to record TV shows using remote personal antennas and recording offered by online service.	[No court decision on fair use.]

APPENDIX E: Map of U.S. Copyright Lawsuits v. AI Companies
Source: ChatGPT Is Eating the World



Testimony of Maxwell V. Pritt

Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training

Hearing Before the
Committee on the Judiciary
Subcommittee on Crime and Counterterrorism
United States Senate

July 16, 2025

Washington, DC

I. Introduction

Chairman Hawley, Ranking Member Durbin, and members of the Subcommittee: thank you for the invitation and opportunity to testify before you today.[1]

Today, the Committee considers what is likely the largest domestic piracy of intellectual property in our nation's history. That piracy includes hundreds of terabytes of data and many millions of works, including, for example, at least 12 books authored by members of this subcommittee (three authored by Chairman Hawley alone). The culprits of this unprecedented piracy are, incredibly, some of our nation's largest technology companies. They want vast troves of written text, including a limitless number of copyrighted works, to develop their artificial intelligence models. But these massively profitable tech companies don't want to pay for it. Perhaps they're Bob Dylan fans: "Steal a little and they throw you in jail / Steal a lot and they make you a king."[2] These decisions to engage in mass piracy were made

---

[1] I am a Partner at the law firm Boies Schiller Flexner LLP, and am in charge of the firm's San Francisco office. I litigate high-stakes cases for plaintiffs and defendants in courts across the country on issues ranging from antitrust to intellectual property to constitutional and contractual rights. I currently represent authors, artists, and programmers in copyright infringement cases against AI companies including Meta, OpenAI, GitHub, and Midjourney. I am a member of the Judicial Council of California and an Appellate Lawyer Representative for the U.S. Court of Appeals for the Ninth Circuit. I previously served on the California State Bar's Judicial Nominees Evaluation Commission, which vets state judicial candidates, and am Chair Emeritus of the Bay Area Lawyer Chapter of the American Constitution Society. I also taught at Stanford Law School and my alma mater, UC Law San Francisco. After law school, I clerked for Ninth Circuit Chief Judge Mary Murguia (on the district court) and Judge Marsha Berzon.

[2] Bob Dylan, "Sweetheart Like You" (Columbia Records 1983).

at the highest levels of the tech companies. At Meta, company documents show that Mark Zuckerberg himself made the call. Company documents at Anthropic also show a blatant disregard for our copyright laws, preferring to pirate books to avoid or delay the "legal/practice/business slog," as Anthropic's co-founder and CEO Dario Amodei put it.

AI companies now want a pass under a limited exception to infringement—the "fair use" doctrine—that Congress codified in Section 107 of the Copyright Act of 1976. But while these tech companies invoke fair use as a shield in litigation, they know piracy is illegal. As one Meta employee put it: "it's the piracy (and us knowing and being accomplices) that's the issue."[3] As others at Meta explained: "if there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues."[4]

As AI companies scrambled to outpace each other, many of them turned to illegal pirate websites—massive repositories of tens of millions of stolen copyrighted works—to get text for their AI models. By pirating these works for free rather than buying or licensing them from copyright owners, AI companies have built a multibillion-dollar industry generally without paying a single cent to either the creatives whose works are powering their products or the publishers responsible for

---

[3] *Kadrey et. al. v. Meta*, No. 3:23-CV-3417, Pl's Mot. for Partial Summary Judgment (N.D. Cal., April 30, 2025), Dkt. 574 at 7.

[4] *See Kadrey et. al. v. Meta*, No. 3:23-CV-3417, Pl's Opp. to Meta's Mot. for Summary Judgment (N.D. Cal., April 30, 2025), Dkt. 575 at 37.

introducing and providing those works to the public.[5] Tech companies' unapologetic use of these illegal websites has also revived international digital piracy by propping up the for-profit, criminal syndicates that use these illicit marketplaces to violate U.S. copyrights abroad.

The cost-benefit analysis was simple, particularly for tech companies caught flat-footed by OpenAI's unexpected release of ChatGPT: Expend time and resources to legally acquire the rights to copyrighted books and articles from those who own the rights; or pirate them all for free now from illegal websites and pay litigation damages later—or, even more appealing, pay nothing at all if they can convince the courts to excuse their unprecedented commercial piracy as fair use. The rapid rise of generative AI technology has thus ushered in a new era of domestic piracy on a scale never before seen and by extraordinarily powerful corporate interests seeking a quick profit off the backs of the creative industries that contribute over $1 trillion in value to the U.S. economy.

I. Illegal Pirate Websites and Peer-to-Peer File Sharing Networks

Digital piracy costs American businesses tens of billions of dollars annually. Each year, countless copyrighted works are made available, downloaded, and distributed from notorious pirate websites such as Library Genesis ("LibGen") and Z-Library through various methods, including decentralized peer-to-peer file-sharing systems like BitTorrent. BitTorrent is a technological method for downloading and

---

[5] Belying the AI companies' arguments, there is a robust and growing licensing market for AI training data. *See Kadrey v. Meta*, Dkt. 575 at 26.

uploading data over the internet. A key feature of BitTorrent is making available and "sharing" data—and to acquire data fast, you share it both simultaneously (called leeching) and after it is downloaded (called seeding). So as the latest pirated Taylor Swift album or Game of Thrones book is copied to your server, you're also making another copy and sending it out into the ecosystem to others. In this regard, torrenting is a system where "the *downloaders* of a file *barter* for chunks of it by uploading and downloading them in a tit-for-tat-like manner to prevent parasitic behavior."[6]

The Judiciary Committee last confronted the specific problem of peer-to-peer file sharing in 2000, when Napster and other music sharing platforms threatened to gut the music industry by offering a platform where users could exchange copyrighted songs for free and without permission of the rightsholders.[7] When the Ninth Circuit held that Napster's wide-ranging piracy violated U.S. copyright law,[8] the company was forced to shut down. Legitimate online music markets proliferated almost immediately: first, iTunes, and later, streaming services like Spotify.[9] For the next quarter of a century, Congress and the Executive worked with copyright holders to combat global piracy, which the U.S. Chamber of Commerce estimates costs the U.S.

---

[6] Johan Pouwelse et al., *The BitTorrent P2P File-Sharing System: Measurements and Analysis,* IPTPS 2005, at 206 (2005) (emphasis in original).

[7] *Music On The Internet: Is There An Upside to Downloading? Before the S. Comm. on the Judiciary,* 106th Cong. (July 11, 2000).

[8] *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004 (9th Cir. 2001).

[9] 1 Lindey on Entertainment, Publ. & the Arts § 2:28 n. 36 (3d ed. 2024).

economy nearly $30 billion in lost revenue each year.[10] As of 2017, before AI companies' mass piracy started, eBook piracy cost U.S. publishers $315 million in annual lost sales.[11]

The recent discovery of U.S. tech companies' mass piracy creates a seismic shift—from targeting criminal enterprises abroad to combatting piracy here at home. Pirate websites function similarly to Napster, but for books, rather than music. Whenever an individual uses an illegal pirate website in lieu of a legitimate bookseller to acquire a book, that book's author and publisher are directly harmed by the loss of a sale in an otherwise functioning and well-established market for books. A 2016 study indicated that pirated eBooks depress legitimate book sales by as much as 14%.[12]

Furthermore, as the Office of the U.S. Trade Representative documented in its survey of "Notorious Markets for Counterfeiting and Piracy,"[13] which included LibGen in two recent editions, the harms of piracy are far-reaching, impacting not just vibrant U.S. economies reliant on legitimate markets. In the case of books,

---

[10] U.S. Chamber of Commerce, *Impacts of Digital Piracy on the U.S. Economy*, (June 15, 2019) https://www.uschamber.com/technology/data-privacy/impacts-of-digital-piracy-on-the-u-s-economy.

[11] Adam Row, *U.S. Publishers Are Still Losing $300 Million Annually To Ebook Piracy*, FORBES (July 28, 2019).

[12] Imke Reimers, *Can Private Copyright Protection Be Effective? Evidence from Book Publishing* (2016), 59 J. L. & ECON. 411, 414 (2016) https://doi.org/10.1086/687521.

[13] *See* Office of the U.S. Trade Representative, *USTR Releases 2024 Review of Notorious Markets for Counterfeiting and Privacy* (Jan. 8, 2025) https://ustr.gov/about/policy-offices/press-office/ustr-archives/2007-2024-press-releases/ustr-releases-2024-review-notorious-markets-counterfeiting-and-piracy.

internet piracy harms artists, graphic designers, bookstores, publishers, printing presses, copy editors, and others working in creative economies. This rampant piracy undermines critical U.S. comparative advantages by hampering innovation and creativity.[14]

Since 2015, federal courts have consistently held that pirate websites such as LibGen violate U.S. copyright law.[15] Pirate websites are also routinely targeted by government enforcement actions. Authorities regularly shut down their domains and even prosecute the perpetrators.[16] It is hard to reconcile from any good-faith policy or legal perspective how illegal websites that traffic in piracy and are permanently enjoined by federal courts can simultaneously exist as legitimate sources for AI companies.

II. Fair Use & Trends in AI Litigation

The last few years have demonstrated both the potential and pitfalls of generative AI. While there is no evidence yet to support the companies' far-reaching marketing claims such as curing disease, the data analysis capabilities of the machines are impressive. In addition to large language models ("LLMs"), image,

---

[14] *Digital Copyright Piracy: Protecting American Consumers, Workers, and Creators, Hearing Before the Subcomm. on Courts, Intel. Prop., and the Internet of the H. Comm. on the Judiciary* 118th Cong. 68 (December 13, 2023) (Statement of the Association of American Publishers, at 3).

[15] *Elsiever Inc. v. www.Sci-Hub.org*, 2015 WL 6657363 (S.D.N.Y. 2015); *see also Cengage Learning, Inc. v. Does 1-50 d/b/a Library Genesis*, Case No. 23-cv-8136-CM, Dkt. 36 at 2-3 (S.D.N.Y. Sept. 24, 2024) (granting permanent injunction against LibGen for copyright infringement).

[16] *See* Indictment, *United States v. Napolsky et al.*, No. 1:22-CR-525 (E.D.N.Y. Nov. 16, 2022), Dkt. 4 (criminal indictment of Z-Library perpetrators).

video, and music generators are beginning to proliferate. But the more we learn about generative AI technology, the more it has become clear that many AI companies have cut corners in their race to be the biggest and the best. Under competitive pressure, AI companies opted to play fast and loose with our intellectual property because they know how valuable it is but think they can get away with not paying for it by claiming "fair use."

Training data is the raw material that powers the AI industry. AI-generated outputs are a direct product of the data on which AI models train. Training on scientific articles, like medical journals, is what allows LLMs to analyze medical issues. Training on software code is what allows LLMs to generate code for programmers. And training on literary works is what allows LLMs to write creatively.

AI companies have long recognized the value of copyrighted books and long-form text as training data for LLMs. But instead of buying and licensing books from authors or publishers, massive tech companies like Meta, OpenAI, and others abandoned efforts to acquire books through legitimate means and instead turned to piracy, sometimes using peer-to-peer file sharing networks to acquire books and, in the case of Meta, even copying and distributing vast amounts of pirated data to others.

A. Exemplar Case: Meta's Mass Piracy

Much of Meta's exploitation of pirated copyrighted works is now public through the efforts of my firm and our co-counsel on behalf of authors. From the very early

days of its GenAI program, Meta concluded that training its LLM, called Llama, on books would improve its performance. Meta also believed its competitors were obtaining superior results due to their use of pirated datasets to train their LLMs.[17] In response, Meta devised a simple strategy to catch up: acquire more books—and fast.[18] In October 2022, Meta's AI team began seeking in-house legal approval for "pure exploration work" regarding the performance benefits that could be achieved by training Llama on books and articles obtained from LibGen.[19] Meta initially planned to use LibGen only to test its value and then set up licensing agreements.[20] When Meta's legal team approved the plan, Meta downloaded 2.2 million books from LibGen via torrenting.[21] Unsurprisingly, books proved valuable and improved model performance.

In early 2023, Meta scrambled to gather more text data for subsequent iterations of the Llama models.[22] Managers stressed the need to get as many books as possible as quickly as possible.[23] Over the next couple of months, Meta briefly

---

[17] *Kadrey v. Meta*, Dkt. 574 at 6.

[17] *Id.*

[18] *Id.*

[19] *Id.*

[20] *Id.*

[21] *Id.* at 6-7.

[22] *Id.* at 7.

[23] *Id.* at 8.

explored licensing books.[24] Licensing budgets as high as $200 million were floated.[25] But then, in April 2023, Meta's short-lived licensing efforts came to a halt.[26] Meta executives instructed Meta's business development team—the team tasked with licensing—to stop all text licensing efforts.[27] And Meta instead resorted to using pirated copyrighted works from LibGen, which it continued to copy and use as a key component of its commercially available Llama models.[28] To conceal its actions, Meta's employees started referring to these pirated datasets as "external" or "publicly available" instead of "pirated."[29] Then, in May 2023, Meta employees torrented even more copyrighted works from LibGen, this time using LibGen's contents to cross-reference against major publishers' catalogues to determine whether licensing efforts would be necessary at all.[30] This "gap approach" showed that Meta viewed pirated books simply as a free substitute for licensed books. And why license if it could pirate for free?

After discovering that LibGen contained most of the books it needed, Meta's executives greenlighted its use as training data for Llama. Documents uncovered during litigation confirmed that after an escalation to Meta CEO Mark Zuckerberg,

---

[24] *Id.*

[25] *Id.*

[26] *Id.* at 8.

[27] *Id.* at 8-9.

[28] *Id.* at 9.

[29] *Id.*

[30] *Id.*

the GenAI team was "approved to use LibGen for Llama 3."[31] That internal memorandum, which documented Mr. Zuckerberg's approval, admits that the approval came despite the fact that LibGen is "a dataset we know to be pirated."[32] Notably, that memo also included in-house counsel, who apparently advised that "in **no case** would we disclose publicly that we had trained on libgen," because "if there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues."[33]

When an August 2023 exposé published in *The Atlantic* revealed broadly that Meta had been training on copyrighted works without permission, one employee worried that the public could realize that Meta was continuing to use pirated data for training: "It's the piracy (and us knowing and being accomplices) that's the issue," she remarked.

Meta went to great lengths to conceal its use of illegal websites and pirated copyrighted works. Senior leadership and engineers knew, but it was only conveyed to others on a "'need to know' basis."[34] Indeed, several employees expressed strong reservations. One Meta researcher commented, "I feel that using pirated material should be beyond our ethical threshold."[35] Another referred to LibGen as an "illegal

---

[31] *Id.* at 10.

[32] *Id.*

[33] *Id.* (Ex. 61) (emphasis in original).

[34] *Id.* at 7.

[35] *Id.* at 10.

pirated website[ ]" and expressed that "it should not go in the training of the published model."[36]

But as 2023 progressed and Meta's models increased in size, Meta's hunger for high-quality data increased. In late 2023, as development started for Llama 4, Meta's engineers explored the use of Anna's Archive, an aggregator of many illegal pirated datasets. Meta engineers confirmed that Anna's Archive contained substantially all of LibGen, nearly all of Z-Library, and over two-thirds of an additional database called Internet Archive.[37] So Meta began downloading and processing copyrighted works from the Anna's Archive aggregation of illegal pirate websites, despite employees calling it a "pretty shady website" that "won't be popular with the lawyers."[38]

To speed up its acquisition of terabytes upon terabytes of pirated works—and ultimately to keep pace with its pirate website-using competitors—Meta resorted to torrenting, the peer-to-peer file sharing protocol discussed above. This protocol optimizes speed by simultaneously making available and transmitting content being downloaded to others.[39] Meta's engineers were well aware of the legal risks posed by torrenting data, but they decided to do it anyway, apparently with the approval of Meta's in-house counsel.[40] One employee told others: "Btw, it would not be trivial to

---

[36] *Id.*

[37] *Id.* at 11.

[38] *Id.*

[39] *Id.* at 11-12.

[40] *Id.* at 12.

download libgen if everything is in torrents," sharing a link to a Quora webpage asking, "What is the probability of getting arrested for using torrents in the USA?"[41] Meta opted to hedge some of this legal risk—and the risk of law enforcement or public discovery generally—by using Amazon Web Services ("AWS") for its torrenting activities, a deviation from Meta's usual practice.[42] When an engineer asked why Meta infrastructure could not be used, he was told it was to avoid risk of tracing the torrenting to Meta's servers.[43] And, although Meta could have changed the default settings on its torrent client to completely prevent uploading, it did not do so, presumably because changing those settings would decrease the speed of Meta's downloads.[44]

Meta's embrace of data piracy increased exponentially over time. Between April and July 2024 alone, Meta downloaded over **134 terabytes** from pirate websites through Anna's Archive and uploaded over **40 terabytes** of pirated data to third parties.[45]

Meta has attempted in litigation to justify its piracy by arguing it had no choice but to pirate the books because legitimate acquisition was prohibitively expensive or time-intensive. But ability-to-pay was never an obstacle for a company like Meta. This massive piracy occurred at the hands of one of the world's richest companies,

---

[41] *Id.*

[42] *Id.* at 14.

[43] *Id.* at 14-15.

[44] *Id.* at 14

[45] *Kadrey v. Meta*, Dkt. 562-50 at 4.

which has invested huge sums of money into other aspects of its AI program, including data centers and talent. Meta plans to spend hundreds of billions of dollars to build AI data centers[46] and reportedly has offered gargantuan recruitment bonuses to individual data scientists,[47] yet Meta has not paid a single cent for the copyrighted works it pirated.

B. Other U.S. Tech Companies' Piracy

Meta is far from alone in its conduct. Piracy has become endemic to the GenAI industry. Anthropic, the company behind the LLM known as Claude, also pirated millions of books to build its LLM. A recent decision in *Bartz v. Anthropic* noted that the company downloaded over seven million pirated copies of books from LibGen and another illegal website called Pirate Library Mirror and paid nothing.[48] OpenAI used LibGen too.

By downloading millions upon millions of books and other copyrighted works from pirate websites and then using those unauthorized copies for their AI products, these companies have committed copyright infringement on a massive scale. These historic acts of domestic piracy have deprived the U.S. creative industry of billions of dollars. Further, Meta and any other company that used a torrenting network like BitTorrent to source pirated works has perpetuated the copyright infringement there

---

[46] Jaspreet Singh and Aditya Soni, *Meta's Zuckerberg pledges hundreds of billions for AI data centers in superintelligence push,* REUTERS (July 14, 2025).

[47] *Sam Altman says Meta offered $100 million bonuses to OpenAI employees*, REUTERS (June 18, 2025).

[48] *Bartz v. Anthropic*, No. 3:24-CV-5417 (N.D. Cal. June 23, 2025), Dkt. 231 at 18.

on both a massive and exponential scale by making and distributing additional copies of pirated works to others on the network, who in turn could distribute those works to other network participants, and so on. In fact, one major pirate websites boasts that AI companies saved the illicit digital books market: "Not too long ago, 'shadow-libraries' were dying. Sci-Hub, the massive illegal archive of academic papers, had stopped taking in new works, due to lawsuits. 'Z-Library', the largest illegal library of books, saw its alleged creators arrested on criminal copyright charges. They incredibly managed to escape their arrest, but their library is no less under threat . . . . Then came AI. Virtually all major companies building LLMs contacted us to train on our data."[49]

III. Piracy Is Incompatible with "Fair Use"

Meta and other AI companies now seek to defend their massive, systemic infringement in court. In defense, the AI companies argue their infringement was "fair use."

Fair use is an *exception* that allows for *limited* use of copyrighted material without permission from the copyright owner under certain conditions. The doctrine is meant to balance the rights of creators with the public interest in freedom of expression and access to information. For example, a musician who creates a parody of a song usually will not be liable for copyright infringement because the fair use doctrine protects her. Why? Because parodies are considered sufficiently

---

[49] Anna's Archive, *Copyright reform is necessary for national security* (January 31, 2025), https://annas-archive.org/blog/ai-copyright.html.

"transformative" (among other reasons, a parody cannot exist without using the copyrighted work it parodies) that they do not operate as a substitute for the original and in so doing, do not harm the market for the original. While courts analyze fair use under four non-exhaustive factors, the analysis often focuses on questions of the purpose of the copying, whether it is transformative, and "market substitution" for actual or potential markets of *each use of each copy* the alleged infringer makes of the work.

But internet piracy is the antithesis of fair use. Downloading millions of pirated works from known illegal databases created by foreign actors to avoid compensating American creators and rightsholders is not "transformative": it does nothing to create new expression or facilitate meaningful dialogue. Internet piracy also functions as pure market substitution—taking for free what otherwise must be purchased.

For over a century, courts have held that unmitigated piracy of copyrighted works, *i.e.*, the duplication of entire works to avoid compensating rightsholders, is not fair use.[50] That through-line has been applied to digital piracy—the illegality of downloading and sharing copyrighted material has been well-established since the

---

[50] See, e.g., *Folsom v. Marsh*, 9 F. Cas. 342, 342-45 (C.C.D. Mass. 1841) (explaining "it is as clear, that if [defendant] thus cites the most important parts of the work, with a view, not to criticise, but to supersede the use of the original work, and substitute the review for it, such a use will be deemed in law a piracy") (Story, J.); see, e.g., *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 550 (1985) ("As Justice Story's hypothetical illustrates, the fair use doctrine has always precluded a use that 'supersede[s] the use of the original.'").

days of Napster.[51] The uncontroversial premise behind this conclusion is that for fair use to apply, the work that was copied must have been lawfully obtained in the first place.[52]

Whatever the merits of generative AI, stealing copyrighted works off the internet for one's own benefit has always been unlawful.[53] While everyone expected these historic AI copyright cases to test the boundaries of the "fair use" defense with respect to training AI models on copyrighted works, no one expected that billion- and trillion-dollar companies would be arguing—and courts would be contemplating—that mass piracy for commercial use could also fall within the ambit of the fair use defense.

As the U.S. Copyright Office recently explained in a seminal report on copyright and AI, "making commercial use of vast troves of copyrighted works to produce expressive content that competes with them in existing markets, *especially where this is accomplished through illegal access*, goes beyond established fair use boundaries."[54] Deeming rampant digital piracy fair use would mark the first time in

---

[51] See, e.g., *Metro-Goldwyn-Mayer Studios Inc. v. Grokster*, Ltd., 545 U.S. 913, at 919 (2005) ("[O]ne who distributes a device with the object of promoting its use to infringe copyright . . . is liable for the resulting acts of infringement by third parties."); *see also United States v. Slater*, 348 F.3d 666, 669 (7th Cir. 2003).

[52] *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832, 843 (Fed. Cir. 1992) ("To invoke the fair use exception, an individual must possess an authorized copy of a literary work.") (emphasis added).

[53] *UMG Recordings, Inc. v. MP3.Com, Inc.*, 2000 WL 710056, at *1 (S.D.N.Y. June 1, 2000) (the "mere fact" that copyright infringement is "clothed in the exotic webbing of" a new technology "does not disguise its illegality") (Rakoff, J.).

[54] United States Copyright Office, Report on Copyright and Artificial Intelligence: Part 3 (May 9, 2025) (pre-publication version) (emphasis added).

history that piracy and trafficking in stolen goods are given a pass under U.S. copyright law. As Judge Alsup put it: "There is no carveout [] from the Copyright Act for AI companies."[55]

IV. AI Companies' Piracy Is a Bipartisan Issue

Tech companies' mass digital piracy affects everyone irrespective of political persuasion. This is a bipartisan issue. For example, because Meta torrented LibGen and Anna's Archive, we know it pirated the following books written by members of this very Subcommittee:

- "The Tyranny of Big Tech," by Senator Josh Hawley
- "Antitrust: Taking on Monopoly Power from the Gilded Age to the Digital Age," by Senator Amy Klobuchar
- "United: Thoughts on Finding Common Ground and Advancing the Common Good," by Senator Cory Booker
- "A Time for Truth: Reigniting the Promise of America," by Senator Ted Cruz
- "Life Equity: Realize Your True Value and Pursue Your Passions at Any Stage in Life," by Senator Marsha Blackburn
- "Manhood: The Masculine Virtues America Needs," by Senator Josh Hawley
- "Theodore Roosevelt: Preacher of Righteousness," by Senator Josh Hawley
- "Justice Corrupted: How the Left Weaponized Our Legal System," by Senator Ted Cruz

---

[55] *Bartz v. Anthropic*, Dkt. 231 at 14.

- "One Vote Away: How a Single Supreme Court Seat Can Change History," by Senator Ted Cruz
- "Unwoke: How to Defeat Cultural Marxism in America," by Senator Ted Cruz
- "The Joy of Politics: Surviving Cancer, a Campaign, a Pandemic, an Insurrection, and Life's Other Unexpected Curveballs," by Senator Amy Klobuchar
- "The Senator Next Door: A Memoir from the Heartland," by Senator Amy Klobuchar

Meta also pirated books written by every President and Vice President in the 21st century, including:

- "The Art of the Deal," by President Donald Trump
- "Hillbilly Elegy: A Memoir of a Family and Culture in Crisis," by Vice President J.D. Vance
- "Promise Me, Dad: A Year of Hope, Hardship, and Purpose," by President Joseph Biden
- "The Truths We Hold: An American Journey," by Vice President Kamala Harris
- "So Help Me God," by Vice President Mike Pence
- "A Promised Land," by President Barack Obama
- "Decision Points," by President George W. Bush

- In My Time: A Personal and Political Memoir," by Vice President Dick Cheney

Meta did not pay for its use of any of these books. Moreover, Meta's indiscriminate torrenting means that countless copies of these books were made and distributed by Meta to other internet pirates—each of which constituted yet another lost sale.

V. Conclusion

In conclusion, I would like to address a common refrain stated publicly by some AI companies—that in order to keep pace with China, the United States has no choice but to excuse widespread digital piracy as fair use. That premise is simply untrue. Protecting intellectual property rights is a core American value and has been one since our founding. As Senator Leahy, who once chaired the Judiciary Committee, said: "the intellectual property generated in our country is the envy of the rest of the world."[56] U.S. intellectual property law has always fostered innovation, not hindered it.

This nation's commitment to protecting intellectual property—including U.S. tech companies' IP—has made us more competitive, not less. The threat of innovation in foreign countries has never been grounds for this country to excuse rampant lawbreaking and abandon fundamental principles of the rule of law. Nor should it be now.

---

[56] *Music On The Internet: Is There An Upside to Downloading? Before the S. Comm. on the Judiciary,* 106th Cong. (July 11, 2000). *See also* U.S. Chamber of Commerce, 2023 International IP Index.

To be clear, a policy of responsible AI simply requires adherence to our nation's most foundational legal principles and the existing laws governing intellectual property rights. No one is above the law—and certainly not billion- and trillion-dollar tech companies. Congress can—and must—promote AI growth and innovation to secure U.S. dominance while also promoting the progress of science and the arts. To maintain U.S. dominance across cultural output in the arts and sciences, Congress must protect our creative industries by, at the very least, aiding enforcement of the copyright laws as they already exist on the books against IP pirates, without favoring those who have shown they are most capable of infringing copyrights on a massive scale.

Written Testimony of Dr. Michael D. Smith
J. Erik Jonsson Professor Of Information Technology And Policy,
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Senate Committee on the Judiciary
Subcommittee on Crime and Counterterrorism

Too Big to Prosecute?:
Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training

July 16, 2025

Introduction:

Chairman Hawley, Ranking Member Durbin, Distinguished Members of the Subcommittee, thank you for giving me this opportunity to testify on "Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training."

My name is Michael Smith and I am the J. Erik Jonsson Professor of Information Technology and Policy at Carnegie Mellon University's Heinz College of Public Policy Management.

My testimony today is informed by over 25 years of empirical research into the impact of technological change on economic markets for creative content. It is also informed by my experience serving on a roundtable of 10 economists convened by the U.S. Copyright Office to study the implications of gen AI on copyright policy.

Economic Evidence on Digital Piracy:

My research on this question started in the early 2000s when digital piracy was a relatively new problem for the creative industries. During that period many in the tech community—including many piracy platforms—argued that piracy was fair use because it would not harm legal sales, was unlikely to harm creativity, and any legislative efforts to curtail piracy would not only be ineffective but would also stifle innovation.

My empirical research over the past 25 years has contributed to a large economic literature studying these three questions. My colleagues Brett Danaher, Rahul Telang and I summarized the findings from this literature in a 2020 Piracy Landscape Study for the U.S. Patent and Trademark Office. Our report drew three broad conclusions:

> First, the peer-reviewed academic literature shows that digital piracy harms creators by reducing their ability to make money from their creative efforts.
>
> Second, the peer-reviewed academic literature shows that that digital piracy harms society by reducing economic incentives for investment in creative output.
>
> Third, the peer-reviewed academic literature shows that legislative interventions implemented worldwide have been effective in reversing these harms to the creative community—while also allowing internet businesses and other legitimate online distribution platforms to flourish.

Applying These Economic Principles to Piracy and Generative AI Training:

In the context of gen AI training, we are now hearing many of the same arguments that we heard in the early days of the Internet: Allowing generative AI companies to use pirated content to train their models is fair use because it won't harm legal sales, is unlikely to harm creativity, and any legislative efforts to curtail the use of pirated materials for training will not only be ineffective, but will also stifle innovation.

It's important to recognize that while the time has changed, the underlying economic principles are the same today as they were in 2000. I think we can learn a great deal from applying those economic principles to today's question. Indeed, I think we'll find many of the same results.

The use of pirated content to train generative AI models will harm sales for creators: Allowing generative AI companies to train their models with pirated content is likely to harm sales for creators in two key ways.

First, the nature of BitTorrent networks is that when someone downloads a file from the network, they also share back pieces of the file to other people downloading the file. This not only increases the download speeds for the person—or in the case of Meta, the company—downloading the

pirated file. It also increases the download speeds for everyone else downloading the file on BitTorrent—making piracy a more attractive option to legal purchases. The economic literature shows that making it easier for consumers to download pirated content will cause direct harm by reducing sales in the legal market.

Second, allowing gen AI companies to obtain unlicensed training data through P2P pirate networks will also harm the market for licensed content. The Copyright Alliance has documented over 70 licensing contracts between gen AI companies and rightsholders including HarperCollins, Universal Music, Reddit, Shutterstock, and the *Wall Street Journal*. So it's clear licensing markets between rightsholders and gen AI companies can work!

However, there are two key economic problem with the current market. The first is that gen AI companies are disincentivized from signing licenses for fear of damaging their fair use defense in court. Discovery in the Kadrey case revealed one Meta employee saying "The problem is that people don't realize that if we license one single book, we won't be able to lean into fair use strategy." The second problem with the current markets is that the sellers are negotiating with a gun held to their head: In a world where gen AI companies know they can pirate with impunity, they can tell sellers the equivalent of "accept my terms or I will just steal your content and you'll get nothing." Imagine how much we could improve outcomes in these markets if we created an environment where buyers were no longer disincentivized from signing licenses, and where buyers and sellers were negotiating on equal terms?

The use of pirated content to train generative AI models will harm society by reducing economic incentives for creators: The economic principle in the early days of piracy—that when creators can make less money, society will see less creative output—holds here as well. It turns out the Founders were onto something when they included Article 1, Section 8, Clause 8 in the U.S. Constitution, giving Congress the power to "promote the Progress of Science and useful Arts."

But there's a new and unique indignity to our current situation: When piracy is used to train gen AI models, we are not only stealing from creators, we are then using the theft of their content to create tools that can flood the market with machine-generated creative output, which could in turn replace many of those creators.

That nightmare scenario for creators—stealing my past creative output to eliminate my future creative output—is not hard to imagine. Already industry leaders and academic research has shown that gen AI tools have replaced workers—particularly "entry level" workers—in other important sectors of the economy. It's perfectly reasonable to believe that gen AI tools will do the same for creative artists, particularly emerging artists—the very people who otherwise would create the next new thing that can benefit our creative ecosystem.

I want to be very clear here: I'm not opposed to technological innovation, but what I've seen in my research is that technological innovation needs to be sustainable. I worry that in our current environment the short-term interests of gen AI companies will come at the expense of the long term interests of a sustainable creative market—a market that benefits creators, society, and the innovation that gen AI companies could create from a sustainable creative market.

The use of pirated content to train generative AI models will also harm creators, technology firms, and lawful society by creating perverse market incentives: The use of piracy to train generative AI models also has the potential to create problems for creators, generative AI companies—and I would argue a lawful society—by putting into place a set of perverse incentives.

One notable perverse incentive is the market incentive to steal instead of license content. Discovery in the Kadrey case, and other similar cases, shows that Meta was pirating content to train their models in part because it believed that everyone else in Silicon Valley was using pirated content to train their models. And indeed there is ample evidence from other cases filed in the courts that Meta was right, many other firms were training their models with pirated content. In short Meta believed that it needed to break the law to maintain competitive parity with its rivals. That's not something we should want to incentivize.

The unrestricted use of pirated content to train gen AI models creates another perverse inventive: The incentive for gen AI firms to launder otherwise licensable content through pirate networks. If training on pirated data is considered legal, then gen AI firms, will have strong incentives to add new content to online repositories of stolen works—content that otherwise would not have been available. Indeed, this the unlicensed use of pirated content could create a new illicit licensing business model for pirate networks: adding new stolen content to their collections, knowing that AI developers will want access to them.

A third notable perverse incentive is the incentive for rightsholders to remove their content from the open web in ways that would harm both society, and the business models of rightsholders. Today rightsholders make free content available as a way to support sales in their other paid or advertising-supported channels. In a world where that content can be used to train gen AI models, rightsholders will have reduced incentives to provide free content—to the detriment of their existing business models and the detriment of the open web. Indeed, Cloudflare's recent announcement that it will block A.I. data scrapers by default for its clients could be the salvo in the war between gen AI harvesting and the open web.

Interventions can reverse these harms: We can—and should—respond to these threats. Consider a world where the Napster and Grokster cases went the other direction—a world where sharing pirated content was "fair use" and was allowed to exist legally. In that world there's a strong argument that licensed markets for creative content like Spotify and Netflix wouldn't exist today — to the detriment of consumers, creators, and technology investors.

I think today we have a similar opportunity to create a win-win-win for society, creators, and tech firms by making it clear that piracy is wrong, and that a vibrant technology economy depends on a vibrant creative economy. We found a way to make licensed streaming and sales channels work for consumers, copyright owners and platforms in the early 2000s, and we must do the same for generative AI.

In short I think we have the potential to create a sustainable system where creators have the economic incentives to continue to innovate, and where consumers benefit from those innovations both directly and through a vibrant generative AI system that is trained on that creative output.

As I said in a recent Harvard Business Review article, Gen AI has the potential to benefit industry and society in many ways. But achieving that potential will require a more robust and transparent partnerships between technology firms and the creative industries. On our current path we risk killing the goose—or in this case the authors, musicians, coders, and filmmakers—who laid the golden eggs that are key to the present and future value of gen AI output.

Statement of
BHAMATI VISWANATHAN
Assistant Professor, New England Law | Boston

before the

SENATE COMMITTEE OF THE JUDICIARY
SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM

July 16, 2025

Bhamati Viswanathan, Assistant Professor of Law at New England Law | Boston, submits this statement for the record concerning the hearing titled Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training before the Senate Judiciary Committee, Subcommittee on Crime and Counterterrorism, on July 16, 2025.

The Powerful and Robust U.S. Creative Economy Is Being Irreparably Harmed by the Proliferation of Digital Piracy Undertaken By So-Called Shadow Libraries

The arts and cultural economic activity in the US, as estimated by the U.S. Bureau of Economic Analysis accounted for 4.2 percent of GDP, or $1.17 trillion, in 2023. It is dependent on strong copyright protection.[1] In Article 1 Section 8 Clause 8, the U.S. Constitution establishes this right, creating an incentive structure that creators rely on. Innovation is the goal of copyright, as set forth in the Constitution: "to promote progress in Science and the Useful Arts."

Piracy circumvents that balance. Piracy is the consumption of unlicensed copyrighted products, differing from counterfeiting, which is the consumption of unlicensed trademarked products. Digital piracy, specifically, mirrors supply chain for physical pirated goods in that intermediaries facilitated discovery of pirated contents by consumers. Here, the distribution of content form providers to consumers, and the flow of payments from consumers to both platforms and providers. However, differing from physical piracy, digital piracy does not require manufacturing steps and is distributed virtually, reducing cost and increasing scope and scale of digital piracy operations.[2]

These shadow libraries play significant roles as illicit actors and continue to be pursued by the Federal Bureau of Investigations (FBI) and the Department of Homeland Security (DHS). For example, in 2022, the FBI seized domains associated with Z-Library and charged two of its operators with criminal copyright infringement, wire fraud, and money laundering.[3] Likewise, in its Operations Intangibles, U.S. Immigration and Customs Enforcement's (ICE) of the DHS have also outlined its commitment to "stop digital piracy and eliminate a vital source of illicit revenue from transnational criminal organizations," citing that these activities have continued to feed a

[1] Arts and Cultural Production Satellite Account, U.S. and States, 2023 | U.S. Bureau of Economic Analysis (BEA)
[2] Brett Danaher, Michael D. Smith, and Rahul Telang, Piracy Landscape Study: Analysis of Existing and Emerging Research Relevant to Intellectual Property Rights (IPR) Enforcement of Commercial-Scale Piracy , https://www.cmu.edu/entertainment-analytics/documents/uspto.pdf
[3] Federal Law Enforcement Arrests and Indicts Z-Library Operators with AG's Assistance - The Authors Guild

criminal enterprise whose profits are used to support other organized criminal endeavors, including violent crime and trafficking.[4]

Generative AI Companies Are Relying on Pirated Materials to Build Their Large Language Models and Thereby Augmenting the Harms That Shadow Libraries Cause

GenAI companies are required to ingest vast amounts of materials in order to build robust large language models (LLMs). This is because LLMs are essentially sophisticated prediction models: they learn structure, syntax, speech patterns, and other linguistic foundations, and then "predict" language sequences based on their acquired learning. GenAI companies must find these vast amounts of materials from available digital sources, databases, and repositories.

It is clear that GenAI companies ingest works from pirate sources.[5] When LLM models are trained on pirated works, they circumvent copyright law by training on works that have already been illicitly reproduced, digitized, and distributed. Evidence shows that GenAI companies have willfully, knowingly and repeatedly trained on pirated materials, despite being aware that their source shadow libraries are circumventing the law to obtain and share those materials.[6]

This is a crime compounding a crime: the initial crime is the illicit copying, making available, and distribution of materials under copyright; and the compounded crime is the ingestion and use of these materials in the creation and development of LLMs by GenAI companies.

When Generative AI Companies Ingest Pirated Materials, They Directly Harm Copyright Holders by Undermining Their Rights and Usurping Their Rewards

The training of LLMs on pirated materials is far from a "victimless" crime. Authors, artists, filmmakers, and photographers are among the creators whose works are taken and used without permission or payment.[7] Publishers, film producers and distributors, newspapers, and media outlets are among the intermediaries whose commercial services are usurped, also without permission or payment.[8] These are real victims: they relied on well-established copyright laws to protect their original works,[9] only to have those works taken *en masse* to build LLMs that in turn can enable the mass production of infringing works. And this harm is more than hypothetical: there is a direct relationship between the rise of e-book piracy and the decline in authors'

---

[4] Operation Intangibles | ICE

[5] *E.g.*, *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC, Dkt. No. 567-45 (Meta email noting "GenAI has been approved to use LibGen for Llama 3" despite acknowledging that LibGen is "a dataset we know to be pirated"); Dkt. No. 567-25 (Meta employee stating that "It's the piracy (and us knowing and being accomplices) that's the issue."); Dkt. No. 567-21 (Meta employee stating, "I feel that using pirated material should be beyond our ethical threshold.").

[6] *Id.*

[7] U.S. Chamber of Commerce, Impacts of Digital Piracy on the U.S. Economy (June 2019), https://www.uschamber.com/technology/data-privacy/impacts-of-digital-piracy-on-the-u-s-economy.

[8] *Id.*

[9] The Authors Guild, Piracy, https://authorsguild.org/advocacy/piracy/, ("Each year, the publishing industry loses hundreds of millions of dollars in lost sales to piracy—and with each lost sale, authors lose royalty income.").

income.[10] Piracy does not just deprive rightsholders of the fruits of their labor, it also erodes morale and trust in the creative sector and normalizes theft of intellectual property, diminishing incentives for future innovation. The business models of entire creative industries are at risk.

When Generative AI Companies Ingest Pirated Materials, They Contribute to and Fuel the Proliferation of Shadow Libraries

Digital shadow libraries directly benefit from the ingestion activities of GenAI companies. When GenAI companies mine their work, they drive digital traffic to the libraries. Further, in at least one case, the shadow libraries derive direct benefits from GenAI companies. In one notable case, a shadow library known as Anna's Archives openly offers to work with AI companies in exchange for a "donation" or data trades. Contrary to the view of at least one district court judge, this offers to trade access to pirated materials for money and/or data indicates that pirate libraries can engage in symbiotic relationships with GenAI companies. In sum, the training of GenAI on materials can contribute to the proliferation and growth of digital piracy. These shadow libraries have continued to proliferate, with some of the largest such as Library genesis now claiming to have more than 2.4 million non-fiction books, 80 million science magazine articles and Anna's Archive with 36 million books and 103 million academic papers. [11]

Innovation Can Be Fostered, But Not at The Expense of Fair Compensation of Creative Economy Stakeholders and Support of Creative Markets

It is widely agreed that GenAI companies are the engines of innovation, and their emerging technologies hold great promise of enhancements in every area of life. None of the proposals raised here are intended to hamper such vital innovation. Indeed, none would hinder productive innovation: GenAI companies have the means to license uses of works just as every user of creative works has licensed uses since copyright was put into place. The music industry, to take just one example, is built on a system of rights clearances and licensing arrangements. Similarly, the film industry regularly engages in complex cross-licensing; as do the biotech, biomed, and pharma industries. Licensing works is standard business practice in every creative industry, and with good reason: it enables markets to operate efficiently and at optimal productivity.

Copyright itself exists to boost innovation, and to incentivize risk-taking in commercial markets. Innovation is the goal of copyright, as set forth in the Constitution: "to promote progress in Science and the Useful Arts." To the contrary, training LLMs on pirated works subverts the copyright system and fosters illicit activity that costs industries millions of dollars in revenues. This runs counter to innovation, as it disincentivizes creation, invention and discovery, and commercial productivity. GenAI companies do not need to train their models on illicit materials. There is already a thriving marketplace for the works they need to ingest; and they have the means to participate in the market for creative works just as every other user-consumer does on a daily basis.

---

[10] The Authors Guild, *Authors Guild Survey Shows Drastic 42 Percent Decline in Authors Earnings in Last Decade* (January 5, 2019), https://authorsguild.org/news/authors-guild-survey-shows-drastic-42-percent-decline-in-authors-earnings-in-last-decade/.

[11] https://greycoder.com/a-list-of-the-largest-shadow-libraries/

Generative AI Companies Are Arguably Engaged in Willful Acts That May Rise to the Level of Criminal Copyright Infringement

Training of GenAI on pirated materials promotes copyright infringement at two stages: the initial acts of digital piracy and the subsequent acts of training on materials under copyright without permission, licensing, or other licit forms of use. During the ingestion stage of these illicit practices, GenAI companies have knowingly, intentionally, and willfully chosen to circumvent copyright law and policy through their recourse to pirate repositories.

Historically, practices circumventing copyright protection have been successfully indicted on the basis of criminal copyright infringement, particularly where the actions were made with willful knowledge of such infringement.[12] Criminal copyright infringement requires a finding that copyright infringement was undertaken "willfully" and "for purposes of commercial advantage or private financial gain."[13] GenAI companies are engaging in ingestion of pirated materials with knowledge, constructive or actual, that the works on which they are building their LLM models are illicit sources. It is inarguable that they are acting for the purpose of commercial advantage. Therefore, as both elements are met, a strong argument can be made that their activities rise to the level of criminal copyright infringement.

Yet even if GenAI companies are not subject to criminal copyright infringement actions, the fact that they are clearly knowing, intentional, willful, and bad faith actors that resort to training on pirated materials should give a strong ground for denying them the ability to claim that their training is defensible under the fair use doctrine.

Congress Can Act by Ensuring That Generative AI Companies Adhere to Licensing Practices That Are Well-Established Practices in Copyright Law and Policy

By ingesting materials drawn from pirated repositories, GenAI companies are doing an end-run around copyright licensing, which is well-established under copyright law as the appropriate means of facilitating lawful access to, and use of, copyright owners' works.

For commercial markets in creative works to function fairly, sustainably, and optimally, licensing practices must be followed. GenAI companies must be required to follow these practices, as is required of all other participants in the creative markets. Courts have not yet offered a clear way to ensure that GenAI companies adhere to proper licensing practices; nor have they shown hold GenAI may be held accountable when they deviate from such well-settled practices. The time is ripe for Congressional action.

This is an area that urgently calls for guardrails and oversight. Congress can step in by requiring GenAI companies to limit their LLM training to legally-obtained and properly-licensed works. Some disclosure of training materials on the part of GenAI companies would allow oversight and, where necessary, course correction. These reasonable measures would simply bring GenAI

[12] United States v. Gordon, 37 F.4th 767 (1st Cir. 2022).
[13] https://www.justice.gov/archives/jm/criminal-resource-manual-1847-criminal-copyright-infringement-17-usc-506a-and-18-usc-2319

companies into line with standard and established licensing practices that exist in every commercial sector.

The time is ripe for Congress to make its voice heard. By requiring GenAI to follow fair and honest practices that are consistent with bedrock copyright laws and policies – including training its LLMs on materials acquired fairly and honestly, and engaging in well-established licensing practices – Congress can simultaneously foster innovative AI, support productive creators, and expand works available to the public. This can and should be a win-win for stakeholders in the technology industries, creative sectors, and the general public.

Senate Judiciary Subcommittee on Crime and Counterterrorism
Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI
July 16, 2025
Questions for the Record
Senator Amy Klobuchar

For Mr. Baldacci:

According to the Pew Research Center, newspaper advertising revenue has plummeted from $37 billion to $9 billion in recent years. Much of this decline is the result of online platforms finding ways to capture that ad revenue.

- In your written testimony, you note that the median income for authors has experienced a similarly dramatic drop, falling 42% in a decade. How does generative AI's use of author content further threaten the livelihood of authors?

The unlicensed use of books and journalism to train generative AI is truly an existential threat to the future of the already precarious writing profession; and that is because it attacks our potential earnings and the incentives to write in several ways at once: 1) it allows users, with just a little prompting, to quickly generate copycat books, books in an author's style and other infringing works that serve as market substitutes; 2) floods the market with cheaply priced AI-generated books, devaluing all books; 3) robs authors of licensing income both for the training and downstream uses; and 4) promotes piracy.

With most authors still striving to make up the losses of the last decade plus—authors' mean writing income in 2022 was only $20,000 (and only half of that from books) according to the Authors Guild's most recent author earnings survey—it does not take much more income loss before many can no longer afford to write. That in turn will impact the breadth of books that are written and published and prevent many great stories, histories, and ideas from seeing the light of day. It also will impact publishers' ability to invest in a broad assortment of new books, forcing them to focus on proven bestsellers, such as celebrity books and books by already established brand name writers, making it even harder for talented, up-and-coming authors writers to be discovered and earn any meaningful income. Those would-be authors will have less incentive to educate and train themselves and become writers, which will impoverish our nation's literature—directly upending the goals that copyright has existed to advance since the framing of the Constitution. What we're witnessing is theft of authors' works by some of the largest companies in the world to develop technologies that will grow their already enormous

profits. Unless guardrails are placed and these are technologies regulated, it will further accelerate the ongoing transfer of wealth from middle-class creators to the tech industry, and to the detriment of us all.

1. Outputs that Serve as Substitutes for our Books

Once trained on a book, an AI model can produce outputs that directly infringe a work, like summaries, excerpts, and derivative works such as video and audio versions of their work, as well as outputs that emulate a particular author's body of work. While AI companies have technologies that allow them to prevent certain types of outputs, they will not do so unless there are legal consequences. I understand that none of the major LLMs today use technologies to prevent requested outputs "in the style of" an author or that emulate an author. As I said in my testimony, ChatGPT was able to write a plot summary of a "David Baldacci novel" that included "my plot lines and twists, character names, narratives, and every other element from my work, as well as my writing style ripped from my copyrightable expression." If "fair use" is allowed to trounce authors' copyrights, AI companies will have no incentive to prevent output uses that unfairly compete with human authored books. This is why licensing is so important. It not only protects copyright, but gives authors and publishers important controls on downstream and output uses of the work.

As also I noted in my testimony, it is now common to see an AI-generated "summary" or other knockoff of an anticipated bestseller appear on Amazon on the day of or even before the release of the real book. These are clearly intended to confuse consumers and unfairly capitalize on the success of human authors by diverting sales away from the books into which they have invested years of their lives.

This means that when trained on authors' works without permission, not only do authors not get paid for that use of their work, but they also lose sales of their books to AI-generated books that infringe or otherwise serve as direct substitutes.

2. Flooding the Market

The inevitable flood of AI-generated copycat books will dilute the market for all human-authored books, and I understand that we are already seeing evidence of this dilution in the markets for certain genres, like romance and other genre fiction. That means more and more aspiring writers competing for fewer and fewer writing jobs.[1] As a result, we will have far fewer talented people who can devote themselves to writing as a profession

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

and, inevitably, fewer great books will be the unfortunate result. This will irrevocably diminish our culture and society. After all, as the Supreme Court has long recognized, copyright exists not just to "secure a fair return for an author's creative labor," but ultimately "to stimulate artistic creativity for the general public good."[2]

As a federal court put it just a few weeks ago:

> Generative AI has the potential to flood the market with endless amounts of images, songs, articles, books, and more. People can prompt generative AI models to produce these outputs using a tiny fraction of the time and creativity that would otherwise be required. So by training generative AI models with copyrighted works, companies are creating something that often will dramatically undermine the market for those works, and thus dramatically undermine the incentive for human beings to create things the old-fashioned way.[3]

3. *Loss of Licensing Income*

When AI companies simply help themselves to books for AI training without permission or compensation, it deprives authors of a valuable and rapidly growing licensing market. In the recent past, when new technologies come into existence, such as e-books, it has led to a decline in authors' book earnings, making it imperative to find other sources of income. Licensing markets are a critically important piece of this. A well-functioning copyright system would allow authors to license their books to AI companies for training and thereby obtain a valuable income stream to mitigate losses from market dilution and copycat books.

Moreover, it is only fair that authors should share in the economic rewards of generative AI, since those systems, which have generated billions for tech companies, depend on our works. As the evidence in the current class action lawsuits against AI companies for the unlicensed use of books, the large AI companies all risked pirating millions of books to train their AI precisely because they understood just how essential books are to the quality of their LLMs. In other words, they are directly profiting off our talents without paying us anything. This result is in direct contravention of U.S. copyright laws, both their letter and spirit.

---

[2] *Twentieth Cent. Music Corp. v. Aiken*, 422 U.S. 151, 155 (1975) (internal quotation marks omitted).
[3] Order Denying the Plaintiffs' Motion for Partial Summary Judgment and Granting Meta's Cross Motion for Partial Summary Judgment, at 1-2, *Kadrey v. Meta*, No. 23-cv-03417, doc. no. 598 (N.D. Cal. June 25, 2025).

Just as these companies are expected to pay for the electricity and infrastructure needed to run their data centers, so too should they have to pay for the vital creative works that are the *sine qua non* of their products.

Indeed, without our books they cannot build their AI platforms. Simply because books have become valuable to the AI community in a way that was not foreseen previously does not give them the right to steal our work. Oil was not that valuable until the invention of the combustion engine. When that happened, no one thought it was okay for the oil companies to steal the very product they needed to build their businesses. The situation is not entirely analogous because, unlike oil, creative works are particularly vulnerable to cheap copying and piracy. That is why the founders, in Article I of the Constitution, gave additional protections to copyright holders.

Yet if AI companies can simply take all the books they need for free, these essential, already-developing licensing markets will evaporate, leaving authors in the grossly unfair position of losing work to AI systems that their own books were used to create, without a penny of compensation.

4. Incentives to Pirate Books

Last, allowing AI companies to continue copying books *en masse* from pirate websites will only incentivize more piracy—already a significant drain on authors' livelihoods. Pirate sites have a ready user base in the richest companies in the world with an insatiable demand for books. It doesn't take an economist to know that, if these companies are free to get their books from such sites, new pirates will emerge and find ways to get their own piece of the action.[4] As Professor Smith explained in his testimony, we know that digital piracy significantly harms authors' incomes in multiple ways.[5] Incentivizing piracy will only make the problem worse, and vindicate patently illegal uses of copyrighted works

****

Together, the harms caused by AI companies' uncontrolled and uncompensated use of our books will add up to enormous losses, devaluing human writing; and will directly

---

[4] *See* Written Testimony of Dr. Michael D. Smith, Carnegie Mellon University at 4 (July 16, 2025) ("Indeed, this the unlicensed use of pirated content could create a new illicit licensing business model for pirate networks: adding new stolen content to their collections, knowing that AI developers will want access to them."), https://www.judiciary.senate.gov/imo/media/doc/64bc45b6-9e04-22e4-34c1-12d0efad69ef/2025-07-16%20-%20Testimony%20-%20Smith.pdf.

[5] *See id.*

impact the incentives to do the hard work it takes to become a good writer and to be able to keep writing.

These harms will also have ripple effects throughout other sectors of the writing profession. The vast majority of authors today rely on other types of writing—freelance journalism, advertising, website content, etc. to make up for the losses in book income. Fewer book publishing deals will mean that more authors will need even more supplemental income while working on their personal writing projects. But AI is already replacing many of those jobs as well, resulting in writers being attacked on multiple flanks. And that means that even our most talented and highly trained writers will have to leave the writing profession altogether.

This is clearly not what the country's founders intended. The AI community is pirating our work, breaking the laws of copyright and now asks to be excused from these abuses, only after their crimes were discovered, by arguing their technology is so transformational that the very laws intended to stop such pillaging should not apply.

One of the first lessons I learned in law school was that the slippery slope is indeed slippery. Thus, if the AI community succeeds in their argument, there is no more copyright protection for anyone. The exception will have swallowed the rule, as everyone in the future will argue that their theft of our work is also transformational. And while Professor Lee at the July 16 hearing argued that we should let the court cases play out, the reality is that the vast majority of authors, or even groups of authors, do not have the time or financial wherewithal to take on the largest corporations in the world and their armies of lawyers in a protracted court battle. Being a member of a class action lawsuit currently, I can attest to how disruptive, costly and time consuming such litigation is. That is why legislation is desperately needed, to stop this juggernaut of copyright thievery before the damage becomes irreversible. Unfortunately, we are already perilously close to that point of no return.

Written Response to Questions from the U.S. Senate Committee on the Judiciary Subcommittee on Crime and Counterterrorism

Hearing on Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training

Aug. 5, 2025

Edward Lee
Professor of Law
Santa Clara University School of Law

Dear Chair Hawley, Ranking Member Durbin, and Committee Members:

I received the following questions from Senator Klobuchar (in italics) and offer my responses below.

AI generated news summaries created from real-time scraping of journalistic sources—sometimes circumventing paywalls and violating terms of service—are not highly transformative and significantly devalue the market for the reporting necessary to make the news. These circumstances may indicate that real-time scraping of this nature may not fall under the fair use exception to copyright infringement.

This important question is raised in ongoing litigation, including the New York Times' and other news media's copyright lawsuits against OpenAI and Microsoft in the Multi-District Litigation and against other AI companies in other lawsuits.[1] It relates to the deployment of retrieval augmented generation (RAG) with AI generators to enable them to incorporate real-time information from the Internet (that was not contained in the datasets used to train the AI models).[2]

[1] *Do you believe that AI-generated news summaries based on copyrighted news reports infringes on copyrighted news content?*

As with most questions of infringement and fair use, I believe the answer will be fact specific. Consequently, I believe it is premature and unwarranted to conclude that AI-generated news summaries are "not highly transformative" or that they "significantly devalue the market for the reporting" without consideration of the evidence both sides will present in the ongoing litigation. For example, different AI generators or chatbots may differ in not only their outputs, but how those outputs are presented to the public, including how links to sources are displayed.[3] A new technology that significantly improves people's ability to conduct research, find relevant facts,

---

[1] *See, e.g.*, Second Am. Compl. ¶¶ 108-23, 186, New York Times Co. v. Microsoft Corp., No. 1:23-c-v-11195-SHS (May 28, 2025); Second Am. Compl. ¶¶ 73-79, 105-09, Dow Jones & Co. v. Perplexity AI, Inc., No. 1:24-cv-078984-KPF (Jan. 28, 2025). The full list of all copyright lawsuits is provided at Master List of Lawsuits v. AI, CHATGPT IS EATING THE WORLD (updated (Jun. 30, 2025).

[2] Kim Martineau, *What is retrieval-augmented generation?*, IBM (Aug. 22, 2023).

[3] For Google's AI mode, see Eugene Levin, *How Google's AI Mode Compares to Traditional Search and Other LLMs [AI Mode Study]*, SEMRUSH (Jun. 24, 2025).

and gain greater access to information serves the Copyright Clause's overriding goal of advancing knowledge in the United States.[4] And an AI generator that is multi-purpose and multi-functional—designed not simply for news summaries—will likely implicate other important considerations for promoting progress in the United States, such as making more accessible new creative tools to a larger segment of the population, including people with disabilities.[5]

With that caveat in mind, I believe the starting point is that facts themselves are not copyrightable. Original expression is. As the Supreme Court explained in the seminal case *Feist*: "The most fundamental axiom of copyright law is that '[n]o author may copyright his ideas or the facts he narrates.'"[6] Facts, including the "news of the day," "are part of the public domain available to every person."[7] This fact-expression dichotomy, along with the idea-expression dichotomy, serves important First Amendment values in our democracy, enabling widespread dissemination of facts and ideas.[8] As Judge Miner explained, "the freedom of access to facts and ideas is the history of democracy."[9] Indeed, as the Supreme Court admonished, the "[First] Amendment rests on the assumption that the *widest possible dissemination of information* from diverse and antagonistic sources is essential to the welfare of the public…."[10] Copyright promotes this First Amendment goal by leaving all facts in the public domain.[11]

Applying the fact-expression dichotomy, copyright law's most fundamental axiom, we must distinguish between (1) copying merely facts, which is not infringement, and (2) copying original expression, which is infringing if substantially similar and not a fair use. Thus, a news summary may be an infringing abridgment of a prior news article if the summary copied and included the *original expression* from the prior article in the summary.[12] However, if the news

---

[4] The AI company's development *of the model* serves a highly t*ransformative purpose, including in* enhancing people's ability to find relevant information and facts potentially more effectively. *See generally* Tim Keary, *Survey: 83% of users prefer AI search over 'traditional' Googling*, INNOVATING WITH AI (Jul. 1, 2025) (poll of IWAI's audience found more than 83% found AI search more efficient for getting answers to questions than traditional search); Golan v. Holder, 565 U.S. 302, 324 (2012) (interpreting the Copyright Clause's reference to the "progress of science" as "refer[ring] broadly to 'the creation and spread of knowledge and learning.'").

[5] *See* Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOU. L. REV. (forthcoming 2025) (manuscript at pp. 190-191) (AI tools offer greater accessibility to creative pursuits for people with disabilities). There is an important distinction between an AI company's development of an AI model and a user's use of an AI generator deploying the model. An AI company's use of copyrighted works to develop an AI model may be highly transformative. See Bartz v. Anthropic PBC, -- F. Supp. 3d --, 2025 WL 1741691, at *7 (N.D. Cal. Jun. 23, 2025); Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *9 (N.D. Cal. June 25, 2025). By contrast, a user's use of an AI generator model might produce a short summary of news—let's say, Congress's recent passage of the GENIUS Act—without copying any copyrighted expression from other sources. Even though the AI-generated article might be simple and merely recount facts (without original expression) from other sources, the article is non-infringing and needs no fair use defense to escape liability. And that simple article a user created using AI would not vitiate the highly transformative purpose of the AI company in developing and training the new model. AI models escaped researchers' successful development for decades. *See* Lee, supra, 63 HOU. L. REV. *(manuscript at pp. 152-55)*

[6] Feist Publns., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 344-45 (quoting Harper & Rose, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 556 (1985)).

[7] *Id.* at 348 (internal citation omitted).

[8] *See* Mary Sarah Bilder, *The Shrinking Back: The Law of Biography*, 43 STAN. L. REV. 299, 313-17 (1991).

[9] Roger J. Miner, *Exploiting Stolen Text: Fair Use or Foul Play?*, 37 J. COPYRIGHT SOC'Y 1, 10 (1989).

[10] Associated Press v. U.S., 326 U.S. 1, 20 (1945) (emphasis added).

[11] *See* Eldred v. Ashcroft, 537 U.S. 186, 219 (2003) ("As we said in *Harper & Row*, this 'idea/expression dichotomy strike[s] a definitional balance between the First Amendment and the Copyright Act *by permitting free communication of facts* while still protecting an author's expression.'") (quoting Harper & Row v. Nation Enters., 471 U.S. 539, 556 (2003) (emphasis added).

[12] *Cf., e.g.*, Barclays Capital Inc. v. Theflyonthewall.com, 700 F. Supp. 2d 310, 328 (S.D.N.Y. 2010) (defendant "no longer disputes, however, that it infringed the copyrights in these seventeen reports" summarizing plaintiffs' financial news content), *rev'd in part on other grounds*, 650 F.3d 876, 880 (2d Cir. 2011) ("Although *the extent to which the Firms'* success on the

summary did not copy any copyrightable expression, but *simply copied facts*, the news summary would not be infringing. Facts are in the public domain—and are free for all to use.[13] Moreover, AI models that are trained to identify merely the unprotected facts from sources without republishing their protected expression has a transformative purpose in using copies of articles to be able to identify the unprotected elements of the works so people can find relevant information, thereby serving the First Amendment interest in the *widest possible dissemination of information*.[14]

That a news summary is generated by AI does not change this analysis. Imagine that a second newspaper wrote a news summary based on an article first reported by the *Washington Post*. If the second newspaper merely copied facts reported by the *Post*, there is no copyright infringement. (Journalistic norms would typically require attribution to the source.) The answer would be the same if the second newspaper article were instead an AI-generated summary.

The Second Circuit's holding in *Hoehling v. Universal City Studios, Inc.* demonstrates this fundamental principle of copyright law. The court held that Michael MacDonald Mooney's and Universal City Studio's unauthorized copying of the "sabotage" interpretation of the Hindenburg's demise offered by A.A. Hoehling's book did not infringe Hoehling's copyright.[15] "Such an historical interpretation, whether or not it originated with Mr. Hoehling, is not protected by his copyright and can be freely used by subsequent authors," the court concluded.[16] "The rationale for this doctrine is that the cause of knowledge is best served when history is the common property of all, and each generation remains free to draw upon the discoveries and insights of the past."[17] It is of no moment that *Hoehling* involved historical facts while news articles typically involve recent facts at the time of publication. As the Supreme Court admonished in *Feist*, "The same is true of all facts—scientific, historical, biographical, and news of the day. '[T]hey may not be copyrighted, and are part of the public domain available to every person.'"[18]

Indeed, this fundamental principle was recognized in the "hot news" case, *INS v. AP*, in which the Court stated:

> [T]he news element—the information respecting current events contained in the literary production—is not the creation of the writer, but is a report of matters that ordinarily are *publici juris*; it is the history of the day. It is not to be supposed that the framers of the Constitution, when they empowered Congress "to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." (Const. Art. I, § 8, par. 8), intended to confer

---

copyright claims has alleviated their overall concerns is not clear, their victory on these claims is secure: Fly has not challenged the resulting injunction on appeal.")

[13] *See* Zalewski v. Cicero Builder Dev., Inc., 754 F.3d 95, 102 (2d Cir. 2014) ("Everything else in the work, the history it describes, the facts it mentions, and the ideas it embraces, are in the public domain free for others to draw upon. It is the peculiar expressions of that history, those facts, and those ideas that belong exclusively to their author.").

[14] *Cf.* Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 600 (9th Cir. 2000); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1527-28 (9th Cir. 1992).

[15] Hoehling v. Universal City Studios, Inc., 618 F.2d 972, 974-77, 978-79 (2d Cir. 1980).

[16] *Id.* at 979.

[17] *Id.* at 974.

[18] Feist Publns., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 348 (internal citation omitted).

> upon one who might happen to be the first to report a historic event the exclusive right for any period to spread the knowledge of it.[19]

My conclusion is further supported by the long line of cases recognizing fair use to create search engines (summarized in my written testimony in Appendices B and C), technologies that help people find relevant information online.[20] When an AI generator helps people in the United States find facts and information, that furthers the goal of advancing knowledge under the Copyright Clause.[21] As long as the summary or output of an AI generator does not copy original expression from online news sources, but copies merely facts, the dissemination of such facts does not produce a cognizable harm under Factor 4 of fair use in "*the protected aspect*" of the underlying work, to borrow Judge Leval's apt analysis in *Authors Guild v. Google*, another important technology fair use case.[22]

But, as I explained in my written testimony, the training of an AI model that routinely produces infringing outputs—such as infringing abridgments or summaries of works—due to inadequate guardrails will not likely be a fair use.[23] Under Factor 3 of fair use, it uses more of the works than is reasonably necessary for the transformative purpose it was intended.

There is also a very narrow claim of state law misappropriation of "hot news" that is not preempted by the Copyright Act. But, under the Second Circuit's five-factor test, it does not apply if the defendant did not publish the hot news "as its own" reporting but, instead gave attribution to the original source.[24] Thus, if an AI-generated article provided links to the sources of any hot news, it would not constitute misappropriation.

Whatever the outcome of the ongoing copyright litigation brought by news media against AI companies, it is also important to bear in mind copyright law does not preclude voluntary measures undertaken by relevant parties. For example, while prevailing in its fair use defense with respect to Google caching search, image search, and Google Books,[25] Google also established a partnership program, with paid licensing to news publishers, to feature their news in a Google News Showcase.[26] Contrary to common fallacy, fair use and licensing are not mutually exclusive.[27]

---

[19] INS v. AP, 248 U.S. 215, 234 (1918).
[20] *See* Edward Lee, Testimony before the U.S. Senate Committee on the Judiciary Subcommittee on Crime and Counterterrorism 14-15 (Jul. 16, 2025) (Appendices B and C).
[21] U.S. CONST. art. I, § 8, cl. 8.
[22] Authors Guild v. Google, Inc., 804 F.3d 202, 224 (2d Cir. 2015) (emphasis in original).
[23] *See* Lee, *supra* note 20, at 2, 4-5.
[24] *See* Barclays Capital Inc. v. Theflyonthewall.com, 650 F.3d 876, 903 (2d Cir. 2011) (explaining NBA v. Motorola, Inc. 105 F.3d 841, 898 (2d Cir. 1997) and INS v. AP, 248 U.S. 215, 239 (1918)).
[25] *See* Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (Google Book search was fair use); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) (Google image search was fair use); Field v. Google, Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006) (Google search of cached copy of website was fair use).
[26] *See, e.g.*, Sundar Pichai, Our $1 billion investment in partnerships with news publishers, GOOGLE (Oct. 1, 2020).
[27] *See, e.g.*, Sega v. Accolade, SEGA RETRO, https://segaretro.org/Sega_v._Accolade ("The two companies reached an out of court settlement which allowed Accolade to continue building their own Mega Drive cartridges, *but as an official licensee.");* Campbell v. Acuff-Rose Music, Inc., WIKIPEDIA *("*On remand, the parties settled the case out of court. According to press reports, under terms of the settlement, Acuff-Rose dismissed its lawsuit, and 2 Live Crew agreed to license the sale of its parody of the song.*").*

[2] *Do you believe that circumventing paywalls and ignoring terms of service to secure content for AI models shares similarities with downloading pirated books?*

Before discussing fair use, it is important to recognize that circumventing paywalls and ignoring terms of service are both addressed by other laws more directly tailored to such conduct. For example, circumventing a paywall to copyrighted content may violate the DMCA anti-circumvention provision.[28] (Relatedly, the Librarian of Congress has recognized a limited exception under its Section 1201 rulemaking authority for text data mining for scholarly research and teaching.[29]) Circumventing a paywall to a website may also violate the Computer Fraud and Abuse Act.[30] Finally, violating terms of service can raise a breach of contract claim.[31] Thus, regardless of how courts weigh the fair use analysis, other laws might more directly address the issue of scraping of online content in contravention of a paywall or terms of use—and create liability for such conduct.

As to how courts should weigh such conduct under fair use, my analysis is the same as my recommendation elaborated in my testimony with respect to use of pirated books.[32] I agree with Judge Chhabria's flexible approach in *Kadrey v. Meta*, in which he ruled that Meta's downloading of copies from shadow libraries was for the highly transformative purpose of training its AI model, but that such use could weigh against fair use if the plaintiffs establish market harm from such downloading.[33] But an unlawfully acquired or possessed copy should not be treated as a per se disqualification of a defendant's ability to raise a defense of fair use.

This fact-specific approach to the fair use analysis is consistent with the text of the fair use provision in the Copyright Act, which does not include any per se requirement for a "lawfully made copy" or a "lawfully possessed copy" as it does for other copyright exceptions.[34] Under a well-established canon of construction, the fair use provision should not be read to impose a limitation Congress expressly included in other copyright exceptions (such as in Section 109), but left out of the fair use provision (Section 107).[35] As Chief Justice Roberts explained for a unanimous Court in an analogous situation involving a notice exception of the Tax Code that

---

[28] *See* 17 U.S.C. § 1201(a)(1); *see also* Theresa M. Troupson, Note, *Yes, It's Illegal to Cheat a Paywall: Access Rights and the DMCA's Anticircumvention Provision*, 90 N.Y.U. L. REV. 325, 350-52 (2015).
[29] *See* 37 CFR 201.20; *Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies*, FED. REG. (Oct. 28, 2024).
[30] *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1198 (9th Cir. 2022) ("*Van Buren*[, 593 U.S. 374 (2021)] stated that the CFAA's password-trafficking provision, section 1030(a)(6), which also uses the word 'authorization,' 'contemplates a 'specific type of authorization—that is, authentication,' which turns on whether a user's credentials allow him to proceed past a computer's access gate, rather than on other, scope-based restrictions."); 18 U.S.C. § 1030(a)(2)(C) ("[w]hoever ... intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains ... information from any protected computer ... shall be punished" by fine or imprisonment.").
[31] Breach of contract is the first claim in Reddit's *suit against Anthropic* for alleged unauthorized scraping in violation of the terms of service. *See* Complaint, Reddit, Inc. v. Anthropic, PBC, No. CGC-25-62582 (Jun. 4, 2025).
[32] *See* Lee, *supra* note 20, at 5-8.
[33] *See* Kadrey v. Meta Platforms, Inc., -- F. Supp. 3d --, 2025 WL 1752484, at *12, *21 (N.D. Cal. June 25, 2025). I disagree with Judge Chhabria's endorsement of a new theory of market dilution in dicta, however. *See* Lee, *supra* note 20, at 9-12.
[34] Compare 17 U.S.C. § 109(a) ("the owner of a particular copy … lawfully made under this title") (emphasis added) with id. § 107 ("fair use of a copyrighted work"); Kirtsaeng v. John Wiley & Sons, Inc., 568 U.S. 519, 537 (2013) (discussing "lawfully made" copy requirement in §§ 109(c) (exception to public display), 109(e) (exception for video games in coin-operated equipment), and 110(1) (in-classroom teaching exception to public display and performance but not if copy "not lawfully made"); see also 17 U.S.C. § 108(c)(2) ("lawful possession of such copy *" by library or archives*) (emphasis added).
[35] Sebelius v. Closer, 569 U.S. 369, 378 (2013) ("We have long held that '[w]here Congress includes particular language in one section of a statute but omits it in another section of the same Act, it is generally presumed that Congress acts intentionally and purposely in the disparate inclusion or exclusion.'") (quoting Bates v. United States, 522 U.S. 23, 29-30 (1997)).

lacked a requirement expressly contained in a following section, "Had Congress wanted to include a legal interest requirement, it certainly knew how to do so. The very next provision—also enacted as part of the Tax Reform Act of 1976—requires the IRS to" follow such a requirement.[36] This same principle applies with equal force here to the Copyright Act of 1976. Section 109(a) imposes a requirement of a "lawfully made copy," but Section 107 does not.

The fact-specific approach to a defendant's initial acquisition of a copy that was unlawfully made is also consistent with the Supreme Court's repeated admonition that fair use is fact-specific and has no bright-line rules.[37] The Supreme Court did not treat as dispositive the *purloined* nature of a manuscript in *Harper & Row*, and, in *Google v. Oracle*, the Court rejected the argument that courts should consider the "bad faith" of the defendant under fair use, preferring instead the view of Judge Leval's influential article recognizing that "[c]opyright is not a privilege reserved for the well-behaved."[38]

This is not to suggest that defendants have a green light to do whatever they want under fair use. They do not. For example, a defendant's circumventing paywalls may provide evidence of cognizable market harm under Factor 4 of fair use in some cases, especially if wide-scale. The overarching point is that courts are well-equipped to weigh all these considerations and the evidence presented by the parties on a case-by-case basis.

Finally, just as with my answer to the first question above, we must recognize that voluntary practices related to scraping of online content are already developing. Many AI companies, including OpenAI, Amazon, Google, and Microsoft,[39] have voluntarily agreed to follow the EU's General Purpose AI Code of Practice. Under this Code of Practice, companies agree to use scraping of "only lawfully accessible copyright-protected content" and "not to circumvent effective technological measures" protecting copyrighted content.[40]

Given the more than 40 copyright lawsuits pending before the courts, other laws that directly address issues related to paywall circumvention and the breach of terms of service, and the development of voluntary practices among AI companies related to scraping of online content, I believe there is no need for Congress to intervene.

---

[36] Polselli v. IRS, 598 U.S. 432, 439 (2023).
[37] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 528 (2023); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 18-19 (2021); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).
[38] Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 32-33 (2021) (quoting Pierre N. Leval, Toward a Fair Use Standard, 103 HARV. L. REV. 1105, 1126 (1990)).
[39] *See Signatories Code of Practice*, EC EUROPA. For more on the development of the Code of Practice, see *Drawing-up a General-Purpose AI Code of Practice*, EUR. COMM'N.
[40] *Code of Practice for General-Purpose AI Models*, Copyright Chapter measure 1.2, EUROPA, at https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai#ecl-inpage-Signatories-of-the-AI-Pact.

Maxwell V. Pritt
Written Answers to Questions for the Record

Senate Judiciary Subcommittee on Crime and Counterterrorism
Hearing Titled "Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training"

August 6, 2025

Questions Submitted by Senator Amy Klobuchar:

Court filings from *Kadrey v. Meta* showed that Meta spoke with multiple companies about licensing training materials, such as books and research papers, but later decided against it because it would be "unreasonably expensive" and "incredibly slow."

1. **Meta employs more than 70,000 people and earned more than $60 billion in profits just last year. Do you believe it is possible for well-resourced companies like Meta to license and pay for high-quality content to train their models?**

Yes. It is not just possible, but in fact is already happening on a large scale. Many AI developers currently license copyrighted material for AI training, including Meta itself in certain circumstances.

Some AI companies have argued that licensing copyrighted works for use with generative AI systems is impossible due to the large amount of material needed to train a model. That self-serving argument ignores the plethora of licensing solutions that already exist and continue developing to meet market demand. The fact that companies like Meta *prefer* to pirate content for free says nothing about the feasibility of paying a fair price for that content. As one federal district court recently put it: "[T]he suggestion that adverse copyright rulings would stop this technology in its tracks is ridiculous. These products are expected to generate billions, even trillions, of dollars for the companies that are developing them. If using copyrighted works to train the models is as necessary as the companies say, they will figure out a way to compensate copyright holders for it."[1]

Meta specifically is better positioned than most companies to pay prevailing market rates for licensing copyrighted content for internal and external uses with its commercial AI models. And that's true even if Meta could substantiate its claim that doing so would be expensive. Internal documents show Meta was prepared to spend hundreds of millions of dollars on licensing copyrighted content for its AI models before it resorted to just pirating that content instead. Meta at one point discussed a $200 million licensing budget, with half of that sum earmarked for licensing books.[2] Apart from data acquisition, Meta has spent astronomical sums on its AI program, including on data infrastructure and talent. Recent reporting shows that Meta pledged hundreds of *billions* of dollars to build AI data centers, invested tens of billions of dollars in deals with AI startups, and

---

[1] *Kadrey, et al. v. Meta Platfor*ms, *Inc*., No. 23-CV-03217-VC, 2025 WL 1752484, at *2 (N.D. Cal., June 25, 2025).

[2] *Kadrey, et al. v. Meta Platforms, Inc.,* No. 23-CV-03417-VC, Dkt. 574 (Pls' Mot. for Partial Summary Judgment) at 8.

offered a $250 million compensation package to a single AI researcher.[3] Meta also projects enormous profit margins for its AI products. Meta's revenue projections for its AI program through 2035 range from a "Base Case" of $460 billion, to a "GenAI Wins Case" of $1.4 *trillion*.[4] Ability-to-pay is not an issue, and the notion that licensing content is prohibitively expensive for a company like Meta is preposterous. If Meta is willing to spend hundreds of millions of dollars to recruit a few AI researchers and hundreds billions of dollars to build AI data centers, then paying hundreds of millions of dollars, or even several billion dollars, to the reporters, authors, publishers, and others in the creative community whose works Meta used to build its AI models is hardly "unreasonably" expensive.

Other large technology companies have already entered into licensing deals to use copyrighted content with their AI systems. For example, in November 2024, Microsoft contracted to license copyrighted works from HarperCollins.[5] More recently, Amazon entered into a similar deal with The New York Times.[6] These contracts demonstrate the feasibility of large-scale licensing. Meta's internal documents also show it knows licensing copyrighted content for use with its commercial AI models is a viable option. The company at one point planned to acquire as much as 20% of its Llama 4 text data corpus through licensed content.[7] However, Meta's licensing strategy remains limited because it still employs what it calls the "gap approach"—pirate as much copyrighted content as possible, and only then use licensing to fill in the gaps of content that cannot be pirated.[8]

With respect to Meta's and other companies' argument that licensing content for use with their AI models is too slow, even assuming the companies devoted adequate resources to licensing (which Meta did not), it is not surprising that respecting intellectual property rights and complying with the law could take longer than breaking it. Naturally, that does not justify the latter.[9] Certainly Meta would not argue that OpenAI could steal its trade secrets because they helped it develop AI

---

[3] *See*, e.g., Jaspreet Singh and Aditya Soni, *Meta's Zuckerberg pledges hundreds of billions for AI data centers in superintelligence push*, REUTERS (July 14, 2025); Billy Perrigo, *How Meta's $14 Billion Scale AI Investment Upended the AI Data Industry*, TIME (June 16, 2025); Mike Isaac, Eli Tan and Cade Metz, *A.I. Researchers Are Negotiating $250 Million Pay Packages. Just Like N.B.A. Stars.*, N.Y. TIMES (July 31, 2025).

[4] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 574 at 4.

[5] Hannah Miller and Dina Bass, *Microsoft Signs AI-Learning Deal With News Corp.'s HarperCollins*, BLOOMBERG (Nov. 19, 2024).

[6] Alexandra Bruell, *Amazon to Pay New York Times at Least $20 Million a Year in AI Deal,* WALL STREET JOURNAL (July 30, 2025).

[7] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 537 at 29.

[8] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 588-1 at 65.

[9] *See*, *e.g.*, *Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd.*, 454 F. Supp. 2d 966, 989 (C.D. Cal. 2006) (StreamCast also blames Plaintiffs for their difficult licensing terms, which StreamCast believes prevented it from launching a successful, legal business with licensed content. . . . Whatever its subjective intentions were about eventually securing licenses and developing revenue streams that did not depend on infringement, the business that actually materialized was one that thrived only because of the massive infringement enabled by Morpheus and OpenNap/MusicCity.")

models faster. The desire to try to keep pace with competitors cannot justify the AI industry's collective decision to "YOLO the legal risk"[10] and commit domestic online piracy at a staggering scale.

2. Are there licensing models that could fairly compensate creators without unnecessarily delaying or hampering AI innovation?

Yes. In addition to a growing number of one-to-one deals between established copyright-holding companies and generative AI developers, collective licensing is available to address issues of scalability. The U.S. Copyright Office conducted a detailed study of this question in its Report on Copyright and Artificial Intelligence titled "Generative AI Training" (the "Report"),[11] finding "available information shows that [licensing] markets exist or are 'reasonable' or 'likely to be developed[.]'"[12]

There is already high demand for corpora of copyrighted works for ingestion by AI systems, and, as discussed above, copyright holders are offering and entering into various licensing agreements. Publishers and copyright holders of scientific and research works such as Elsevier, JSTOR, the Copyright Clearance Center, and many others have either offered or entered into licensing agreements that allow for text and data mining (TDM) or other generative AI uses. Getty Images has struck several licensing deals with generative AI companies for use of portions of its catalog of stock images for training. Multiple news organizations, including NewsCorp, the Associated Press, The Atlantic, The New York Times, and the Financial Times, have reached deals with various AI developers. The list goes on and on, with new licensing deals being announced almost daily.[13]

Importantly, collective licensing is nothing new—it has proven feasible in many contexts and has readily adapted to new uses. With respect to literary works, as just one example, the Copyright Clearance Center was founded in 1978 with the aim of facilitating photocopying permissions in academic settings, and it has been undeniably successful at distributing royalties at scale.[14] Similarly, Performing Rights Organizations (PROs) collect and distribute monies for

---

[10] "YOLO" being a common slang term for "you only live once", so "why worry about the consequences?" *See Tremblay v. OpenAI*, No. 3:23-cv-0322 (N.D. Cal. March 13, 2025), Dkt. 392-8 (Pl's Proposed Second Amended Consolidated Complaint) at 15.

[11] U.S. Copyright Office Copyright and Artificial Intelligence, Part 3: Generative AI Training (May 9, 2025), at 70.

[12] *Id.* at 70.

[13] Copyright Alliance, *AI Licensing for Creative Works*, https://copyrightalliance.org/artificial-intelligence-copyright/licensing/.

[14] Mark Seeley, *Evolution of Copyright Law from Guild and Printing Monopolies to Human and Natural Rights*, https://www.copyright.com/wp-content/uploads/2021/01/CCC_CreatingSolutionsTogether_Ebook_2020.pdf, at 25.

musicians where it would otherwise be difficult or inefficient to directly license public performance permissions.[15]

While licensing for internal and external uses in connection with generative AI systems is still in its early stages, the information already available shows there is a clear path towards voluntary licensing that would allow copyright owners to control their works and earn incremental revenue for commercial exploitation of their works by the AI industry. While there isn't a one-size-fits-all solution to licensing for AI systems, there is no reason to doubt that major industry players can develop mutually beneficial solutions so that creators and rightsholders can share in the massive profits expected by the generative AI industry.[16] The feasibility of collective licensing is also demonstrated by models like Audible and Spotify Audiobooks, which already license books at scale.

As happened in the music industry in the 2000s, once online piracy is legally prohibited, market forces react naturally by developing legitimate alternatives. Shortly after Napster was enjoined, record companies made deals with Internet platforms and streaming services to distribute their music.[17] Apple's iTunes proliferated immediately in Napster's aftermath. Streaming models like Pandora and Spotify followed shortly thereafter. The lesson from the music industry is clear: once major participants in pirated markets are forced to use legitimate alternatives to obtain copyrighted content, those markets develop rapidly, including functional systems of collective licensing.[18] In light of the already growing market for licensing copyrighted books and other content

---

[15] Issues Related to Performing Rights Organizations, Comments of the Copyright Alliance, https://copyrightalliance.org/wp-content/uploads/2025/04/AS-SUBMITTED-Copyright-Alliance-Comments_NOI-PRO.pdf, at 2.

[16] *See Kadrey v. Meta*, 2025 WL 1752484, at *22 ("Meta argues that the 'public interest' would be 'badly disserved' by preventing Meta (and other AI developers) from using copyrighted text as training data without paying to do so. Meta seems to imply that such a ruling would stop the development of LLMs and other generative AI technologies in its tracks. This is nonsense. As mentioned earlier, a ruling that certain copying isn't fair use doesn't necessarily mean the copier has to stop their copying—it means that they have to get permission for it. So where copying for LLM training isn't fair use, LLM developers (including Meta) won't need to stop using copyrighted works to train their models. They will need only to pay rightsholders for licenses for that training. Presumably, where copying for AI training isn't fair use, AI developers will simply figure out a way to license the works they wish to use as training data. Meta's contention that markets for this licensing can't or won't develop is hard to believe. If books are as good for LLM training as Meta says they are, then it seems nearly certain that LLM developers would be willing to pay for licenses. (Indeed, Meta itself was willing to pay to license books—it just found licensing too logistically difficult.) Even if the value of any particular book as training data is too low to justify negotiating licensing deals book by book, LLM developers would still presumably be interested in licensing large numbers of books at once .. . . So if it isn't fair use for Meta and other LLM developers to use copyrighted books as training data without permission, they won't have to stop working on their LLMs altogether. They'll just have to pay for licenses or use books that aren't copyrighted. Either way, it may be that LLM companies move somewhat more slowly or make somewhat less money. But the suggestion that the growth of LLM technology would come to a halt (or anything close) doesn't pass the straight face test.").

[17] *See* 1 Lindey on Entertainment, Publ. & the Arts § 2:28 n. 36 (3d ed. 2024).

[18] *See* Jonathan M. Barnett, *The Big Steal: Ideology, Interest, and the Undoing of Intellectual Property* 337 (2024) ("As illustrated by the rise of licensed music and video streaming services, the performance of real-world digital content environments shows that well-functioning markets that support a robust flow of content

for use with AI systems, there is little reason to doubt that a thriving licensing market will continue to develop.

---

production are generally compelled to assemble a property-rights infrastructure—understood broadly to encompass legal, technological, and contractual devices that enable content owners to regulate and price access to some significant extent. The same argument can be made for licensed platforms in electronic books, digital images, and other creative media.").

A P P E N D I X

The following submissions are available at:

https://www.govinfo.gov/content/pkg/CHRG-119shrg61891/pdf/CHRG-119shrg61891-add1.pdf

Submitted by Chair Hawley:

Submitted by Ranking Member Durbin:

Submitted by Senator Klobuchar:

Submitted by Senator Coons: