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.