Statement of

Keith Kupferschmid
Chief Executive Officer
Copyright Alliance

before the

SENATE COMMITTEE ON THE JUDICIARY
SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM

July 16, 2025

The Copyright Alliance, on behalf of our membership, 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 Copyright Alliance is a non-profit, non-partisan public interest and educational organization that is dedicated to advocating policies that promote and preserve the value of copyright, and to protecting the rights of creators and innovators. We represent the copyright interests of over 2 million individual creators, including established authors and artists, performers and photographers, and software coders and songwriters, as well as a new generation of creators. Some of these creators are career professionals, while others are hobbyists. Some have years of experience, while others are just embarking on their burgeoning careers. Some are critically acclaimed, while others toil in relative obscurity or have limited audiences.

We also represent the copyright interests of over 15,000 organizations in the United States, across the spectrum of copyright disciplines. These include motion picture and television studios, record labels, music publishers, book and journal publishers, newspaper and magazine publishers, video game companies, software and technology companies, visual media companies, sports leagues, radio and television broadcasters, database companies, standard development organizations and many more. While each of these organizations may come to the Copyright Alliance with somewhat different experiences, views, and interests, they all fall under the Copyright Alliance umbrella for one unifying reason—their strong support for the value and importance of copyright and protecting the rights of human creators and copyright owners.

Copyright Alliance members—whether they are an individual creator or an organization, whether they are big or small, or whether they are more traditional creators/copyright owners or a new generation of creators/copyright owners—share at least two things in common: (1) they rely on copyright law to protect their efforts and investments in the creation, reproduction, distribution, and adaptation of works for the public to enjoy, and (2) they are harmed by piracy and are concerned about copyright infringement and piracy-related issues raised by generative AI.

Many Copyright Alliance members and others in the creative industries are already using artificial intelligence as a tool to assist with their content creation, just as they use other technologies, and undoubtedly will use more sophisticated versions of AI as they develop. We therefore support efforts to advance innovation. But all development and use of technology must be done legally and responsibly.

This is consistent with the position the first Trump Administration took that "[t]he AI technologies we develop must also reflect [the] fundamental American values" in "freedom, guarantees of human rights, the rule of law, stability in our institutions, rights to privacy,

respect for intellectual property, and opportunities to all to pursue their dreams."[1] Abiding by these principles will be critical to making American implementation the gold standard for AI and to protecting and expanding the United States' dominant global positions in *both* AI and the copyright industries.

The United States will not win the AI race with China if it comes at the expense of good copyright policy. Indeed, not only would undermining copyright law hurt the United States as the global leader in the creative industries, but abiding by good copyright policy will help ensure that users at home and abroad can trust and benefit from American AI technology over implementations from AI developers from other countries. That means using trusted, curated datasets that are the most accurate, providing consumers and businesses in the United States and across the globe the confidence that they can use American AI tools without fear of "hallucinations" or other flaws that can cause harm. Indeed, as we discuss in more detail below, a recent report demonstrating that chatbots already include an alarming amount of Chinese propaganda seeded by the Communist Party demonstrates the harm that comes from relying on indiscriminate scraping of the internet to train AI systems as opposed to licensing reliable content.[2]

The two copyright issues of most interest to Copyright Alliance members and the copyright community more broadly are online piracy and generative AI. These two issues may appear to be independent from one another, but recent AI copyright infringement lawsuits have shown that they are very much intertwined. Indeed, those lawsuits have demonstrated that some AI companies have leveraged the proliferation of copyrighted works in unauthorized centralized databases to their benefit, for instance, by accessing so-called "shadow libraries," like Z-Library or Lib-Gen, they copy and ingest pirated works that they use to train their AI systems

[1]The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update (emphasis added), https://trumpwhitehouse.archives.gov/ai/ai-american-values/.

[2]*See* COURTNEY MANNING, MONIQUE SHUM & JOSEY WALDEN, AMERICAN SECURITY PROJECT, EVIDENCE OF CCP CENSORSHIP, PROPAGANDA IN U.S. LLM RESPONSES (June 25, 2025) (stating that "AI-powered chatbots in the United States now regurgitate CCP propaganda in Chinese and English when prompted on certain topics, posing significant ramifications for global AI development and U.S. national security"), https://cdn.prod.website-files.com/67919c3b2972e57c613c2ea2/685b1a27a830fb5b6e7ff511_Sentinel%20Brief%20-%20Evidence%20of%20CCP%20Censorship%20in%20LLM%20Responses.pdf.

instead of licensing them from copyright owners. Therefore, the focus of this hearing, which is to address that overlap, is of significant importance to the Copyright Alliance.

Piracy Harms America's Strong Creative Economy

America's innovation and creative economies continue to be the best in the world. There is a reason for that U.S. dominance—it is U.S. copyright law. Strong and effective copyright protection is critical to and fuels the U.S. economy, culture, trade and employment. A report on the economic impact of copyright by the International Intellectual Property Alliance (IIPA) found that, in 2023, the core copyright industries contributed more than $2 trillion to the U.S. gross domestic product (GDP) (accounting for 7.66% of the U.S. economy) and employed 11.6 million workers (or 5.43% of the workforce).[3] In addition to growing at a rate more than three times that of the rest of the economy, the report notes that the core copyright industries: (1) make up an increasingly large percentage of value added to GDP; (2) create more and better paying jobs than other sectors of the U.S. economy; (3) grow faster than the rest of the U.S. economy; (4) contribute substantially to U.S. foreign sales and exports, outpacing many industry sectors; and (5) make significantly large contributions to the digital economy, which does not even encompass the full scope of the copyright industries' digital activities.[4] The IIPA report also found that copyright-reliant industries contribute over 63% to the economic value of the digital services sector, which underscores the critical importance of copyright law to America's digital economy.[5]

While the growth of the internet over the last twenty-five years has revolutionized the way that creative works are legally made available and reach their intended audience, it has also facilitated massive amounts of copyright piracy that causes tremendous harm to creators, copyright owners, and consumers. The widespread theft of copyrighted works online is a persistent and evolving problem affecting virtually all types of works and all types of copyright

[3] Robert Stoner & Jéssica Dutra, *Copyright Industries in the U.S. Economy: The 2024 Report*, INT'L INTELL. PROP. ALL. (Feb. 2025).

[4] *Id.* at 21.

[5] *Id.* at 20.

owners in the digital age, and it undermines the rights of creators, the value of copyright, and our creative economy.

Global online piracy of copyright-protected works results in billions of dollars of economic loses each year, hundreds of thousands of lost jobs, and immeasurable harm to the safety of consumers through the spread of malware, phishing scams, and identity theft.[6] A study by the Global IP Center found that digital video piracy alone deprives the U.S. economy of a minimum of $29.2 billion in reduced revenue each year.[7] This type of piracy not only causes lost revenues to the U.S. creative sectors, it also results in losses to the U.S. economy of between a quarter million and half million jobs and between $47.5 billion and $115.3 billion in reduced gross domestic product (GDP) each year.[8] Piracy also poses a threat to investments in creative works by unjustly enriching bad actors who make no investment and take no risk, at the expense of the creators.

In an article outlining the effects of online piracy, one of today's panelists, Professor Michael Smith, explains that digital piracy harms creators by reducing their ability to commercialize their creative efforts.[9] He points to a broad consensus in peer-reviewed academic literature that confirms "that online piracy does exactly what one would expect: it makes it harder for creators and rights owners to make a fair market return on their investments in content creation and dissemination."[10] In addition to the harms caused to copyright owners, the article summarizes the harms caused to society by reducing creators' economic incentives to invest in creative output. It explains that economic theory reinforces the Constitutional principle[11] that the public

[6] Impacts of Digital Video Piracy on the U.S. Economy, GLOBAL INNOVATION POLICY CENTER (June 2019), available at: https://www.uschamber.com/technology/data-privacy/impacts-of-digital-piracy-on-the-u-s-economy.

[7] *Id.*

[8] *Id.* at 14.

[9] Michael D. Smith, *What the Online Piracy Data Tells Us About Copyright Policymaking*, HUDSON INSTITUTE (April 12, 2023), https://www.hudson.org/intellectual-property/what-online-piracy-data-tells-us-about-copyright-Policymaking.

[10] *Id.*

[11] U.S. Const. art. I, § 8, cl. 8.

interest is promoted by ensuring creators can pursue their own private interests, and that reduced incentives "cause significant problems for both creators and the broader society that benefits from their talents."[12] Finally, the article cites to significant empirical evidence in the academic literature that the losses in revenues that result from online piracy has harmed consumers by reducing both the quantity and quality of creative output that would have occurred absent piracy.[13]

AI Companies' Mass Ingestion of Pirated Works from Illicit Sources to Train Their AI Systems is Extremely Harmful to the U.S. Creative Economy

AI companies obtain copyrighted works and other content they use to train their AI systems in many different ways—some legal and some illegal. For example, some companies legally use proprietary materials that they have created themselves or otherwise own.[14] Some may train on material that is in the public domain or not protected by copyright. But the most common example of legal training is through licensing deals that have been and continue to be struck between copyright owners (or their representatives) and AI developers.[15] The robust and ever-expanding licensing-for-training market spans the spectrum of creative works and generative AI models. Many copyright owners are in the unique position of being able to organize, curate, and apply metadata to high quality works and many AI companies recognize the value in clean, ethically sourced, liability free works over indiscriminately scraped masses of material. *The*

[12] Smith, *supra* note 6.

[13] *Id.*

[14] *Adobe unveils AI video generator trained on licensed content*, Benj Edwards, ARS TECHNICA (Oct. 14, 2024), https://arstechnica.com/ai/2024/10/adobe-unveils-ai-video-generator-trained-on-licensed-content/.

[15] There is already high demand for corpora of copyrighted works for ingestion by AI systems, and copyright owners are offering and entering into various licensing agreements. Publishers and copyright owners 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 organization, 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. *See* Copyright Alliance, "Generative AI Licensing Isn't Just Possible, It's Essential," *available at* https://copyrightalliance.org/generative-ai-licensing/ (citing original sources and media coverage). See https://copyrightalliance.org/artificial-intelligence-copyright/licensing/.

current licensing landscape is clear evidence of a free-market licensing system that is working and must be fostered and not disrupted by illegal activity by some bad actors.

In some cases, licenses for AI training are direct, voluntary agreements reached between an AI developer and copyright owner. There are also a growing number of voluntary collecting rights organizations, including Protege, Created by Humans, and ProRata.ai, that voluntarily license the works of many copyright owners to AI companies. Many of these organizations are small companies created within the last several years for the purpose of making voluntary licensing of copyrighted works for training much easier for AI companies and creators by allowing an AI company to voluntarily license many copyrighted works from many creators at one time, through one license agreement. The collecting rights organizations model is not a new one. It has been around for many decades and proven to be successful in many other areas of copyright licensing.

While some AI companies claim that licensing works for training is impossible due to the large amount of material they desire to train a model, that is a self-serving argument that ignores the many different voluntary licensing solutions that have already been (and continue to be) established to meet market demand. A judge recently made this point in a generative AI infringement case, explaining that:

> "[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."[16]

Just as with other technological advancements like the internet or streaming that temporarily disrupted the copyright marketplace, new and robust voluntary licensing market solutions

[16] *Kadrey v. Meta Platforms, Inc.,* 3:23-cv-03417, Dkt. 601 at 3 (N.D. Cal.).

support the development of generative AI that benefits both copyright owners and AI companies.

Unfortunately, AI companies do not always license their training materials and instead choose to cut corners by opting for quicker, less expensive, illicit routes. Indeed, using illicitly sourced material for AI training is a problem identified in numerous lawsuits brought by copyright owners against AI companies for the unauthorized mass ingestion of copyrighted works to train their generative AI models. In these lawsuits, the defendant AI companies have not denied their use of such illicit source materials. It is now common knowledge that some AI companies routinely engage in indiscriminate mass ingestion of copyrighted works for the training of their AI systems, which inevitably implicates either or both of copyright owners' reproduction rights by copying works from piratical websites and services (where copies of those works were made available without the owner's authorization) and/or violating the anticircumvention provisions in copyright law by intentionally breaching firewalls and other technical measures for the purpose of illegally downloading copyrighted works that were only intended for authorized consumer use. While some AI datasets involve more discriminating curation, many of the most popular repositories used for large language model (LLM) training contain hundreds of thousands, and in some cases millions, of copies of pirated works that have been knowingly ingested from illicit sources.[17] We want U.S. AI companies to dominate over their foreign counterparts, but engaging in that type of large scale, commercial, and harmful activity is not the way to establish American AI dominance.

There are many different illicit ways some AI developers get pirated material for training. The list below is nonexclusive and includes methods that often overlap with one another.

- *Using pirated works sourced from shadow libraries*: So-called shadow libraries are online repositories containing massive amounts of pirated works, including books, research papers, and other literary works. Many shadow libraries have repeatedly been

[17] *The Unbelievable Scale of AI's Pirated-Books Problem*, Alex Reisner, THE ATLANTIC (Mar. 20, 2025), https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/.

found to be illegal and against the public interest, and have had their domains shut down and operators arrested.[18] In fact, illicit services like Library Genesis (aka LibGen), Z-Library, Sci-Hub, and Bibliotik have already been indicted for criminal copyright infringement or listed in the Office of the U.S. Trade Representative's annual review of Notorious Markets for Counterfeiting and Piracy.[19] Despite this, many AI developers continue to seek out and use these criminal enterprises as a source to build their AI models.

- *Scraping copyrighted works from websites:* Using bots to indiscriminately scrape the internet for copyrighted works that can be used as training material inevitably involves scraping pirated works off illicit websites, whether made available through the above-mentioned shadow libraries or any other online repositories of pirated works. Many AI companies have used and continue to use massive datasets compiled by organizations like Common Crawl, which uses automated software programs to systematically crawl and copy, or "scrape," the entire internet. Additionally, some web crawlers are designed to bypass firewalls and access legitimate websites by simulating human behavior, further obstructing copyright owners' protection efforts.

- *Peer-to-peer (P2P) torrenting*: P2P torrenting is a process used to download large amounts of material while often simultaneously redistributing them (through a process known as "seeding" or "leeching") to other P2P users. In some ongoing lawsuits, AI companies have been shown to have knowingly availed themselves of P2P networks that are notorious hotbeds of digital piracy and copied an untold number of pirated copies of copyrighted works using torrent services. In at least some cases, it is alleged

[18] Brief for the International Association of Scientific, Technical, and Medical Publishers as Amicus Curiae Supporting Plaintiffs at 2*, Kadrey v. Meta Platforms, Inc.*, 3:23-cv-03417, (N.D. Cal.).

[19] *See generally* OFFICE OF THE U.S. TRADE REPRESENTATIVE, *2024 Review of Notorious Markets for Counterfeiting and Piracy*, (Jan. 8, 2025) https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Counterfeiting%20and%20Piracy%20(final).pdf; *see also* OFFICE OF THE U.S. TRADE REPRESENTATIVE, *2023 Review of Notorious Markets for Counterfeiting and Piracy*, at 27 & 30 (Jan. 30, 2024), https://ustr.gov/sites/default/files/2023_Review_of_Notorious_Markets_for_Counterfeiting_and_Piracy_Notorious_Markets_List_final.pdf.

that an AI company has not only downloaded but uploaded the works through seeding or leeching. Such redistribution of massive amounts of copyrighted works is incredibly harmful and possibly criminal.

The harm online piracy already causes to copyright owners is no doubt exacerbated when pirated works are reproduced, and in some instances redistributed, by AI companies that employ some or all of the mechanisms above to amass training material. Specifically, the harm caused by illicit shadow libraries, which is already massive, is compounded by AI companies that access the pirated works through torrenting because it enables further downstream dissemination. It also contributes to the distribution of illegitimate copies of scientific and medical works that contain misinformation or have been retracted or updated.

Mass Ingestion of Pirated Works for AI Training Should Weigh Heavily Against Fair Use

In the most recent report in its ongoing study of the copyright issues raised by generative AI, the U.S. Copyright Office addresses the mass ingestion of copyrighted works for training by AI developers. Ultimately, in the Office concludes that, "the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use."[20] Judges in two generative AI infringement cases in the Northern District of California recently issued orders on fair use related to the use of pirated works for AI training.[21] Similar to the Copyright Office, both Judges expressed serious concerns over the defendants AI companies' mass ingestion of pirated works. In one of the cases, *Bartz v. Anthropic*, the Judge's order describes how Anthropic downloaded over 7 million copies of pirated books and concludes, "[p]iracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded."[22] While the Judge in the other case, *Kadrey v. Meta*, explained that "[i]n the vast majority of cases, this sort of peer-to-peer file-sharing will

[20] *Id.*

[21] *Bartz v. Anthropic PBC*, 3:24-cv-05417, Dkt. 231 (N.D. Cal.); *Kadrey v. Meta Platforms, Inc*., 3:23-cv-03417, Dkt. 598 (N.D. Cal.).

[22] *Bartz*, at 19.

constitute copyright infringement."[23]

These two court cases and many other pending AI infringement cases are distinguished from past fair use cases that are relied on by AI companies. Those past cases involved the use of one or many *legitimate* copies, and that use was found to be fair use because the copy was not otherwise available. In contrast, many of the pending AI training infringement cases involve the use of *millions of pirated* copies. Never has a court confronted the use of pirated copies on such a massive scale. Moreover, copyrighted works that AI companies wish to use for training are largely available through legitimate online sources, which is in stark contrast to past cases where works were not licensable or otherwise available. Sourcing mass amounts of pirated copyrighted works is not necessary for the development of generative AI models, it is simply the least expensive and fastest way for some AI companies to get the vast quantity of works they want.

In sum, the major distinction between these AI cases and past fair use cases is (i) the sheer scale of what is being ingested by AI companies, (ii) the fact that AI companies are using pirated copies, not legitimately made copies, and (iii) the sheer scale of competing outputs that are generated by using these pirated works. In these AI infringement cases, there are millions of pirated works being illegally copied and used to produce millions of outputs that may compete with and harm the market for the original works. The fact that AI companies intentionally use these illicit sites and services does not presumptively disqualify an AI developer from claiming fair use is antithetical to the foundations of our copyright system and the rights it guarantees.

Strong Enforcement of Copyright Laws is Integral to Our National Security

Some AI companies claim that limiting the use of copyrighted material for AI training amounts to a national security threat because it could impede AI innovation in the United States, potentially granting geopolitical competitors like China a technological advantage. Those claims could not be more wrong.

Promoting strong copyright laws and the Constitutional guarantees of the protection of human

[23] *Kadrey*, at 20.

creators is critical to maintaining American AI dominance, which in turn will diminish national security threats. American culture and arts are key tools for spreading American values and advancing democracy and freedom across the globe. The global popularity of American books, movies, music, and art provides vehicles by which we are able to project American values to the rest of the world. Tearing down American creativity leaves us and the rest of the world vulnerable to Chinese propaganda and is an obvious and dangerous step in the wrong direction.

Before suggesting an unwarranted upheaval to U.S. copyright law for the purposes of national security, AI companies should clean their own house. For example, it has been widely reported that China has been covertly using "ChatGPT to spread propaganda, manipulate social media engagement, and target journalists and politicians in a coordinated AI-powered influence campaign."[24] And ChatGPT is not alone. A report issued at the end of June by the American Security Project concluded that:

> "The Chinese Communist Party's aggressive censorship laws and disinformation campaigns have resulted in a proliferation of propaganda and censorship across the global AI data marketplace. AI-powered chatbots in the United States now regurgitate CCP propaganda in Chinese and English when prompted on certain topics, posing significant ramifications for global AI development and U.S. national security."[25]

If AI companies want to talk about national security, the conversation must start with this very real threat to our national security, instead of the contrived threat of copyright law.

Additional Final Points

It is important to understand that copyright owners' rights are implicated when their works are

[24] *OpenAI Reveals China Covertly Used ChatGPT To Spread Propaganda, Manipulate Social Media Engagement, And Target Journalists And Politicians In A Coordinated AI-Powered Influence Campaign*, Ezza Ijaz, WCCFTECH (June 6, 2025), https://wccftech.com/openai-reveals-china-covertly-used-chatgpt-to-spread-propaganda-manipulate-social-media-engagement-and-target-journalists-and-politicians-in-a-coordinated-ai-powered-influence-campaign/.

[25] *Evidence of CCP Censorship, Propaganda in U.S. LLM Responses*, Courtney Manning, Monique Shum, and Josey Walden, THE AMERICAN SECURITY PROJECT (June 25, 2025), https://ai.americansecurityproject.org/research/ccp-censorship-in-llm-responses.

reproduced to create datasets or ingested by generative AI systems, regardless of whether the AI systems generate infringing output or distribute infringing copies to end users. Some AI companies like to argue that there is no harm to copyright owners if their models do not generate infringing outputs, but that is only half the story. The mass ingestion of pirated works indisputably implicates a copyright owner's right of reproduction, which is the foremost right guaranteed by Section 106 of the Copyright Act. AI developers' choice to ingest massive amounts of pirated works to train their AI systems instead of entering into licensing agreements with copyright owners may also harm the copyright owners whose works have been ingested by negatively impact their ability to commercialize their works, recoup investments, and the incentivization to create new works.

Finally, a separate issue we would like to bring to the attention of the Subcommittee is legislation intended to block access to large-scale commercial foreign piracy sites. The Copyright Alliance strongly supports the establishment of a judicial blocking system that would allow copyright owners to petition a federal court to issue an order requiring a U.S. internet service provider (ISP) to prevent foreign-based websites and services that have a primary purpose of providing access to infringing material from providing pirated content to U.S. consumers. Over 50 nations across the globe have established similar systems that have been proven to effectively curb piracy and promote legitimate services without harming free expression or breaking the internet.[26] Efforts to establish a judicial blocking system have bipartisan and bicameral support, and we hope to soon see the introduction of legislation that will help copyright owners combat harmful foreign-based piracy.

[26] *Blocking Access to Foreign Pirate Sites: A Long-Overdue Task for Congress*, Rodrigo Balbontin, INFORMATION TECHNOLOGY & INNOVATION FOUNDATION (June 9, 2025), https://itif.org/publications/2025/06/09/blocking-access-to-foreign-pirate-sites-a-long-overdue-task-for-congress/.