UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 1: Digital Replicas

A REPORT OF THE REGISTER OF COPYRIGHTS

JULY 2024

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 1: Digital Replicas

A REPORT OF THE REGISTER OF COPYRIGHTS

JULY 2024

ABOUT THIS REPORT

This Report by the U.S. Copyright Office addresses the legal and policy issues related to artificial intelligence ("AI") and copyright, as outlined in the Office's August 2023 Notice of Inquiry ("NOI").

The Report will be published in several Parts, each one addressing a different topic. The first Part addresses the topic of digital replicas—the use of digital technology to realistically replicate an individual's voice or appearance. Subsequent Parts will turn to other issues raised in the NOI, including the copyrightability of works created using generative AI, training of AI models on copyrighted works, licensing considerations, and allocation of any potential liability. To learn more, visit www.copyright.gov/AI.

ABOUT THE U.S. COPYRIGHT OFFICE

The U.S. Copyright Office is the federal agency charged by statute with the administration of U.S. copyright law. The Register of Copyrights advises Congress, provides information and assistance to courts and executive branch agencies, and conducts studies on national and international issues relating to copyright, other matters arising under Title 17, and related matters. The Copyright Office is housed in the Library of Congress. Its mission is "to promotes creativity and free expression by administering the nation's copyright laws and by providing impartial, expert advice on copyright law and policy for the benefit of all." For more information, visit www.copyright.gov.

FOREWORD FROM THE REGISTER OF COPYRIGHTS

The recent emergence of sophisticated generative artificial intelligence ("AI") models available for use by consumers constitutes a major leap forward in technology. It presents both exciting opportunities and complex challenges for society as a whole, which have captured the attention of policymakers around the world, as well as the press and the public.

One of the areas affected is intellectual property. Copyright issues in particular have risen to the forefront, due to their visibility, immediacy, and relevance to the average person. By the fall of 2022, millions of Americans were utilizing generative AI systems and services to produce an astonishing array of expressive material, including visual art, text, and music. Almost weekly, tremendous strides have been announced in the technology's capabilities. Artists have harnessed the power of AI to find new ways to express themselves and new ways of connecting with audiences. At the same time, AI-generated deepfakes have proliferated online, from celebrities' images endorsing products to politicians' likenesses seeking to affect voter behavior. Over the past year or so, the resulting debates have intensified, with enthusiasm about the promise of extraordinary technical potential tempered by concern about the impact on individuals' livelihoods and reputations.

AI raises fundamental questions for copyright law and policy, which many see as existential. To what extent will AI-generated content replace human authorship? How does human creativity differ in nature from what AI systems can generate, now or in the future? What does this mean for the incentive-based foundation of the U.S. copyright system? In what ways can the technology serve as a valuable tool to amplify human creativity and ultimately promote science and the arts? How do we respect and reward human creators without impeding technological progress?

For copyright, this is the latest chapter in a symbiotic relationship with technology. Throughout history, technological innovation has shaped the evolution of copyright law and policy, with new forms of expression, such as photography, motion pictures, and computer programs; new methods of copying, such as photocopiers and video-cassette recorders; and new means of distribution, such as radio, television, and the internet. In recent decades, the pace of change has sharply accelerated, and today's generative AI tools have picked it up even further. The late 20th century saw the Copyright Act amended to respond to the challenge of digital networked technology. History has shown that the copyright system is resilient and continues to evolve as needed.

In response to the importance and urgency of the copyright issues, in early 2023 the Copyright Office initiated the study that led to this Report. Our work is just one part of a broader national and global conversation. In the United States, the Biden Administration's

October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence addresses how AI technologies can be deployed safely and responsibly while taking into account concerns about fraud, bias, and transparency, as well as the impact on intellectual property rights. Other agencies are examining issues within their own areas of jurisdiction, and Congress is debating the need for legislation. Governments of other countries are similarly grappling with the potential impact of AI in all of its forms.

As with all of the Copyright Office's studies, our analysis is guided by the Constitutional goal of promoting creativity in order ultimately to benefit the public. This requires an appropriate balance, enabling technology to move forward while ensuring that human creativity continues to thrive. It is our hope that this Report will further productive discussions in Congress, the courts, and the executive branch, to help achieve that balance.

Shira Perlmutter
Register of Copyrights and Director
U.S. Copyright Office

PREFACE

In early 2023, the U.S. Copyright Office announced a broad initiative to explore the intersection of copyright and artificial intelligence.

In March of that year, the Office released a policy statement with registration guidance for works incorporating AI-generated content. Over the spring and summer, we hosted a series of online listening sessions, presented educational webinars, and engaged with numerous stakeholders to enhance our understanding of the technology and how it is used, the copyright implications, and the potential impact on businesses and individuals.

These activities culminated in an August 2023 Notice of Inquiry, formally seeking public input on the full range of copyright issues that had been raised. In response, we received more than 10,000 comments representing a broad range of perspectives, including from authors and composers, performers and artists, publishers and producers, lawyers and academics, technology companies, libraries, sports leagues, trade groups and public interest organizations, and even a class of middle school students. Comments came from all 50 states and from 67 countries. That valuable and extensive input, supplemented by additional Office research and information received from other agencies, forms the basis for the discussion and recommendations in this Report.

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence
Part 1: Digital Replicas

A REPORT OF THE REGISTER OF COPYRIGHTS

JULY 2024

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This first Part of the Copyright Office's Report on copyright and artificial intelligence ("AI")[1] addresses the topic of digital replicas. From AI-generated musical performances to robocall impersonations of political candidates to images in pornographic videos, an era of sophisticated digital replicas has arrived. Although technologies have long been available to produce fake images or recordings, generative AI[2] technology's ability to do so easily, quickly, and with uncanny verisimilitude has drawn the attention and concern of creators, legislators, and the general public.

As part of a broad AI Initiative, the Copyright Office sought comments on these developments. We asked whether existing laws provide sufficient protection against unauthorized digital replicas or if new protection is needed at the federal level. In response, numerous commenters called for a new federal law to protect individuals from the appropriation of their persona. They provided extensive input into the justifications for and the appropriate parameters of such a law.

In the months since the Office's inquiry was launched, unauthorized digital replicas have continued to make headlines, and have triggered Congressional activity. During this time, we analyzed the comments received, performed additional research, and consulted with other agencies on their relevant areas of expertise. Based on all of this input, we have concluded that a new law is needed. The speed, precision, and scale of AI-created digital replicas calls for prompt federal action. Without a robust nationwide remedy, their unauthorized publication and distribution threaten substantial harm not only in the entertainment and political arenas, but also for private individuals.

Section I summarizes the context and history of the Office's study of the digital replicas issue. Section II.A outlines the main existing legal frameworks: state rights of privacy and publicity, including recent legislation specifically targeting digital replicas, and at the federal level, the Copyright Act, the Federal Trade Commission Act, the Communications Act, and the Lanham Act.

---

[1] For purposes of the Copyright Office's Report on Copyright and Artificial Intelligence, "AI" or "Artificial Intelligence" is a general classification of automated systems designed to perform tasks typically associated with human intelligence or cognitive functions. Artificial Intelligence Study: Notice of Inquiry, 88 Fed. Reg. 59942, 59948 (Aug. 30, 2023) ("NOI"). *See also* John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. 115–232, § 238(g)(2), 132 Stat. 1636, 1697–98 (2018) (defining "artificial intelligence" to include systems "developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action").

[2] "Generative AI" refers to applications of AI used to generate outputs in the form of expressive material such as text, images, audio, or video. NOI at 59948–49.

In Section II.B, we explain why existing laws do not provide sufficient legal redress for those harmed by unauthorized digital replicas and propose the adoption of a new federal law. We make the following recommendations regarding its contours:

- Subject Matter. The statute should target those digital replicas, whether generated by AI or otherwise, that are so realistic that they are difficult to distinguish from authentic depictions. Protection should be narrower than, and distinct from, the broader "name, image, and likeness" protections offered by many states.
- Persons Protected. The statute should cover all individuals, not just celebrities, public figures, or those whose identities have commercial value. Everyone is vulnerable to the harms that unauthorized digital replicas can cause, regardless of their level of fame or prior commercial exposure.
- Term of Protection. Protection should endure at least for the individual's lifetime. Any postmortem protection should be limited in duration, potentially with the option to extend the term if the individual's persona continues to be exploited.
- Infringing Acts. Liability should arise from the distribution or making available of an unauthorized digital replica, but not the act of creation alone. It should not be limited to commercial uses, as the harms caused are often personal in nature. It should require actual knowledge both that the representation was a digital replica of a particular individual and that it was unauthorized.
- Secondary Liability. Traditional tort principles of secondary liability should apply. The statute should include a safe harbor mechanism that incentivizes online service providers to remove unauthorized digital replicas after receiving effective notice or otherwise obtaining knowledge that they are unauthorized.
- Licensing and Assignment. Individuals should be able to license and monetize their digital replica rights, subject to guardrails, but not to assign them outright. Licenses of the rights of minors should require additional safeguards.
- First Amendment Concerns. Free speech concerns should expressly be addressed in the statute. The use of a balancing framework, rather than categorical exemptions, would avoid overbreadth and allow greater flexibility.
- Remedies. Effective remedies should be provided, both injunctive relief and monetary damages. The inclusion of statutory damages and/or prevailing party attorney's fees provisions would ensure that protection is available to individuals regardless of their financial resources. In some circumstances, criminal liability would be appropriate.
- Relationship to State Laws. Given well-established state rights of publicity and privacy, the Office does not recommend full federal preemption. Federal law should provide a floor of consistent protection nationwide, with states continuing to be able

to provide additional protections. It should be clarified that section 114(b) of the Copyright Act does not preempt or conflict with laws restricting unauthorized voice digital replicas.

Section III discusses protection against AI outputs that deliberately imitate an artist's style. We acknowledge the seriousness of creators' concerns and identify legal remedies available to address this type of harm. We do not, however, recommend including style in the coverage of new legislation at this time.

The Office appreciates and has benefitted from the extensive and thoughtful comments we received on this important topic. We remain available to assist as Congress continues to consider legislative solutions.

I. INTRODUCTION

In April of 2023, a new song featuring the voices of Drake and The Weeknd drew over fifteen million views on social media and six hundred thousand listens on Spotify.[3] Yet neither artist was aware of the song before its release, because the vocals were unauthorized, AI-generated replicas.[4]

The viral hit "Heart on My Sleeve," commonly referred to as the "Fake Drake" song, is a high-profile example of a burgeoning subgenre of sound recordings using generative AI systems[5] to create vocals that can pass for those of a favorite artist.[6] Vocal tracks are merely one form of increasingly realistic replicas of individuals' voices, images, and artistic styles.[7] In a short period of time, generative AI technology has become so sophisticated, and so accessible, that minimal expertise is required to rapidly produce such replicas.[8] On social media and other internet platforms, their volume has skyrocketed.[9]

---

[3] Bill Donahue, *Fake Drake & The Weeknd Song — Made With AI — Pulled From Streaming After Going Viral*, BILLBOARD (Apr. 17, 2023), https://www.billboard.com/pro/fake-ai-drake-the-weeknd-song-pulled-streaming/.

[4] Colin Stutz, *The Fake Drake AI Song Earned Millions of Streams — But Will Anyone Get Paid?*, BILLBOARD (Apr. 19, 2023), https://www.billboard.com/pro/fake-drake-ai-song-earned-millions-streams-get-paid/.

[5] An "AI System" is a software product or service that substantially incorporates one or more AI models and is designed for use by an end-user. NOI at 59948; *see also* James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, Pub. L. 117–263, § 7223(4)(A), 136 Stat. 2395, 3669 (2022) (defining "artificial intelligence system" as "any data system, software, application, tool, or utility that operates in whole or in part using dynamic or static machine learning algorithms or other forms of artificial intelligence").

[6] Joe Coscarelli, *An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World*, N.Y. TIMES (Apr. 19, 2023), https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html.

[7] *See, e.g.*, Melissa Heikkilä, *This Artist Is Dominating AI-Generated Art. And He's Not Happy About It*, MIT TECH. REV. (Sept. 16, 2022), https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/; Jane Friedman, *I Would Rather See My Books Get Pirated Than This (Or: Why Goodreads and Amazon Are Becoming Dumpster Fires)*, JANEFRIEDMAN BLOG (Aug. 20, 2023), https://janefriedman.com/i-would-rather-see-my-books-pirated/.

[8] *See* Federal Trade Commission ("FTC") Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Oct. 30, 2023) ("FTC Initial Comments") ("Although policymakers have debated the disruptive potential of AI for decades, the pace of the technology's development and rollout has accelerated in recent years . . . ."); *Fast, Relax, & Turbo Modes*, MIDJOURNEY, https://docs.midjourney.com/docs/fast-relax ("Wait times for Relax are dynamic but generally range between 0–10 minutes per job. . . . Turbo Mode is available for subscribers who want extremely quick image generation. . . . Jobs run in Turbo mode generate up to four times faster. . . .") (last visited July 21, 2024); Karen X. Cheng (@karenxcheng), INSTAGRAM (Apr. 11, 2022), https://www.instagram.com/p/CcN5nBSpO9W/ (demonstrating how to generate an image in seconds using DALL·E).

[9] *E.g.*, Don Philmlee, *Practice Innovations: Seeing is no longer believing — the rise of deepfakes*, THOMSON REUTERS (July 18, 2023), https://www.thomsonreuters.com/en-us/posts/technology/practice-innovations-deepfakes/.

A. AI and Digital Replicas

This Report uses the term "digital replica" to refer to a video, image, or audio recording that has been digitally created or manipulated to realistically but falsely depict an individual. A "digital replica" may be authorized or unauthorized and can be produced by any type of digital technology, not just AI. The terms "digital replicas" and "deepfakes" are used here interchangeably.[10]

Digital replicas may have both beneficial and harmful uses. On the positive side, they can serve as accessibility tools for people with disabilities,[11] enable "performances" by deceased or non-touring artists,[12] support creative work,[13] or allow individuals to license, and be compensated for, the use of their voice, image, and likeness.[14] In one noted example, musician

[10] Although the term "deepfake" is often associated with unauthorized or deceptive uses, especially in explicit imagery, *see infra* notes 22–23, some dictionary definitions are broader. *See deepfake*, MERRIAM-WEBSTER DICTIONARY, https://www.merriam-webster.com/dictionary/deepfake ("an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said") (last updated July 20, 2024); *deepfake*, CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/us/dictionary/english/deepfake ("a video or sound recording that replaces someone's face or voice with that of someone else, in a way that appears real"). In popular media too, the term has been used to describe authorized uses as well as malicious ones. *See* Nilesh Christopher & Varsha Bansal, *Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve*, WIRED (May 20, 2024), https://www.wired.com/story/indian-elections-ai-deepfakes/ ("Politicians are using audio and video deepfakes of themselves to reach voters—who may have no idea they've been talking to a clone.").

[11] *E.g.*, Press Release, Office of Congresswoman Jennifer Wexton, Wexton Shares Video Debuting New AI Voice Model (July 10, 2024), https://wexton.house.gov/news/documentsingle.aspx?DocumentID=952 ("Today, Congresswoman Jennifer Wexton (D-VA) shared a video debuting a new Artificial Intelligence-generated model of her voice as it was before being impacted by her Progressive Supranuclear Palsy (PSP) condition.").

[12] Universal Music Group ("UMG"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5–6 (Oct. 30, 2023) ("UMG Initial Comments"); *see also* Elias Leight, *Will AI Be Used to Raise Musicians From the Dead*?, BILLBOARD (Nov. 29, 2023), https://www.billboard.com/pro/ai-bring-back-dead-artists-musicians-estate-managers/.

[13] *See, e.g.*, Letter from Motion Picture Association ("MPA"), Summary of *Ex Parte* Meeting on May 13, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (May 20, 2024) ("MPA highlighted the importance of this issue to our members, given the ubiquity of depiction of individuals in docudramas, biopics, and similar works. And we argued that use of digital-replica technology is simply an evolution of the type of technology our members have long used to make actors more closely resemble the people they portray, including make-up and prosthetics.").

[14] *See, e.g.*, American Association of Independent Music ("A2IM") et al., Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1–2 (Dec. 6, 2023) ("A2IM-Recording Academy-RIAA Joint Reply Comments"); William Morris Endeavor Entertainment, LLC ("WME"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Oct. 30, 2023) ("WME Initial Comments").

Randy Travis, who has limited speech function since suffering a stroke, was able to use generative AI to release his first song in over a decade.[15]

At the same time, a broad range of actual or potential harms arising from unauthorized digital replicas has emerged. Across the creative sector, the surge of voice clones and image generators has stoked fears that performers and other artists will lose work or income.[16] There have already been film projects that use digital replica extras in lieu of background actors,[17] and situations where voice actors have been replaced by AI replicas.[18] Within the music industry, concerns have been raised that the use of AI in sound recordings could lead to the "loss of authenticity and creativity" and displacement of human labor.[19] Numerous commenters,

---

[15] Dylan Smith, *Randy Travis Harnesses AI to Release His 'First New Music in More Than a Decade' – Another Song Is Already Being Created*, DIGIT. MUSIC NEWS (May 6, 2024), https://www.digitalmusicnews.com/2024/05/06/randy-travis-new-song.

[16] *See, e.g.*, Screen Actors Guild-American Federation of Television Artists ("SAG-AFTRA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 4 (Oct. 30, 2023) ("SAG-AFTRA Initial Comments"); Writers Guild of America ("WGA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2–3 (Oct. 30, 2023) ("WGA Initial Comments"); Christi Carras, *Which Entertainment Jobs Are Most Likely to Be Disrupted by AI? New Study Has Answers*, L.A. TIMES (Jan. 30, 2024), https://www.latimes.com/entertainment-arts/business/story/2024-01-30/ai-artificial-intelligence-impact-report-entertainment-industry; Sam O'brien, *AiArt: Why Some Artists Are Furious About AI-Produced Art*, IEEE COMPUT. SOC'Y (Nov. 29, 2023), https://www.computer.org/publications/tech-news/trends/artists-mad-at-ai; Ari's Take, *Is AI Music Taking Royalties From Musicians and Composers? — The New Music Business Podcast*, YOUTUBE, at 20:57 (Jan. 17, 2024), https://www.youtube.com/watch?v=3I9dnxzjEzg.

[17] Jeremy Dick, *Disney Gets Roasted for 'Creepy' AI Extras in Disney+ Movie Prom Pact*, CBR (Oct. 12, 2023), https://www.cbr.com/disney-prom-pact-ai-actors/; *see also* Bobby Allyn, *Movie Extras Worry They'll be Replaced by AI. Hollywood Is Already Doing Body Scans*, NPR (Aug. 2, 2023), https://www.npr.org/2023/08/02/1190605685/. As we discuss below, there have been steps taken to address these concerns through private and collective bargaining agreements. *See infra* Section II.A.3.

[18] *See, e.g.*, FTC Initial Comments, Attach. at 37 (statement of Tim Friedlander, Nat'l Ass'n of Voice Actors); Cade Metz, *What Do You Do When A.I. Takes Your Voice?*, N.Y. TIMES (May 16, 2024), https://www.nytimes.com/2024/05/16/technology/ai-voice-clone-lawsuit.html; Ed Nightingale, *Baldur's Gate 3 Actors Reveal the Darker Side of Success Fuelled by AI Voice Cloning*, EUROGAMER (Apr. 12, 2024), https://www.eurogamer.net/baldurs-gate-3-actors-reveal-the-darker-side-of-success-fuelled-by-ai-voice-cloning (addressing concerns by video game voice actors).

[19] *See, e.g.*, CVL ECONOMICS, FUTURE UNSCRIPTED: THE IMPACT OF GENERATIVE AI ON ENTERTAINMENT INDUSTRY JOBS 39–40 (Jan. 2024), https://animationguild.org/wp-content/uploads/2024/01/Future-Unscripted-The-Impact-of-Generative-Artificial-Intelligence-on-Entertainment-Industry-Jobs-pages-1.pdf ("With the capability to recreate melodies and replicate musicians' voices convincingly and quickly, it is becoming easier than ever to generate a music track without any direct human involvement."); Jem Aswad, *Billie Eilish, Nicki Minaj, Stevie Wonder, Dozens More Call on AI Developers to Respect Artists' Rights*, VARIETY (Apr. 2, 2024), https://variety.com/2024/music/news/billie-eilish-nicki-minaj-ai-respect-artists-rights-1235957451/.

among them many SAG-AFTRA members, stressed the importance of performers being able to prevent such displacement as well as the resulting impacts on their careers and livelihoods.[20]

While digital replicas depicting well-known individuals often attract the most attention, anyone can be vulnerable.[21] Beyond the creative sector, the harms from unauthorized digital replicas largely fall into three categories. First, there have been many reports of generative AI systems being used to produce sexually explicit deepfake imagery.[22] In 2023, researchers concluded that explicit images make up 98% of all deepfake videos online, with 99% of the individuals represented being women.[23] Instances of students creating and posting deepfake explicit images of classmates appear to be multiplying.[24]

---

[20] *See, e.g.*, Morgan Keaton, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Oct. 30, 2023); Allie Radice, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Oct. 29, 2023); Gregory Schott, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Oct. 23, 2023).

[21] *E.g.*, Arian Garshi, *Deepfakes in 2022: How individual non-celebrities are targeted,* MEDIUM (Oct. 17, 2022), https://ariangarshi.medium.com/deepfakes-in-2022-how-individual-non-celebrities-are-targeted-a7dab59cac3a.

[22] *See, e.g.*, Caroline Haskins, A *Deepfake Nude Generator Reveals a Chilling Look at Its Victims*, WIRED (Mar. 25, 2024), https://www.wired.com/story/deepfake-nude-generator-chilling-look-at-its-victims/; Katherine Noel, *Journalist Emanuel Maiberg Addresses AI and the Rise of Deepfake Pornography*, INST. OF GLOBAL POL. (Apr. 22, 2024), https://igp.sipa.columbia.edu/news/rise-deepfake-pornography. On social media platform X alone, one sexually explicit deepfake image of Taylor Swift was "viewed 47 million times before the account [where the image was posted] was suspended." Kate Conger & John Yoon, *Explicit Deepfake Images of Taylor Swift Elude Safeguards and Swamp Social Media*, N.Y. TIMES (Jan. 26, 2024), https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html; Ashley King, *Megan Thee Stallion the Latest Victim of Deepfake Porn — X/Twitter 'Proactively Removing' Clips*, DIGI. MUSIC NEWS (June 14, 2024), https://www.digitalmusicnews.com/2024/06/14/megan-thee-stallion-deepfake-porn-x-twitter-removing.

[23] HOME SECURITY HEROES, 2023 STATE OF DEEPFAKES at Key Findings 2–3 (2023), https://www.homesecurityheroes.com/state-of-deepfakes/#key-findings. *See also* Katherine Noel, *Journalist Emanuel Maiberg Addresses AI and the Rise of Deepfake Pornography*, INST. OF GLOBAL POL. (Apr. 22, 2024), https://igp.sipa.columbia.edu/news/rise-deepfake-pornography ("It is almost exclusively young women who are nonconsensually being undressed and put into AI-generated porn.") (internal quotation marks omitted).

[24] *E.g.*, Natasha Singer, *Teen Girls Confront an Epidemic of Deepfake Nudes in Schools*, N.Y. TIMES (Apr. 8, 2024), https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html; Caroline Haskins, *Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates*, WIRED (Mar. 8, 2024), https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/; Cameron Sires, *Schools Navigate The New World of Explicit AI-Generated Images*, ISSAQUAH REPORTER (Apr. 16, 2024), https://www.issaquahreporter.com/news/schools-navigate-the-new-world-of-explicit-ai-generated-images/; Miranda Ceja, *AI-Generated Nude Photos Of High Schoolers Investigated In South OC*, MSN NEWS (Apr. 2, 2024), https://www.msn.com/en-us/news/us/ai-generated-nude-photos-of-high-schoolers-investigated-in-south-oc/ar-BB1kXrkt.

Second, the ability to create deepfakes offers a "potent means to perpetrate fraudulent activities with alarming ease and sophistication."[25] The media has reported on scams in which defrauders replicated the images and voices of a multinational financial firm's CEO and its employees to steal $25.6 million;[26] replicated loved ones' voices to demand a ransom;[27] and replicated the voice of an attorney's son asking him to wire $9,000 to post a bond.[28] Digital replicas of celebrities have been used to falsely portray them as endorsing products.[29]

Finally, there is a danger that digital replicas will undermine our political system and news reporting by making misinformation impossible to discern. Recent examples involving politicians include a voice replica of a Chicago mayoral candidate appearing to condone police brutality;[30] a robocall with a replica of President Biden's voice discouraging voters from participating in a primary election;[31] and a campaign ad that used AI-generated images to depict former President Trump appearing with former Director of the National Institute of Allergy and Infectious Diseases, Anthony Fauci.[32] Deepfake videos were even used to influence a high

---

[25] Giuseppe Ciccomascolo, *Deepfakes Make Up 66% of AI Fraud While Crypto Scams Halved*, CCN (Apr. 25, 2024), https://www.ccn.com/news/technology/deepfakes-ai-fraud-crypto-scams/.

[26] Heather Chen & Kathleen Magramo, *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'*, CNN (Feb. 4, 2024), https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html ("The elaborate scam saw the [finance] worker duped into attending a video call with what he thought were several other members of staff, but all of whom were in fact deepfake recreations . . . . [T]he worker put aside his early doubts after the video call because other people in attendance had looked and sounded just like colleagues he recognized . . . .").

[27] Charles Bethea, *The Terrifying A.I. Scam That Uses Your Loved One's Voice*, NEW YORKER (Mar. 7, 2024), https://www.newyorker.com/science/annals-of-artificial-intelligence/the-terrifying-ai-scam-that-uses-your-loved-ones-voice.

[28] Samantha Manning, *Father Warns Congress About AI scammer Who Sounded Just Like His Son*, KIRO7 (Nov. 16, 2023), https://www.kiro7.com/news/local/father-warns-congress-about-ai-scammer-who-sounded-just-like-his-son/KA7BXJJ2OJB3NHDDM4EGB5L24M/.

[29] *See, e.g.*, Megan Cerullo, *AI-generated Ads Using Taylor Swift's Likeness Dupe Fans With Fake Le Creuset Giveaway*, CBS NEWS (Jan. 16, 2024), https://www.cbsnews.com/news/taylor-swift-le-creuset-ai-generated-ads/; *Tom Hanks Says AI Version of Him Used In Dental Plan Ad Without His Consent*, THE GUARDIAN (Oct. 1, 2023), https://www.theguardian.com/film/2023/oct/02/tom-hanks-dental-ad-ai-version-fake; Jamey Tucker, *Fake Ads Made With Artificial Intelligence Exploit Celebrities on Social Media*, WPSD LOCAL 6 (Mar. 11, 2024), https://www.wpsdlocal6.com/news/fake-ads-made-with-artificial-intelligence-exploit-celebrities-on-social-media/article_7a94d8a6-dfba-11ee-bb47-7f2861477d8e.html.

[30] Tiffany Hsu & Steven Lee Myers, *A.I.'s Use in Elections Sets Off a Scramble for Guardrails*, N.Y. TIMES (June 25, 2023), https://www.nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html.

[31] Shannon Bond, *AI Fakes Raise Election Risks As Lawmakers and Tech Companies Scramble to Catch Up*, NPR (Feb. 8, 2024), https://www.npr.org/2024/02/08/1229641751.

[32] Nicholas Nehamas, *DeSantis Campaign Uses Apparently Fake Images to Attack Trump on Twitter*, N.Y. TIMES (June 8, 2023), https://www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html.

profile union vote by falsely showing a union leader urging members to oppose the contract that he had "negotiated and . . . strongly supported."[33]

Summarizing the challenges to the information ecosystem, one digital forensics scholar cautioned, "[i]f we enter a world where any story, any audio recording, any image, any video can be fake . . . then nothing has to be real."[34] As AI technology continues to improve, researchers predict that it will become increasingly difficult to distinguish between digital replicas and authentic content.[35]

B. Background of This Study

In early 2023, the Copyright Office announced an initiative to examine the copyright issues raised by AI. Over the following months, we hosted public listening sessions and engaged in extensive outreach to better understand the issues, including those related to generative AI's ability to produce digital replicas.[36]

The topic of digital replicas does not fall neatly under any one area of existing law. While some characterize it as a form of intellectual property, protection against the use of unauthorized digital replicas raises overlapping issues including privacy, unfair competition, consumer protection, and fraud. It relates to copyright in a number of ways: creators such as artists and performers are particularly affected; copyrighted works are often used to produce digital replicas; and the replicas are often disseminated as part of larger copyrighted works. Moreover, the noncommercial harms that may be caused are similar to violations of moral rights protected in part through the copyright system.[37]

In August 2023, the Office published a Notice of Inquiry on AI and Copyright that sought input on "the treatment of generative AI outputs that imitate the identity or style of human artists," among other topics.[38] The NOI asked what existing laws apply to AI-generated material that features the voice or likeness of a particular person; whether Congress should enact a new federal law that would protect against unauthorized digital replicas; and, if so,

---

[33] *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Duncan Crabtree-Ireland, National Executive Director, SAG-AFTRA).

[34] Weekend Edition Sunday, *As Tech Evolves, Deepfakes Will Become Even Harder to Spot*, NPR (July 3, 2022), https://www.npr.org/2022/07/03/1109607618 (Interview with Professor Hany Farid of U.C. Berkeley).

[35] *E.g.*, Weekend Edition Sunday, *As Tech Evolves, Deepfakes Will Become Even Harder to Spot*, NPR (July 3, 2022), https://www.npr.org/2022/07/03/1109607618 (Interview with Professor Hany Farid of U.C. Berkeley).

[36] *See, e.g.*, Music and Sound Recordings Listening Session Tr. at 11:11–12; 16:19–17:4; 73:11–77:20 (May 31, 2023) (statements by Nathaniel Bach, Music Artists Coalition ("MAC"); Kenneth Doroshow, Recording Industry Association of America ("RIAA"); Rohan Paul, Controlla; Garrett Levin, Digital Media Association ("DiMA")).

[37] *See infra* note 41.

[38] NOI at 59945.

what its contours should be. We also inquired whether there are or should be protections against AI systems generating outputs that imitate artistic style.[39] Finally, we sought views on how, for sound recordings, section 114(b) of the Copyright Act relates to state laws protecting against the imitation of an individual's voice.[40]

The Office received approximately one thousand comments responding to this group of questions, over 90% of them from individuals. The majority advocated for the enactment of new federal legislation. The scope, duration, and assignability of the right to be provided, as well as its relationship to existing state laws, were the subject of greater disagreement.

The copying of an individual's identity is not an entirely new topic for the Copyright Office. In 2019, we published a report on the moral rights of attribution and integrity[41] in the United States, in which we recommended that Congress consider adopting a federal right of publicity.[42] The current study has a narrower focus—assessing the need for federal protection specifically with respect to unauthorized digital replicas.

The Office concludes that the time has come to adopt such a law at the federal level. Based on our analysis of the comments received, independent research, and a review of work being done at other agencies, we believe there is an urgent need for a robust nationwide remedy beyond those that already exist. In the sections below, we review the protections available under current laws and the gaps in their capacity to respond to today's threats, explain the reasons for new federal protection, and provide recommendations regarding its contours.

We then address requests for protection against AI outputs that mimic or appropriate an artist's style. While the Office acknowledges the seriousness of this concern, we believe that existing laws may provide sufficient protection at this time.

---

[39] *Id.* at 59945, 59948.

[40] *Id.* at 59948.

[41] Moral rights are non-economic rights in copyrighted works that are considered personal to the authors. The two most commonly recognized are the right of attribution (being credited as the author) and the right of integrity (preventing distortions of the work). U.S. COPYRIGHT OFFICE, AUTHORS, ATTRIBUTION, AND INTEGRITY: EXAMINING MORAL RIGHTS IN THE UNITED STATES 6 (2019) ("MORAL RIGHTS REPORT"), https://copyright.gov/policy/moralrights/full-report.pdf. The United States provides these moral rights through a combination of federal and state laws, most of which are described below, including the Lanham Act, certain provisions of the Copyright Act, and state laws relating to privacy and publicity, contracts, fraud and misrepresentation, unfair competition, and defamation. *See id.* at 7.

[42] *Id.* at 110–19.

II. PROTECTION AGAINST UNAUTHORIZED DIGITAL REPLICAS

A. Existing Legal Frameworks

A variety of legal frameworks provide protection against the unauthorized use of aspects of an individual's persona. Some exist at the state level, including statutory and common law rights of privacy and publicity. Others are based on federal law, including the Copyright Act, the Federal Trade Commission Act, the Lanham Act, and the Communications Act.

1. State Common and Statutory Law

The most directly relevant state laws are the long-standing rights of publicity and privacy. In response to the accelerating pace of replicas created by generative AI systems, many states are also considering or have enacted new legislation specifically directed at unauthorized digital replicas.

a) Right of Privacy

The common law right of privacy emerged in the late 19th century and has been described as protecting against unreasonable intrusions into individuals' private lives, safeguarding their autonomy, dignity, and personal integrity.[43] Privacy rights are considered personal to the individual and typically apply only to the living.[44] Most states recognize some form of the right of privacy, either through statutory or common law.[45]

The common law right of privacy has been described as a complex of torts,[46] with the torts of false light and of appropriation of name and likeness most relevant here.[47] False light invasion of privacy protects the reputation of individuals, "with the same overtones of mental

---

[43] Samuel Warren and Louis Brandeis first argued for a law to protect the right of privacy in their 1890 article, *The Right to Privacy*, describing "the next step which must be taken for the protection of the person, and for securing to the individual . . . the right 'to be let alone.'" Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890).

[44] RESTATEMENT (SECOND) OF TORTS § 652I (AM. L. INST. 1977).

[45] 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY §§ 6:1, 6:7 (2d ed. 2024).

[46] William L. Prosser, *Privacy*, 48 CALIF. L. REV. 383, 389 (1960) ("It is not one tort, but a complex of four."). The four-tort complex includes: (1) intrusion upon seclusion or solitude, (2) disclosure of embarrassing private facts, (3) false light, and (4) appropriation of a person's name or likeness for the defendant's advantage. DAN B. DOBBS, PAUL T. HAYDEN & ELLEN M. BUBLICK, THE LAW OF TORTS § 578 (2d ed. 2024).

[47] *See* Nicholas Schmidt, *Privacy Law and Resolving 'Deepfakes' Online*, IAPP (Jan. 30, 2019), https://iapp.org/news/a/privacy-law-and-resolving-deepfakes-online/; *see also* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1794–95 (2019).

distress as in defamation."[48] Liability arises when someone "gives publicity to a matter concerning another that places [them] before the public in a false light," if the false light is "highly offensive to a reasonable person," and if "the actor had knowledge of or acted in reckless disregard as to the falsity."[49] For example, courts have found liability where a defendant spread false statements that "attribut[ed] a lewd fantasy" to a woman and claimed she agreed to appear nude in an adult magazine,[50] as well as where a defendant used individuals' names and likenesses in promotions for strip clubs without their consent.[51] The majority of jurisdictions have recognized this tort,[52] with a few incorporating it by statute.[53]

False light invasion of privacy may provide some legal protection against unauthorized digital replicas when they are used to depict an individual participating in offensive conduct.[54] It would appear particularly appropriate to address deepfake pornography. However, the objective "highly offensive" standard will limit its applicability to other uses of unauthorized digital replicas, such as depictions that are merely untruthful.[55]

---

[48] William L. Prosser, *Privacy*, 48 CALIF. L. REV. 383, 400 (1960). *See also* 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:112 (2d ed. 2024) ("The difference between false light invasion of privacy and defamation is still unclear. While the false light tort primarily focuses upon indignity and defamation focuses upon reputation, the distinction is a subtle one."). Although some courts view defamation as duplicative of false light invasion of privacy, *see, e.g.*, *Denver Pub. Co. v. Bueno*, 54 P.3d 893, 894 (Colo. 2002), defamatory statements are not necessary for an individual to be placed in a false light. *See* RESTATEMENT (SECOND) OF TORTS § 652E (AM. L. INST. 1977).

[49] RESTATEMENT (SECOND) OF TORTS § 652E (AM. L. INST. 1977). "The courts uniformly adopt the Restatement of Torts list of elements." 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:114 n.1 (2d ed. 2024).

[50] *Wood v. Hustler Mag., Inc.*, 736 F.2d 1084, 1089, 1093 (5th Cir. 1984).

[51] *Longoria v. Kodiak Concepts LLC*, 527 F. Supp. 3d 1085, 1102 (D. Ariz. 2021); *Johnson v. J.P. Parking, Inc.*, No. 4:22-cv-00146, 2024 WL 676770, at *18 (S.D. Iowa Feb. 20, 2024).

[52] *Welling v. Weinfeld*, 866 N.E.2d 1051, 1055 (Ohio 2007) ("A majority of jurisdictions in the United States have recognized false-light invasion of privacy as a distinct, actionable tort."). However, some states, such as Colorado, Florida, Minnesota, New York, North Carolina, and Texas, have rejected or do not recognize the false light tort. 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:115 (2d ed. 2024).

[53] *E.g.*, 9 R.I. GEN. LAWS § 9-1-28.1(a)(4) (2024) (providing for a "right to be secure from publicity that reasonably places another in a false light before the public," and allowing recovery if "[t]here has been some publication of a false or fictitious fact which implies an association which does not exist" and "[t]he association which has been published or implied would be objectionable to the ordinary reasonable man under the circumstances"); NEB. REV. STAT. § 20-204 (2024) (following the *Restatement (Second) of Torts*' formulation).

[54] *See* Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99, 115–16 (2019).

[55] *See, e.g.*, *De Havilland v. FX Networks, LLC*, 230 Cal.Rptr.3d 625, 630, 644 (Cal. Ct. App. 2018) (rejecting a false light claim for a docudrama's fictionalized interview because it would not "subject a person to hatred, contempt, ridicule, or obloquy").

The related tort of invasion of privacy by appropriation involves "appropriation of the plaintiff's identity or reputation, or some substantial aspect of it, for the defendant's own use or benefit."[56] On its face, this tort appears well suited to protect against unauthorized digital replicas,[57] although not every state recognizes it.[58] Courts, however, have not interpreted the tort consistently. Several states require that the appropriative act be for commercial purposes or purposes of trade,[59] excluding claimants harmed by noncommercial uses. Although in most jurisdictions the tort is available to any member of the public, some require a showing that the name or likeness has "intrinsic value," limiting protection to individuals who are well-known.[60]

b) Right of Publicity

The right of publicity addresses the use of individuals' personas[61] in commercial contexts, aiming to prevent others from profiting from unauthorized uses. The right evolved from the tort of invasion of privacy by appropriation to protect celebrities and well-known figures.[62] In 1953, the Second Circuit coined the term "right of publicity" in *Haelan Laboratories, Inc. v. Topps Chewing Gum, Inc.*, where it held that "in addition to and independent of that right

---

[56] DAN B. DOBBS, PAUL T. HAYDEN & ELLEN M. BUBLICK, THE LAW OF TORTS § 579 (2d ed. 2024). Common law elements of invasion of privacy include (1) use, without permission, of "some aspect of the plaintiff's identity or persona in such a way that plaintiff is identifiable from defendant's use," and (2) that the use "causes some damage to plaintiff's peace of mind and dignity, with resulting injury measured by plaintiff's mental or physical distress and related damage." 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:62 (2d ed. 2024) (footnote omitted).

[57] Shannon Reid, *The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections*, 23 U. PA. J. CONST. L. 209, 215 (2021).

[58] *See, e.g.*, *Hougum v. Valley Mem'l Homes*, 1998 ND 24, ¶ 12, 574 N.W.2d 812, 816 ("This Court has not decided whether a tort action exists in North Dakota for invasion of privacy."); *Nelson v. J.C. Penney Co.*, 75 F.3d 343, 347 (8th Cir. 1996).

[59] *See, e.g.*, *Barbieri v. News-J. Co.*, 56 Del. 67, 70, 189 A.2d 773, 774 (1963); *Fergerstrom v. Hawaiian Ocean View Ests.*, 441 P.2d 141, 144 (1968).

[60] *See* 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:62 (2d ed. 2024). Right of privacy laws generally do not address other issues, such as secondary liability or First Amendment exceptions, at the level of detail that right of publicity laws do, as discussed below. We also received fewer comments focused on these issues from a right of privacy perspective.

[61] The term "persona" in right of publicity law "is increasingly used as a label to signify the cluster of commercial values embodied in personal identity as well as to signify that human identity 'identifiable' from defendant's usage. There are many ways in which a 'persona' is identifiable: from name, nickname and voice, to picture or performing style and other indicia which identify the 'persona' of a person." *Id.* § 4:46.

[62] *Id.* § 1:25; Melville B. Nimmer, *The Right of Publicity*, 19 L. & CONTEMP. PROBS. 203, 203–04 (1954) ("[T]he [privacy] doctrine, first developed to protect the sensibilities of nineteenth century Brahmin Boston, is not adequate to meet the demands of the second half of the twentieth century, particularly with respect to the advertising, motion picture, television, and radio industries. Well known personalities connected with these industries do not seek the 'solitude and privacy' which Brandeis and Warren sought to protect.").

of privacy . . . , a man has a right in the publicity value of his photograph, i.e., the right to grant the exclusive privilege of publishing his picture. . . ."[63]

A majority of states now recognize the right of publicity by statute, common law, or both.[64] Because the harms that the right of publicity and the tort of invasion of privacy address are similar, some jurisdictions use the terms interchangeably,[65] while others treat them as distinct.[66]

Intended to protect aspects of an individual's identity, the right of publicity may be the most apt state law remedy for unauthorized digital replicas.[67] Numerous commenters noted, however, that the contours of the right differ considerably from state to state.[68] As to the subject matter, in some states the law sweeps more broadly than digital replicas, capturing aspects of

---

[63] 202 F.2d 866, 868 (2d Cir. 1953).

[64] 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 6:2 (2d ed. 2024). Some states, such as Alaska, Kansas, Maryland, and North Carolina, have neither statutory nor common law rights of publicity. *See* Jennifer Rothman, *Rothman's Roadmap to the Right of Publicity*, https://rightofpublicityroadmap.com/law/ (last visited July 21, 2024). There are also a few states, such as Colorado, Delaware, and Oregon, that have no statutory right of publicity but where the existence of a common-law right is unclear. *Id. See generally* 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 1:2 (2d ed. 2024).

[65] *E.g.*, *Rosa & Raymond Parks Inst. for Self Dev. v. Target Corp.*, 812 F.3d 824, 830 (11th Cir. 2016) ("The last category of invasion of privacy—misappropriation of a person's name or likeness—is commonly referred to as a violation of the 'right of publicity.'"); *In re Jackson*, 972 F.3d 25, 34 (2d Cir. 2020) (using the *Restatement (Second) of Torts*' description of liability for invasion of privacy by appropriation to describe right of publicity liability).

[66] *See, e.g.*, *Minnifield v. Ashcraft*, 903 So. 2d 818, 826 (Ala. Civ. App. 2004) ("[W]e cannot say that the commercial-appropriation invasion-of-privacy tort in Alabama bases its liability solely on commercial rather than psychological interests. To do so would, in effect, substitute the commercial-appropriation invasion-of-privacy tort with the tort of violating the right to publicity.").

[67] A number of commenters agreed that existing rights of publicity would apply to certain uses of digital replicas. *See, e.g.*, Jennifer Rothman, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2–3 (Oct. 25, 2023) ("Jennifer Rothman Initial Comments") (noting that, in the context of an AI-generated Tom Hanks appearing in an advertisement and the AI-generated voices of Drake and The Weeknd appearing in a song, "[a]bsent jurisdictional hurdles, . . . each have straightforward lawsuits under state right of publicity laws for the uses described"); SAG-AFTRA Initial Comments at 6–7; *see also* Russell Spivak, *"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 383–85 (2019); Alexandra Curren, *Digital Replicas: Harm Caused by Actors' Digital Twins and Hope Provided by the Right of Publicity*, 102 TEX. L. REV. 155, 164, 166–67 (2023). Indeed, courts have applied standard right of publicity laws to other forms of digital likenesses. *See, e.g.*, *Hart v. Elec. Arts, Inc.*, 717 F.3d 141 (3d Cir. 2013) (digital avatars of football players); *No Doubt v. Activision Publ'g, Inc.*, 122 Cal. Rptr. 3d 397 (Cal. Ct. App. 2011) (digital avatars of the musical group No Doubt).

[68] *See, e.g.*, The Honorable Marsha Blackburn, U.S. Senator, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2–3 (Oct. 30, 2023) ("Senator Marsha Blackburn Initial Comments"); Computer & Communications Industry Association, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 24 (Oct. 30, 2023); Daniel Gervais, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 9 (Oct. 30, 2023).

identity that merely evoke or call to mind the protected individual. In one well-known example, the Ninth Circuit found that a robotic depiction of a blonde woman in a long gown turning large block letters on a game-show set sufficiently "evoked" Vanna White, even without using her name or image, to state a claim under California's common law right of publicity.[69] Further, some states protect additional aspects of identity, such as gestures and mannerisms in Indiana[70] or "any attribute of an individual that serves to identify that individual to an ordinary, reasonable viewer or listener" in Illinois.[71]

In other cases, the laws are written too narrowly to cover all types of digital replica uses. Some states restrict the right to limited groups of individuals, from professional performers,[72] to soldiers[73] or the deceased.[74]

Protection for Postmortem Rights. The treatment of postmortem rights of publicity is one of the areas of greatest variation.[75] Twenty-seven states currently provide postmortem rights of publicity, 19 by statute and 8 by common law.[76] The durations vary from as short as 20 years in Virginia, to 100 years in Indiana, and indefinitely in Tennessee, as long as the right is continuously exploited.[77] The postmortem term can differ depending on the ongoing

---

[69] *White v. Samsung Electronics America, Inc.*, 971 F.2d 1395, 1399 (9th Cir. 1992), *as amended* (Aug. 19, 1992) ("The identities of the most popular celebrities are not only the most attractive for advertisers, but also the easiest to evoke without resorting to obvious means such as name, likeness, or voice."). *See also* Stacy L. Dogan, *An Exclusive Right to Evoke*, 44 B.C. L. REV. 291, 292 (2003) (citing cases where "non-proprietary symbols" were held to call to mind, and thus violate the right of publicity of, various celebrities).

[70] IND. CODE § 32-36-1-6 (2024).

[71] 765 ILL. COMP. STAT. 1075/5 (West 2024).

[72] *See infra* Section II.A.1.c.

[73] ARIZ. REV. STAT. ANN. § 12-761 (2024).

[74] TEX. PROP. CODE ANN. § 26.002 (West 2023). As mentioned above, New York's digital-replicas-specific amendment applies only to deceased performers. N.Y. CIV. RIGHTS LAW § 50-f(1)(a)–(b), (2)(b) (McKinney 2024).

[75] *See, e.g.*, CAL. CIV. CODE §§ 3344, 3344.1 (West 2024) (protecting against the unconsented commercial use of a person's name, voice, signature, photograph, or likeness, affording a 70-year postmortem term, and requiring registration by the successor in interest); KY. REV. STAT. ANN. § 391.170 (West 2024) (recognizing a right in name and likeness, and protecting against the commercial use of the name or likeness of a "public figure" for 50 years after death); *see also* Mary LaFrance, *Choice of Law and the Right of Publicity: Rethinking the Domicile Rule*, 37 CARDOZO ARTS & ENT. L.J. 1, 2 (2019).

[76] 2 J. THOMAS MCCARTHY AND ROGER E. SCHECHTER, RIGHTS OF PUBLICITY AND PRIVACY § 9:17 (2d ed.).

[77] *See* VA. CODE ANN. § 8.01-40 (2024); IND. CODE § 32-36-1-8(a) (2024); TENN. CODE ANN. § 47-25-1104 (2024).

commercial exploitation of the individual's identity, its commercial value at the time of death, and whether the estate complied with statutory registration requirements.[78]

Commercial Use Requirement. Many right of publicity laws only protect against unauthorized commercial uses. These uses may include advertising campaigns, product endorsements, merchandising, and sponsored content,[79] and may extend to newer forms of commercial exploitation facilitated by digital platforms, such as influencer marketing and brand partnerships on social media.[80]

Secondary and Intermediary Liability. Most state statutes do not specify rules for potential secondary liability.[81] However, courts have interpreted these laws as incorporating ordinary tort law principles of aiding and abetting liability, so that a party may be secondarily liable for infringing the right where it has knowledge of the illegal acts and provides substantial assistance.[82]

Several states explicitly limit liability for certain types of intermediaries, where they lack knowledge of the unauthorized acts. California, Pennsylvania, Ohio, and New York exempt advertising media from liability so long as they do not have knowledge that the use of the name, image, or likeness is unauthorized.[83] Arkansas, borrowing concepts from federal copyright law,[84] exempts the "service provider of a system or network" if the service provider does not have actual knowledge that the use is unlawful and is not aware of facts and

---

[78] *See, e.g.*, CAL. CIV. CODE § 3344.1(h) (West 2024) (applying to deceased personalities "whose name, voice, signature, photograph, or likeness has commercial value at the time of his or her death, or because of his or her death").

[79] *See, e.g.*, *Onassis v. Christian Dior-New York, Inc.*, 472 N.Y.S.2d 254 (1984), *judgment aff'd*, 488 N.Y.S.2d 943 (1985) (finding that the use of a look-alike in an advertisement infringed Jacqueline Onassis's right of publicity); *Beverley v. Choices Women's Med. Ctr., Inc.*, 532 N.Y.S.2d 400 (1991) (holding that the use of a female physician's photo on a calendar distributed by and advertising the defendant's business infringed the physician's right of publicity).

[80] Grace Greene, *Instagram Lookalikes and Celebrity Influencers: Rethinking the Right of Publicity in the Social Media Age*, 168 U. PA. L. REV. ONLINE 153, 189–92 (2020) (citing several (settled) right of publicity suits by online influencers seeking to protect their exact images as well as their influencer personas).

[81] *See* Alexandra Curren, *Digital Replicas: Harm Caused by Actors' Digital Twins and Hope Provided by the Right of Publicity*, 102 TEX. L. REV. 155, 164, 166–67 (2023).

[82] *See, e.g.*, *Perfect 10, Inc. v. Cybernet Ventures, Inc.*, 213 F. Supp. 2d 1146, 1183–84 (C.D. Cal. 2002); *Keller v. Elecs. Arts, Inc.*, No. 09-cv-1967, 2010 WL 530108, at *3 (N.D. Cal. Feb. 8, 2010) (allowing civil conspiracy claims for violation of California right of publicity to proceed based on defendant's alleged direction of users to infringing websites).

[83] *See* CAL. CIV. CODE §§ 3344(f), 3344.1(a)(l) (West 2024); 42 PA. STAT. AND CONS. STAT. § 8316(d) (2024) (described as those "in the business of producing, manufacturing, publishing or disseminating material for commercial or advertising purposes by any communications medium"); OHIO REV. CODE ANN. § 2741.02(E) (West 2024); N.Y. CIV. RIGHTS LAW § 50-f(9) (McKinney 2024) (adding a "by prior notification" element to knowledge).

[84] *See infra* Section II.B.3.d.

circumstances that make a violation apparent.[85] A number of courts have found intermediaries not liable for state right of publicity violations where they served as "mere conduits" for the unlawful activity.[86]

First Amendment Protections. States have adopted varied approaches to accommodating First Amendment concerns, either by statute or judicial interpretation.[87] A number of statutes provide carveouts for categories of conduct likely to implicate protected speech, such as news reporting, sports broadcasts, political campaigns, commentary, and satire.[88] California's law, for example, permits unauthorized uses of an individual's voice or likeness "in connection with any news, public affairs, or sports broadcast or account, or any political campaign."[89] Some carveouts also cover broader categories of expressive works, such as in Ohio, which exempts a "literary work, dramatic work, fictional work, historical work, audiovisual work, or musical work regardless of the media in which the work appears or is transmitted."[90]

Others, however, are silent on this issue.[91] When interpreting the common law, or a statute without an express exemption, courts analyze the extent to which the claim at hand implicates First Amendment rights.[92]

Jurisdiction and Remedies. State right of publicity statutes apply varying jurisdictional requirements. Some restrict the law's protections to those domiciled in the state; others are

---

[85] ARK. CODE ANN. § 4-75-1110(a)(1)(F) (2024).

[86] *See Jane Doe No. 1 v. Backpage.com, LLC*, 817 F.3d 12, 27–28 (1st Cir. 2016) (upholding dismissal of statutory misappropriation claims against a classifieds website for images appearing in an advertisement, as it is a "mere[] conduit" and does not benefit from the appropriation); *Almeida v. Amazon.com, Inc.*, 456 F.3d 1316, 1326 (11th Cir. 2006) (highlighting that Amazon did not make "editorial choices" when displaying a book cover that included an unauthorized image on the book's sales page, and that the display was incidental to its role as an internet bookseller).

[87] JENNIFER E. ROTHMAN, THE RIGHT OF PUBLICITY 145, 147 (2018) ("At least five balancing approaches have been applied to evaluate First Amendment defenses in right of publicity cases. . . . This panoply of tests used to determine whether the First Amendment allows and protects uses of a person's identity has led to bizarre and conflicting outcomes in cases with similar facts.").

[88] *See, e.g.*, CAL. CIV. CODE § 3344 (West 2024); ARK. CODE ANN. § 4-75-1110 (2024); N.Y. CIV. RIGHTS LAW § 50-f(2)(d) (McKinney 2024); LA. STAT. ANN. § 51:470.5 (2024); NEV. REV. STAT. § 597.790 (2023).

[89] CAL. CIV. CODE § 3344(d) (West 2024).

[90] OHIO REV. CODE ANN. § 2741.09(A)(1)(a) (West 2024).

[91] *See, e.g.*, KY. REV. STAT. ANN. § 391.170 (West 2024); UTAH CODE ANN. § 45-3-3 (West 2024); VA. CODE ANN. § 8.01-40 (2024).

[92] *See, e.g.*, *Daly v. Viacom*, 238 F. Supp. 2d 1118, 1123 (N.D. Cal. 2002) (the First Amendment protected the use of the plaintiff's likeness in advertisements for a television show in which the plaintiff appeared; the advertisement was found to be an expressive work); *Hoffman v. Capital Cities/ABC, Inc.*, 255 F.3d 1180, 1186–87 (9th Cir. 2001) (right of publicity claims targeting noncommercial uses of an individual's name or likeness may receive heightened First Amendment scrutiny).

more generous.[93] Remedies available to a successful plaintiff also vary across states, although all provide for some form of both injunctive and monetary relief.[94]

c) New State Regulation of Digital Replicas

In response to the emergence of AI-created digital replicas, a number of states have taken steps to either amend existing right of publicity statutes or adopt new laws.[95] Tennessee, for example, recently extended its right of publicity statute to encompass voice simulations.[96] It also expanded the law's scope beyond solely commercial conduct (*i.e.*, "purposes of advertising") to include all acts of unauthorized publishing, performing, distributing, transmitting, or making available to the public.[97]

Two other states, Louisiana and New York, recently passed laws targeting the use of digital replicas.[98] The Louisiana statute applies only to living, professional performers, and prohibits the use of their digital replicas "in a public performance of a scripted audiovisual work, or in a live performance of a dramatic work, if the use is intended to create, and creates, the clear impression that the professional performer is actually performing in the role of a fictional character."[99] New York's amendment of its existing right of publicity prohibits unauthorized digital replicas of deceased professional performers "in a scripted audiovisual

---

[93] *Compare* WASH. REV. CODE § 63.60.010 (2024) (applying to "all individuals and personalities, living and deceased, regardless of place of domicile or place of domicile at time of death"), *with* OHIO REV. CODE ANN. § 2741.03 (West 2024) (limiting to individuals whose domicile or residence is or was in the state).

[94] *See, e.g.*, ARK. CODE ANN. § 4-75-1109 (2024) (providing for injunctive relief and recovery of monetary damages and profits); OHIO REV. CODE ANN. § 2741.07 (West 2024) (providing for injunctive relief and recovery of actual damages including profits, statutory damages, punitive damages, attorney's fees, and treble damages).

[95] In 2023, state legislators introduced 191 AI-related bills, 37 of them addressing deepfakes. *2023 State AI Legislation Summary*, BSA|THE SOFTWARE ALLIANCE (2023), https://www.bsa.org/files/policy-filings/09222023statelegai.pdf. Six deepfake bills were passed targeting nonconsensual deepfake porn and use of deepfakes in politics. *Id. See, e.g.*, S.D. CODIFIED LAWS § 22-24A-2 (2024); 2024 Utah Laws Chs. 127 (H.B. 148), 146 (S.B. 66), 142 (H.B. 238); UTAH CODE ANN. § 20A-11-1104 (West 2024). New Mexico updated its elections laws in 2024 to require a disclaimer of any "materially deceptive media" generated by AI in the context of certain campaign advertisements, and claims may be brought by the Attorney General, a district attorney, falsely represented individual, candidate, or any organization that represents the interests of potentially deceived voters. 2024 N.M. Laws Ch. 57 (H.B. 182).

[96] Ensuring Likeness, Voice, and Image Security Act of 2024, Tenn. Pub. Acts ch. 588.

[97] *Id.*

[98] Other states seem likely to follow suit. In California, for example, a pending bill would amend the postmortem right of publicity statute to establish liability for the production, distribution, or making available of a deceased personality's digital replica. Assemb. B. 1836, Reg. Sess. (Cal. 2024).

[99] LA. STAT. ANN. § 51:470.4(C) (2024).

work as a fictional character or for the live performance of a musical work" if the use is likely to deceive the public.[100]

Both laws incorporate categorical exemptions to accommodate free speech concerns. New York specifies a list of uses excluded from protection:

> [I]f the work is of parody, satire, commentary, or criticism; works of political or newsworthy value, or similar works, such as documentaries, docudramas, or historical or biographical works, regardless of the degree of fictionalization; a representation of a deceased performer as himself or herself, regardless of the degree of fictionalization, except in a live performance of a musical work; de minimis or incidental; or an advertisement or commercial announcement for any of the foregoing works.[101]

Louisiana's law similarly exempts several categories of uses: those made in connection with "a news, public affairs, sports transmission or account, or political campaign," in works of "political, public interest, educational, or newsworthy value," and in "a play, book, magazine, newspaper, literary work, musical composition, single and original work of art or photograph, or visual work."[102] The law also exempts "a sound recording, audiovisual work, motion picture, or radio or television program," but not if they include unauthorized digital replicas that substitute for a professional performer who did not actually appear in the work.[103]

2. Federal Law

While no federal statute focuses solely on the use of an individual's image, likeness, or voice, several serve to limit the creation or use of digital replicas in particular circumstances. We outline below the most relevant laws and regulatory schemes: the Copyright Act, the Federal Trade Commission Act, the Lanham Act, and the Communications Act. In those areas where the Copyright Office does not have special expertise, we summarize the descriptions by the expert agency or commenters of the statutes and their application to digital replicas.[104]

---

[100] N.Y. CIV. RIGHTS LAW § 50-f(1)(a)–(b), (2)(b) (McKinney 2024).

[101] N.Y. CIV. RIGHTS LAW § 50-f(2)(d)(ii) (McKinney 2024).

[102] LA. STAT. ANN. § 51:470.5(B) (2024).

[103] *Id.*; *id.* § 51:470.2(11) (defining "performance" as "the use of a digital replica to substitute for a performance by a professional performer in a work in which the professional performer did not actually appear"). The statute also provides that it "does not affect rights and privileges recognized under other state or federal laws, including those privileges afforded under the 'fair use' factors in the United States Copyright Act of 1976." *Id.* § 51:470.5(A).

[104] Other agencies are also working on aspects of digital replica issues. For example, the Federal Elections Commission ("FEC") has sought public comments on amending 11 C.F.R. section 110.16 to clarify that candidates

a) Copyright Act

Copyright protects original works of authorship, including the material—photographs or audio or video recordings—from which a digital replica might be constructed.[105] The Copyright Act provides copyright owners with a bundle of exclusive rights, including the rights to reproduce a work and to prepare derivative works.[106]

Digital replicas that are produced by ingesting copies of preexisting copyrighted works, or by altering them—such as superimposing someone's face onto an audiovisual work or simulating their voice singing the lyrics of a musical work—may implicate those exclusive rights.[107] If the depicted individual is an owner of the copyrighted work, he or she could have a copyright claim for infringement of the work as a whole. Copyright does not, however, protect an individual's identity in itself, even when incorporated into a work of authorship.[108] A replica of their image or voice alone would not constitute copyright infringement.

b) Federal Trade Commission Act

The Federal Trade Commission Act prohibits "[u]nfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce."[109] FTC

---

and their agents may not use deliberately deceptive AI in campaign ads. *Comments sought on amending regulation to include deliberately deceptive Artificial Intelligence in campaign ads*, FEC (Aug. 16, 2023), https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads/.

[105] *See* 17 U.S.C. § 102.

[106] *Id.* § 106. Under the Copyright Act, a "derivative work" is a work "based upon one or more preexisting works" in which the original is "recast, transformed, or adapted." *Id.* § 101. Examples of derivative works in the Act's definition include, but are not limited to, "a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation." *Id.*

[107] The Office will address the legal issues involved in use of copyrighted works in AI systems in a subsequent Part of this Report.

[108] *See Downing v. Abercrombie & Fitch*, 265 F.3d 994, 1004 (9th Cir. 2001) ("A person's name or likeness is not a work of authorship within the meaning of 17 U.S.C. § 102."); *Midler v. Ford Motor Co.*, 849 F.2d 460, 462 (9th Cir. 1988) ("A voice is not copyrightable. The sounds are not 'fixed.' What is put forward as protectible here is more personal than any work of authorship."); *see also* 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 5:41 (2d ed. 2024) ("While a recorded aspect of these features, such as a facial photograph or a video, is subject to protection under federal copyright law, the human identity that they identify is not protected by copyright.").

[109] 15 U.S.C. § 45(a)(1). Many states have unfair competition laws that target similar business practices and prohibit deceptive or misleading conduct in commercial activities. *See, e.g.*, CAL. BUS. & PROF. CODE § 17200 (West 2024) (providing that "unfair competition shall mean and include any unlawful, unfair or fraudulent business act or practice and unfair, deceptive, untrue or misleading advertising"); MASS. GEN. LAWS ch. 93A, § 11 (2024) (providing a

rules against deceptive advertising and unfair trade practices encompass the misleading use of individuals' identities.

The FTC submitted comments in response to the Office's NOI. It explained that it is empowered to protect the public against deceptive and unfair uses of AI technologies that harm competition, and "there is no AI exemption from the laws on the books."[110] According to the FTC, the use of a digital replica that mimics an individual's voice and likeness might qualify as an unfair method of competition or an unfair or deceptive practice, particularly if it "deceives consumers, exploits a creator's reputation or diminishes the value of her existing or future works, reveals private information, or otherwise causes substantial injury to consumers."[111]

The FTC is also exploring issues related to digital replicas in its ongoing rulemaking to amend its Rule on Impersonation of Government and Businesses.[112] Concurrent with the promulgation of the Impersonation Rule,[113] it issued a supplemental notice requesting comments on the Rule's expansion to prohibit the impersonation of individuals and to extend liability to parties who provide goods and services with knowledge or reason to know that they will be used in impersonations that are unlawful under the Impersonation Rule.[114] The proposed prohibition is meant to address misrepresentations that the person is, or is affiliated with, the impersonated individual, including those that use "identifying information, or insignia or likeness of an individual."[115] Digital replicas, including voice cloning, would be covered.[116]

---

cause of action for those who engage in trade or commerce who suffer loss "as a result of the use or employment by another person who engages in any trade or commerce of an unfair method of competition or an unfair or deceptive act or practice"). Because these laws largely parallel the protections provided by the FTC Act and the Lanham Act, we do not discuss them separately.

[110] FTC Initial Comments at 3–4, 6, 8 ("The FTC is empowered under Section 5 of the FTC Act to protect the public against unfair methods of competition, including when powerful firms unfairly use AI technologies in a manner that tends to harm competitive conditions.").

[111] *Id.* at 4–6.

[112] Trade Regulation Rule on Impersonation of Government and Businesses, 89 Fed. Reg. 15072 (Mar. 1, 2024) ("Impersonation Rule"), https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-03793.pdf.

[113] Trade Regulation Rule on Impersonation of Government and Businesses, 89 Fed. Reg. 15017 (Mar. 1, 2024) (to be codified at 16 C.F.R. pt. 461), https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04335.pdf.

[114] Impersonation Rule at 15072. The supplemental notice proposes defining "Individual" in 16 C.F.R. section 461.4 to mean "a person, entity, or party, whether real or fictitious, other than those that constitute a business or government under this Part." This definition may include deceased persons.

[115] *Id.* at 15077.

[116] *Id.* at 15082 n.98 ("[T]he use of voice cloning for purposes of impersonation is covered where its use satisfies the Rule's prohibitions. Audio deepfakes, including voice cloning, are generated, edited, or synthesized by artificial intelligence, or 'AI,' to create fake audio that seems real.").

In a statement accompanying the final rule on government and business impersonation and describing the supplemental notice, FTC Chair Lina Khan, joined by Commissioners Rebecca Kelly Slaughter and Alvaro Bedoya, highlighted the proliferation of AI-enabled fraud, such as voice cloning used to impersonate individuals regardless of whether they are celebrities.[117] This statement noted that the extension of "means and instrumentalities" liability could apply to persons or entities, including AI developers, "who knew or should have known that their AI software tool designed to generate deepfakes of IRS officials would be used by scammers to deceive people about whether they paid their taxes."[118]

c) Lanham Act

The Lanham Act is the federal trademark law and addresses certain acts of unfair competition. It prohibits deceptive and misleading uses of marks and unfair competition, and fraud and deception in commerce, among other things.[119] Several commenters noted that third-party uses of a digital replica without authorization could constitute false endorsement under the Lanham Act.[120] They cited cases where Lanham Act claims were successful based on unauthorized uses of aspects of plaintiffs' identities, such as soundalikes and lookalikes in advertising.[121] In some circumstances a celebrity or performer may be able to demonstrate that

---

[117] 89 Fed. Reg. 15017, 15030–31 (to be codified at 16 C.F.R. pt. 461) (statement of Chair Lina M. Khan joined by Comm'r Rebecca Kelly Slaughter and Comm'r Alvaro M. Bedoya).

[118] *Id.* at 15031 (to be codified at 16 C.F.R. pt. 461) (statement of Chair Lina M. Khan joined by Comm'r Rebecca Kelly Slaughter and Comm'r Alvaro M. Bedoya).

[119] 15 U.S.C. § 1127. In March 2024, the U.S. Patent and Trademark Office held a public symposium on intellectual property and AI, which included a panel discussion titled "AI, NIL, and the Lanham Act," which addressed the legal and policy considerations related to name, image, and likeness, and the intersection with generative AI. *See Public Symposium on AI and IP*, U.S. PAT. AND TRADEMARK OFF., https://www.uspto.gov/about-us/events/public-symposium-ai-and-ip (last visited July 21, 2024).

[120] Section 43(a) of the Lanham Act establishes liability for using in commerce "any word, term, name, symbol, or device, or any combination thereof, or any false designation of origin, false or misleading description of fact, or false or misleading representation of fact, which[] is likely to cause confusion, or to cause mistake, or to deceive as to the affiliation, connection, or association of such person with another person, or as to the origin, sponsorship, or approval of his or her goods, services, or commercial activities by another person," or which "in commercial advertising or promotion, misrepresents the nature, characteristics, qualities, or geographic origin of his or her or another person's goods, services, or commercial activities." 15 U.S.C. § 1125(a)(1). For comments related to the Lanham Act, see, *e.g.*, International Trademark Association ("INTA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 30, 2023) ("INTA Initial Comments"); Jennifer Rothman Initial Comments at 3; American Intellectual Property Law Association, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 15 (Oct. 30, 2023); SAG-AFTRA Initial Comments at 5; Kernochan Center for Law, Media and the Arts, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 16 (Oct. 30, 2023).

[121] *E.g.*, UMG Initial Comments at 93 (citing *Waits v. Frito-Lay, Inc.*, 978 F.2d 1093 (9th Cir. 1992)).

their voice or a particular pose has achieved trademark status as the basis for a successful claim.[122]

Both false endorsement and trademark infringement claims require proof of commercial use and a likelihood of consumer confusion, mistake, or deceit. The Lanham Act specifies that the defendant's use must be "likely to cause confusion, or to cause mistake, or to deceive as to the affiliation, connection, or association of such person with another person, or as to the origin, sponsorship, or approval of his or her goods, services, or commercial activities by another person."[123] It may be difficult for many individuals, including artists and performers, to prove that the challenged conduct is likely to confuse consumers regarding the plaintiff's association with, or approval of, the defendant's commercial activities. And as INTA noted, AI-generated "revenge porn" would likely fall beyond its reach.[124]

d) Communications Act

The Federal Communications Commission ("FCC") has taken action to regulate digital replicas and to authorize state Attorneys General to do the same. In 2023, it published a Notice of Inquiry on the use of AI-generated voice clones in robocall scams targeting consumers.[125] Following this inquiry, pursuant to the Telephone Consumer Protection Act, the FCC unanimously adopted a declaratory ruling "mak[ing] voice cloning technology used in common robocall scams targeting consumers illegal,"[126] and giving state Attorneys General authority to enforce the rule. FCC Chair Jessica Rosenworcel explained, "Bad actors are using AI-generated

---

[122] *See, e.g.*, INTA Initial Comments at 7 (quoting *ETW Corp. v. Jireh Publ'g, Inc.*, 332 F.3d 915, 922 (6th Cir. 2003), for the "general rule" that "a person's image or likeness cannot function as a trademark," unless "a particular photograph was consistently used on specific goods"); Law Office of Seth Polansky LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 37 (Oct. 12, 2023) ("Seth Polansky Initial Comments"); *Presley's Est. v. Russen*, 513 F. Supp. 1339, 1364–65 (D.N.J. 1981) (noting that while the assertion that Elvis's likeness and image serve as a service mark is too broad, that "a picture or illustration of Elvis Presley dressed in one of his characteristic jumpsuits and holding a microphone in a singing pose is likely to be found to function as a service mark," and ultimately finding a likelihood of success on the merits of an infringement claim as to that mark).

[123] 15 U.S.C. § 1125(a).

[124] INTA Initial Comments at 8.

[125] FCC, Notice of Inquiry on Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts, CG Docket No. 23-362 (Nov. 16, 2023), https://docs.fcc.gov/public/attachments/FCC-23-101A1.pdf.

[126] *FCC Makes AI-Generated Voices in Robocalls Illegal*, FCC (Feb. 8, 2024), https://docs.fcc.gov/public/attachments/DOC-400393A1.pdf; *see* FCC, Declaratory Ruling on Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts, CG Docket No. 23-362 (Feb. 2, 2024), https://docs.fcc.gov/public/attachments/FCC-24-17A1.pdf.

voices in unsolicited robocalls to extort vulnerable family members, imitate celebrities, and misinform voters. We're putting the fraudsters behind these robocalls on notice."[127]

3. Private Agreements

Beyond these statutory and common law protections, private contracts can be negotiated to govern the use of individuals' names or likenesses. Performer service agreements, for example, often include terms proscribing whether and how the other contracting party can use the performer's identity. These usually cover use of the performer's name, image, voice, or likeness for the purpose of promoting the works in which the performer appeared.[128] They may be structured to allow only limited uses, for instance through time-limited grants and restrictions to a particular performance or context, or they may grant broad control, such as an assignment in perpetuity on an exclusive basis.[129] With the advent of AI, some agreements now include specific terms for the use of digital replicas. Talent agency WME, for example, described deals for the use of its clients' likenesses and personalities in connection with AI experiences and products.[130]

In the entertainment field, collective bargaining agreements that establish baseline employment terms have begun to include provisions on the treatment of AI-generated replicas. In December 2023, SAG-AFTRA ratified a multi-year agreement with the Alliance of Motion Picture and Television Producers ("AMPTP") that incorporates new provisions related to the

---

[127] *FCC Makes AI-Generated Voices in Robocalls Illegal*, FCC (Feb. 8, 2024), https://docs.fcc.gov/public/attachments/DOC-400393A1.pdf.

[128] 2 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 10:48 (2d ed. 2024) ("When an actor or performer contracts with a producer to perform in a motion picture or record a phonorecord, the actor or performer commonly signs a contract which includes a 'grant of rights' clause. A 'grant of rights' clause typically assigns copyright in the work to the producer and exclusively licenses the producer to use the actor or performer's identity in advertising and promotion of the work.").

[129] *See* JENNIFER E. ROTHMAN, THE RIGHT OF PUBLICITY 120 (2018); *see also id.* at 122 ("Although many of these voluntary assignments are limited in various ways, to particular time periods, or to the context of telecasts, or to a particular photograph, they are often broader—and can be perpetual and cover all uses of a person's identity in any context."). Contract terms made public through litigation provide some examples of the range and breadth of such agreements. *See, e.g.*, *In re Jackson*, 972 F.3d at 31 (recording agreement granting a label term-limited exclusive rights, and non-exclusive rights thereafter, to use the artist's name and likeness for advertising and marketing covered sound recordings and videos); *Lugosi v. Universal Pictures*, 25 Cal. 3d 813, 816 n.2 (Cal. 1979) (grant of rights clause for a film included, in part, the right to use and publicize "the artist's name and likeness, photographic or otherwise, and to recordations and reproductions of the artist's voice and all instrumental, musical and other sound effects produced by the artist hereunder, in connection with the advertising and exploitation of said photoplay" (emphasis omitted)).

[130] WME Initial Comments at 2 ("WME has already worked with its clients to negotiate AI-specific deals . . . [such as] a deal to lend Snoop Dogg's voice to the AI app Artifact, and deals between Meta and WME's clients to lend their likenesses and personalities to a series of AI-powered chatbots.").

creation and use of digital replicas produced by AI.[131] The ratification was the culmination of a months-long strike, in which this was among the issues "at the forefront."[132] The agreement's final terms establish guidelines related to consent, compensation, and exceptions for replicas created outside the scope of employment if the intended use is protected by the First Amendment.[133] First Amendment-protected uses are specified to include those "for purposes of comment, criticism, scholarship, satire or parody, or . . . use in a docudrama, or historical or biographical work."[134] Similar protections have been negotiated for voice actors and recording artists with respect to voice replicas.[135]

As AI technology continues to evolve, tailored private agreements are likely to become more common.[136] It may be unrealistic, however, to expect such agreements to extend to many other industries, particularly outside of the collective bargaining context.

B. The Need for Federal Legislation

The Copyright Office concludes that new federal legislation is urgently needed. As numerous commenters noted, generative AI technology enables the production and dissemination of digital replicas at a speed and scale that calls for a national response.[137]

---

[131] Memorandum of Agreement Between the SAG-AFTRA and the AMPTP 60–76 (2023) ("SAG-AFTRA 2023 Agreement"), https://www.sagaftra.org/files/2023_Theatrical_Television_MOA.pdf.

[132] SAG-AFTRA Initial Comments at 1 ("[M]any of our members have identified AI as their number one issue, more important to them than increases in wages or improvements in other working conditions, because it poses an existential threat to their very livelihoods.").

[133] SAG-AFTRA 2023 Agreement at 60–76.

[134] *Id.* at 67.

[135] In January 2024, SAG-AFTRA entered into an agreement with Replica Studios covering the use of digital voice replicas by video game studios, and in April 2024, its members ratified an agreement with several major record labels regarding the use of voice replicas in sound recordings. Both agreements include provisions regarding consent and compensation. *SAG-AFTRA and Replica Studios Introduce Groundbreaking AI Voice Agreement at CES*, SAG-AFTRA (Jan. 9, 2024), https://www.sagaftra.org/sag-aftra-and-replica-studios-introduce-groundbreaking-ai-voice-agreement-ces; *SAG-AFTRA Members Ratify 2024 Sound Recordings Code Contract*, SAG-AFTRA (Apr. 30, 2024), https://www.sagaftra.org/sag-aftra-members-ratify-2024-sound-recordings-code-contract.

[136] More recently, in June 2024, the International Alliance of Theatrical Stage Employees ("IATSE") and AMPTP reached a tentative agreement covering requests and consent to "scan" employees. *Tentative 2024–2027 Basic Agreement Summary* at 5, IATSE (2024), https://iatse.net/wp-content/uploads/2024/06/2024-SUMMARY-OF-BASIC-AGREEMENT-NEGOTIATIONS_6.28.24-FINAL.pdf.

[137] *See*, *e.g.*, American Society of Composers, Authors and Publishers, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 13 (Oct. 30, 2023) ("[T]he ubiquity and scale of this new technology requires a robust federal law ensuring that creators' rights are adequately protected."); Songwriters of North America ("SONA") et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of

The impact is not limited to a select group of individuals, a particular industry, or a geographic location. And as described below, existing laws fail to provide fully adequate protection.

1. Shortcomings of Existing Laws

State laws are both inconsistent and insufficient in various respects. As described above, some states currently do not provide rights of publicity and privacy,[138] while others only protect certain categories of individuals.[139] Multiple states require a showing that the individual's identity has commercial value.[140] Not all states' laws protect an individual's voice; those that do may limit protection to distinct and well-known voices, to voices with commercial value,[141] or to use of actual voices without consent (rather than a digital replica).[142]

State right of publicity laws typically apply only where the infringement occurs in advertising, on merchandise, or for other commercial purposes.[143] They do not address the harms that can be inflicted by non-commercial uses, including deepfake pornography, which are particularly prevalent in the internet environment.[144] Different jurisdictional requirements create discrepancies as to who may seek relief.[145] Finally, some of these laws incorporate broad exceptions that may go beyond First Amendment requirements and place many unauthorized uses outside their scope.[146] As numerous commenters noted, the result is a patchwork of protections, with the availability of a remedy dependent on where the affected individual lives or where the unauthorized use occurred.

Inquiry at 10–11 (Oct. 30, 2023) ("SONA-MAC-BMAC Joint Initial Comments") ("We feel strongly that new federal legislation is needed for the protection of a person's identification – their image, voice, characterization, and other likenesses."); WME Initial Comments at 5–6; *Artificial Intelligence and Intellectual Property: Part II—Identity in the Age of AI: Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Jennifer Rothman, Professor of Law, University of Pennsylvania, in response to Questions for the Record ("QFRs") from Rep. Darrell Issa) ("Recent advancements in AI . . . highlight some preexisting challenges because the scale of the problem of unauthorized uses of a person's identity has grown.").

[138] 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY § 6:2 (2d ed. 2024).

[139] *See supra* notes 72–74 and accompanying text.

[140] *See, e.g.*, 42 PA. STAT. AND CONS. STAT. ANN. § 8316 (2024); OHIO REV. CODE ANN. §§ 2741.01, 2741.02 (West 2024).

[141] *See, e.g.*, 42 PA. STAT. AND CONS. STAT. ANN. § 8316 (2024); OHIO REV. CODE ANN. §§ 2741.01(A), 2741.02 (West 2024).

[142] *See, e.g., Midler*, 849 F.2d at 463 (concluding that the California right of publicity statute does not encompass voice imitations but holding that, under California common law, "when a distinctive voice of a professional singer is widely known and is deliberately imitated in order to sell a product, the sellers have appropriated what is not theirs and have committed a tort in California.").

[143] *See, e.g.*, 765 ILL. COMP. STAT. ANN. 1075/30 (2024); VA. CODE ANN. § 8.01-40 (2024); FLA. STAT. § 540.08 (2024).

[144] *See supra* Section II.A.1.b.

[145] *See id.*

[146] *See, e.g.*, OHIO REV. CODE ANN. § 2741.09(A)(1)(a) (West 2024); *see also infra* Section II.B.3.f.

Existing federal laws are too narrowly drawn to fully address the harm from today's sophisticated digital replicas. As explained above, the Copyright Act protects original works of authorship but does not prevent the unauthorized duplication of an individual's image or voice alone,[147] and the targeted individual may not be an owner of copyright in the work as a whole.[148]

The Federal Trade Commission Act prohibits unfair or deceptive acts or practices in or affecting commerce.[149] While it can be applied to cases where digital replicas are used in commercially misleading ways, it does not provide comprehensive protection in other circumstances. Similarly, under the Lanham Act, claims such as false endorsement involving a digital replica are limited to unauthorized *commercial* uses, and most federal courts also require a showing of consumer awareness of the depicted individual in order to establish a likelihood of confusion, limiting the Lanham Act's protection to well-known figures and commercial circumstances. It may be difficult for many individuals, including less famous artists and performers, to prove that the challenged conduct is likely to confuse consumers regarding the plaintiff's association with, or approval of, the defendant's commercial activities. And issues like AI-generated "revenge porn" would likely fall beyond its reach.[150]

Nor can federal communications law address all of the issues raised by unauthorized digital replicas. It only provides the FCC with enforcement powers related to its authority over common carrier services, transmissions, and cable services.[151] The agency's efforts to combat robocall scams stem from its authority related to telephony issues and can help with that particular context. It does not offer a comprehensive solution that could extend more broadly to situations where the use and dissemination of digital replicas may be common, but are not under the FCC's enforcement purview, such as websites featuring user-generated content.

2. Congressional Activity

The Office's recommendations here are presented against the backdrop of ongoing congressional activity.[152] Members of Congress have warned that AI-generated digital replicas have the potential to exacerbate problems of copyright infringement, as well as labor

[147] *See* 17 U.S.C. § 102; *see also supra* Section II.A.2.a.

[148] *See* 17 U.S.C. § 102(a); *id.* § 201(a).

[149] *See* 15 U.S.C. § 45(a)(1).

[150] INTA Initial Comments at 8.

[151] *Comcast Corp. v. FCC*, 600 F.3d 642, 645 (D.C. Cir. 2010).

[152] This discussion reflects relevant Congressional activity that has occurred before July 22, 2024.

displacement and election misinformation.[153] At a hearing last year on AI and copyright, Senator Chris Coons inquired "whether changes to our copyright laws or whole new protections like a federal right of publicity may be necessary to strike the right balance between creators' rights and AI's ability to enhance innovation and creativity."[154]

Legislation has been introduced to address unauthorized digital replicas in various contexts, including political advertisements and communications[155] and sexually explicit images.[156] These bills include the Preventing Deepfakes of Intimate Images Act,[157] which would make it a crime to intentionally disclose or threaten to disclose AI-generated intimate digital depictions; the REAL Political Advertisements Act,[158] which would require political advertisements to disclaim the use of AI-generated sounds or images; and the Protect Elections from Deceptive AI Act,[159] which would make it a crime to distribute deceptive AI-generated media relating to federal elections.

---

[153] The Office is aware that name, image, and likeness issues related to college athletes have also received recent congressional attention. These issues differ from those examined in various aspects here and are beyond the scope of this Report. They arise out of the 2021 Supreme Court decision, *National Collegiate Athletic Association v. Alston*, 141 S. Ct. 2141 (2021), which held in part that NCAA limits on college athlete compensation violated antitrust laws. While the *Alston* case was pending, many states enacted laws to recognize and regulate publicity rights for college athletes. *See generally* Maureen A. Weston, *Off the Guardrails: Opportunities and Caveats for Name Image Likeness and the [Student] Athlete Influencer*, 11 TEXAS A&M L. REV. 911 (2024), https://ssrn.com/abstract=4734794.

[154] *Artificial Intelligence and Intellectual Property—Part II: Copyright: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2023) (statement of Sen. Chris Coons, Chair, S. Subcomm. on Intell. Prop.). The House Judiciary Subcommittee on Courts, Intellectual Property, and the Internet and the Senate Judiciary Subcommittee on Intellectual Property both held hearings focused on the misuse of AI technology with respect to the likeness, voice, and other identifying characteristics of individuals. *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary,* 118th Cong. (2024); *Artificial Intelligence and Intellectual Property: Part II—Identity in the Age of AI: Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024); *Oversight of AI: Election Deepfakes: Hearing Before the Subcomm. on Priv., Tech., & the Law of the S. Comm. on the Judiciary*, 118th Cong. (2024); *see also At Senate Judiciary Subcommittee Hearing, Klobuchar Urges Action to Address Deepfakes in Elections*, AMY KLOBUCHAR, SENATOR (Apr. 16, 2024), https://www.klobuchar.senate.gov/public/index.cfm/2024/4/at-senate-judiciary-subcommittee-hearing-klobuchar-urges-action-to-address-deepfakes-in-elections.

[155] REAL Political Advertisements Act, H.R. 3044, 118th Cong. (2023); Candidate Voice Fraud Prohibition Act, H.R. 4611, 118th Cong. (2023); REAL Political Advertisements Act, S. 1596, 118th Cong. (2023); Protect Elections from Deceptive AI Act, S. 2770, 118th Cong. (2023).

[156] Preventing Deepfakes of Intimate Images Act, H.R. 3106, 118th Cong. (2023); Protect Victims of Digital Exploitation and Manipulation Act of 2024, H.R. 7567, 118th Cong. (2024); DEFIANCE Act of 2024, S.3696, 118th Cong. (2024); TAKE IT DOWN Act of 2024, S.4569, 118th Cong. (2024).

[157] Preventing Deepfakes of Intimate Images Act, H.R. 3106, 118th Cong. (2023).

[158] REAL Political Advertisements Act, S. 1596, 118th Cong. (2023); REAL Political Advertisements Act, H.R. 3044, 118th Cong. (2023)

[159] Protect Elections from Deceptive AI Act, S. 2770, 118th Cong. (2023).

To date, two congressional proposals would address the unauthorized use of digital replicas more broadly: the No Artificial Intelligence Fake Replicas And Unauthorized Duplications ("No AI FRAUD") Act,[160] and the discussion draft of the Nurture Originals, Foster Art, and Keep Entertainment Safe ("NO FAKES") Act of 2023.[161] A number of commenters specifically referenced these two proposals and were generally supportive.[162]

a) No AI FRAUD Act

Introduced in early 2024, the No AI FRAUD Act would establish intellectual property rights in voice and likeness[163] and protect against the use of unauthorized digital voice replicas and digital depictions that readily identify an individual.[164] The bill would allow these rights to be transferred during the individual's lifetime and would make them descendible.[165] Rights would endure at least ten years after the death of the individual, even if they had not been used commercially during their lifetime, and would continue until either (a) proof that they had not been used commercially in a two-year period by an executor, transferee, heir, or devisee; or (b) the death of all executors, transferees, heirs, or devisees.[166]

The legislation would require any authorization to use a digital depiction or digital voice replica to be in writing and valid only if the individual is represented by counsel. If the individual is a minor, the agreement must be approved by a court in accordance with state

---

[160] No AI FRAUD Act, H.R. 6943, 118th Cong. (2024).

[161] Sen. Chris Coons et al., NO FAKES Act Discussion Draft (2023), https://www.coons.senate.gov/imo/media/doc/no_fakes_act_draft_text.pdf. The COPIED Act, introduced in July 2024, establishes rules regarding the attachment of content provenance information for synthetic content, such as digital replicas, but does not provide new rights to individuals. *See* The COPIED Act, S. 4674, 118th Cong. (2024).

[162] *See, e.g.*, SAG-AFTRA Initial Comments at 7–8; Sandra Aistars, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 20 (Oct. 30, 2023); Independent Music Publishers International Forum, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2–3 (Oct. 30, 2023); Dina LaPolt, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 9–10 (Oct. 30, 2023) ("Dina LaPolt Initial Comments").

[163] No AI FRAUD Act, H.R. 6943, 118th Cong. § 3(b)(1)–(2) (2024).

[164] A "digital depiction" is a "replica, imitation, or approximation" of an individual's "likeness," which is defined as an actual or simulated image or likeness that is "readily identifiable as the individual." *Id.* § 3(a)(2), (6) (2024). The bill defines "voice" as an actual or simulated voice that is "readily identifiable" as the depicted individual, while a "digital voice replica" is an audio rendering that includes "replications, imitations, or approximations of an individual that the individual did not actually perform." *Id.* § 3(a)(4)–(5).

[165] *Id.* § 3(b)(2)–(3).

[166] *Id.* § 3(b)(2)–(3).

law.[167] Authorization would also be valid if governed by the terms of a collective bargaining agreement.[168]

The bill would impose direct liability for disseminating a digital voice replica or digital depiction with knowledge that it is not authorized,[169] and for trafficking in a "personalized cloning service" designed to produce digital voice replicas or digital depictions of particular individuals.[170] It would establish secondary liability for any person or entity who "materially contributes to, directs, or otherwise facilitates" directly infringing activity with knowledge that the rightsholder has not consented.[171]

To accommodate the First Amendment, the bill provides a list of factors for a court to consider in balancing the public interest against the private digital replica right.[172] This balancing framework is not required, however, if the digital depiction "includes child sexual abuse material, is sexually explicit, or includes intimate images."[173]

Potential remedies include statutory or actual damages, whichever is greater, lost profits, punitive damages, and attorney's fees.[174] The bill categorizes the law as intellectual property for the purposes of Section 230 of the Communications Decency Act.[175] It expressly does not preempt any state or federal laws.[176]

---

[167] *Id.* § 3(b)(4)(A).

[168] *Id.* § 3(b)(4)(B).

[169] Specifically, any person or entity who "publishes, performs, distributes, transmits, or otherwise makes [it] available to the public." *Id.* § 3(c)(1)(B).

[170] *Id.* § 3(a)(3), (c)(1)(A). The bill does not incorporate a knowledge requirement for this violation.

[171] *Id.* § 3(c)(1)(C). The bill does not provide safe harbors, and a disclaimer is not a defense for any infringing activity. *Id.* § 3(c)(2)(D).

[172] *Id.* § 3(d). These factors include whether "(1) the use is commercial; (2) the individual whose voice or likeness is at issue is necessary for and relevant to the primary expressive purpose of the work in which the use appears; and (3) the use competes with or otherwise adversely affects the value of the work of the owner or licensee of the voice or likeness rights at issue." *Id.*

[173] *Id.* § 3(e)(3).

[174] *Id.* § 3(c)(2)(A)–(C).

[175] *Id.* § 3(j); *see infra* Section II.B.3.d.iii.

[176] No AI FRAUD Act, H.R. 6943, 118th Cong. § 3(g) (2024).

b) NO FAKES Act Discussion Draft

The NO FAKES Act discussion draft provides for a right "to authorize the use of the image, voice, or visual likeness of the individual in a digital replica."[177] The right is a descendible and licensable property right that continues for 70 years after the individual's death, even if it is not exploited during their lifetime.[178] Licensing of the right is valid only if the individual is represented by counsel; the agreement is in writing; or the license is governed by a collective bargaining agreement.[179]

The draft bill imposes liability for producing and disseminating a digital replica without consent.[180] It conditions liability on "knowledge that the digital replica was not authorized by the applicable individual or rights holder."[181] The draft includes a list of categorical exclusions from liability, including the use of digital replicas in news, public affairs, or sports broadcasts; in documentary, historical, or biographical works; for comment, criticism, scholarship, satire, or parody; and where the use is *de minimis* or incidental.[182]

Potential remedies include statutory or actual damages, whichever is greater; punitive damages; and attorney's fees.[183] The bill categorizes the law as an intellectual property law for the purposes of Section 230 of the Communications Decency Act.[184] It expressly does not preempt other state or federal laws.[185]

3. The Contours of a New Right

In response to our NOI, the Office received extensive input on the contours of a new digital replica right. After reviewing the comments, existing law, and the current legislative proposals, we have identified the following critical elements: (1) the definition of "digital

---

[177] Sen. Coons et al., NO FAKES Act Discussion Draft § 2(b)(1) (2023), https://www.coons.senate.gov/imo/media/doc/no_fakes_act_draft_text.pdf. The draft bill defines a "digital replica" as a representation that "is [nearly indistinguishable]" from an individual's actual image, voice, or visual likeness, and is fixed in a sound recording or audiovisual work. *Id.* § 2(a)(1).

[178] *Id.* § 2(b)(2)(A).

[179] *Id.* § 2(b)(2)(B).

[180] *Id.* § 2(c)(2).

[181] *Id.* § 2(c)(2)(B). The draft bill does not provide safe harbors for any activity. Displaying a disclaimer or not having participated in the "creation, development, distribution, or dissemination" of the digital replica is not a defense. *Id.* § 2(d)(3).

[182] *Id.* § 2(c)(3).

[183] *Id.* § 2(d)(4).

[184] *Id.* § 2(f); *see infra* Section II.B.3.d.iii.

[185] Sen. Chris Coons et al., NO FAKES Act Discussion Draft § 2(e) (2023), https://www.coons.senate.gov/imo/media/doc/no_fakes_act_draft_text.pdf.

replica;" (2) the persons protected; (3) the term of protection; (4) prohibited acts; (5) secondary liability; (6) licenses and assignments; (7) accommodation of First Amendment concerns; (8) remedies; and (9) interaction with state laws.[186]

a) Subject Matter

In this Report, we have defined "digital replica" as "a video, image, or audio recording that has been digitally created or manipulated to realistically but falsely depict an individual."[187] A new bill will have to include text that precisely prescribes the subject matter it seeks to protect. Although the Office did not receive comments proposing definitions, in our view the new right should not sweep too broadly. As discussed above, state rights of publicity have been interpreted to cover a broad range of imitations or evocations, including catch phrases or caricatures.[188] But the conduct that now demands federal attention—such as voice cloning in music and the creation of a video or image that appears to depict a real person—involves replicas that do not merely evoke an individual but are difficult to distinguish from reality. We recommend that federal law target replicas that convincingly appear to be the actual individual being replicated.

b) Persons Protected

As discussed above, state rights of publicity and related laws vary significantly in the persons that are protected. Some protect only those who can demonstrate that they are famous or that their identities have commercial value.[189]

Multiple commenters advocated for a federal right that extends protection to all individuals regardless of their level of fame or the commercial value of their identities.[190] They

---

[186] There are other issues on which the NOI did not seek comment and which this Report does not discuss in detail. These include the statute of limitations, retroactivity, and whether federal courts should have exclusive jurisdiction.

[187] *See supra* Section I.A.

[188] *See supra* Section II.A.1.

[189] *See, e.g.*, IND. CODE ANN. § 32-36-1-6 (2024) (defining subject of right of publicity protections as a natural person who possesses an attribute (such as name, voice, image, or likeness) that has commercial value); OHIO REV. CODE §§ 2741.01(A), 2741.02 (2024) (protecting "an individual's name, voice, signature, photograph, image, likeness, or distinctive appearance, if any of these aspects have commercial value").

[190] *See, e.g.*, David Newhoff, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Oct. 7, 2023) ("David Newhoff Initial Comments") ("If ROP law is expanded, it should . . . apply to all persons, not just celebrities . . . ."); Walker Wambsganss et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Oct. 26, 2023) ("Walker Wambsganss et al. Initial Comments") ("[M]inimum standards should be set again including but not limited to, . . . obtain[ing] consent of an individual's likeness of AI-generated content regardless of their public or private status . . . ."); Internet Archive, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11–12 (Oct. 30, 2023) ("Internet

pointed out that everyone has a legitimate interest in controlling the use of their likenesses, and harms such as blackmail, bullying, defamation, and use in pornography are not suffered only by celebrities.[191] While a famous performer might be more susceptible to an AI-generated sound recording topping the music charts, any member of the public could be on the receiving end of a robocall imitating a close family member, or the subject of an explicit image used to humiliate them.[192] Protecting all individuals is consistent with the common law right of privacy, which typically requires neither fame nor commercial value.[193]

The Office believes that the goal of enacting a federal digital replica law is to ensure that everyone has adequate protection and recommends that the law cover all individuals.

c) Term of Protection

The appropriate term of protection is the subject of some debate. Should protection continue after death, allowing heirs or assigns to control exploitation of the deceased's voice and image? As discussed above, a number of states provide postmortem protection for rights of publicity, with protections varying in duration and conditions such as continuing commercial exploitation.

The Office received several comments on this issue. Talent agency WME argued in favor of postmortem rights, stating that "[u]nauthorized deepfakes threaten to usurp estates' valid interests in preserving and strengthening artists' legacies through the *legitimate* use of AI" and may detract from the authenticity, credibility, and commercial value of an artist's body of work.[194] Some courts and commentators have reasoned that postmortem rights promote

---

Archive Initial Comments") ("The public deserves the right to exist in online spaces without being constantly surveilled, and the right to not have our likeness used in ways that humiliate, harass, or abuse us."); INTA Initial Comments at 10–11 (Basic standards for a federal right of publicity should include that "[a]n individual claimant need not make commercial use of his or her persona to have a right of publicity.").

[191] *See, e.g.*, Jennifer Rothman Initial Comments at 5 ("Many ordinary people have their names, likenesses, and voices used without their permission in ways that cause significant harm, including reputational and commercial injuries. There should be no requirement to have a commercially valuable identity to bring a claim."); Internet Archive Initial Comments at 12 ("It would be unfair to hand out rights to celebrities that would allow them to control how their image is collected, processed and used, but not have any protections for the general public.").

[192] *See supra* Section I.A.

[193] *See supra* Section II.A.1.a.

[194] WME Initial Comments at 4. A number of other groups and individuals also supported postmortem protection. *See, e.g.*, INTA Initial Comments at 11; Letter from RIAA, Summary of *Ex Parte* Meeting on April 23, 2024 Regarding the Office's AI Study, to U.S. Copyright Office (Apr. 29, 2024); *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of FKA twigs, Singer, Songwriter, Producer, Dancer, and Actor, in response to QFRs from Sen. Thom Tillis) ("[T]he NO FAKES Act[] should also provide for protection of the artist's rights and image after their death in perpetuity.").

investment in the deceased's legacy,[195] protect the value of assignments made before death,[196] prevent exploitation that heirs and assigns find objectionable or offensive,[197] and conform to the treatment of other types of property.[198]

Support for a postmortem right, however, was not unanimous. Others asserted that there is a less compelling government interest in such protection, making its application to expressive content more vulnerable to First Amendment challenge.[199] It has also been argued that postmortem rights offer little value as a motivating force for creative endeavors.[200] Moreover, to the extent that rights in one's image, voice, and likeness are personal in nature and rooted in privacy, the interests protected do not survive death and generally are not descendible

---

[195] *See Martin Luther King Jr., Ctr. for Soc. Change v. Am. Heritage Prods.*, 694 F.2d 674, 682 (11th Cir. 1983) ("If the right of publicity dies with the celebrity, the economic value of the right of publicity during life would be diminished because the celebrity's untimely death would seriously impair, if not destroy, the value of the right of continued commercial use."); *State ex rel. Elvis Presley Int'l Mem'l Found. v. Crowell*, 733 S.W.2d 89, 99 n.11 (Tenn. Ct. App. 1987) (holding that the law should recognize a celebrity's expectation that he or she is creating a valuable capital asset for the benefit of heirs after death). *Cf.* Peter Felcher & Edward Rubin, *The Descendibility of the Right of Publicity: Is There Commercial Life After Death?*, 89 Yale L.J. 1125, 1128–29 (1980).

[196] *See Martin Luther King Jr., Ctr. for Soc. Change*, 694 F.2d at 705.

[197] Marc A. Lieberstein, *Why a Reasonable Right of Publicity Should Survive Death: A Rebuttal* at 9, 10, NYSBA BRIGHT IDEAS (2008) ("Without a post-mortem right of publicity, [Marilyn] Monroe's name or likeness could show up on portable toilets. Such offensive, unauthorized uses of Monroe's persona are a real possibility absent reasonable legislation that would permit the heirs and/or other authorized entities to regulate use of the publicity right after death.").

[198] *See, e.g.*, *State ex rel. Elvis Presley Int'l Mem'l Found.*, 733 S.W.2d at 97–98 ("If a celebrity's right of publicity is treated as an intangible property right in life, it is no less a property right at death.").

[199] *E.g.*, Letter from MPA, Summary of *Ex Parte* Meeting on May 13, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 5 (May 20, 2024).

[200] As the Sixth Circuit put it in *Memphis Development Foundation v. Factors Etc., Inc.*, before Tennessee adopted postmortem rights legislation, "[t]he desire to exploit fame for the commercial advantage of one's heirs is . . . a weak principle of motivation." 616. F.2d 956, 959 (6th Cir. 1980).

to one's heirs.[201] One commenter warned that postmortem rights could reduce investment in living artists by shifting it to "digitally-resurrected" celebrities.[202]

In addition to these policy arguments, postmortem rights can pose practical challenges. For example, identifying the individual or entity that controls these rights may be difficult, depending on how the rights were bequeathed. INTA suggested addressing the uncertainty around the holder of postmortem rights though a non-mandatory registration system that would provide "public notice that such rights are being claimed, and provide contact information for the use of such rights."[203]

Taking into account all of these points, the Office makes the following recommendation: A federal digital replica right should prioritize the protection of the livelihoods of working artists, the dignity of living persons, and the security of the public from fraud and misinformation regarding current events. For these purposes, a postmortem term is not necessary.

At the same time, we recognize that there is a reasonable argument for allowing heirs to control the use of and benefit from a deceased individual's persona that had commercial value at the time of death. If postmortem rights are provided in a new federal law, we would recommend an initial term shorter than twenty years, perhaps with the option of extending it if

---

[201] This point is consistent with the approach taken in the Visual Artists Rights Act, which protects the moral rights of artists to control their work and reputation via copyright law only during the life of the author. 17 U.S.C. § 106A(d)(1). *See Lugosi v. Universal Pictures*, 603 P.2d 425, 431 (Cal. 1979) ("We hold that the right to exploit name and likeness is personal to the artist and must be exercised, if at all, by him during his lifetime."); *see also The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Benjamin Sheffner, Senior Vice President & Associate General Counsel, MPA) ("Any interest in a performer's reputation or dignity is already governed by defamation and privacy law, which is personal to the individual at issue. But recognizing dignitary interests of deceased individuals, and giving heirs or corporate successors the ability to sue over them, would represent a radical change in centuries of American law, under which 'there can be no defamation of the dead.'" (quoting RESTATEMENT (SECOND) OF TORTS § 560 (AM. L. INST. 1977))).

[202] Jennifer Rothman Initial Comments at 6 ("A federal postmortem right may shore up the replacement of up-and-coming performers with long-dead celebrities."). *Cf.* Mark Bartholomew, *A Right to Be Left Dead*, CALIF. L. REV. (forthcoming 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4610679 ("Because dead celebrities no longer have the capacity to make unpredictable choices in their personal lives that can jeopardize their corporate sponsor's relationship with the public, they represent a more stable investment than their living counterparts. . . . Dead stars also come cheaper as company spokespersons than live ones—an obvious point in their favor when constructing a marketing campaign on a limited budget.").

[203] INTA Initial Comments at 10–11 ("Where practicable, a non-mandatory post-mortem registration system would assist . . . in proving public notice that such rights are being claimed, and provide contact information for the use of such rights. . . . There could be incentives to register the claim of rights, such as reserving the ability to obtain monetary relief to only those valid rights holders who registered their claim prior to the commencement of the unauthorized use.").

the persona continues to be commercially exploited.[204] This approach would not pose the same burden to free expression interests or raise as many practical challenges as a long-term or perpetual right.[205] We note that to the extent the federal law is not fully preemptive, as discussed below, states could still offer a longer term.

d) Infringing Acts

The Office received relatively few comments addressing the scope of the conduct to be prohibited and the allocation of liability.

Regarding the baseline acts that a law should cover, we recommend proscribing activities that involve dissemination to the public—in copyright terms, the acts of distribution, publication, public performance, display, or making available. In our view, this is the type of conduct likely to cause harm to the individual whose image or voice is being replicated.

In contrast, the creation of a digital replica in itself could be part of an artist's experimental process or for a consumer's personal entertainment. Such purely personal use would ordinarily be innocuous and can foster further creativity.[206] If Congress were to impose liability for the mere act of creation, it would be advisable to include a defense for legitimate and reasonable private uses. This does not mean, however, that the act of creation could not be the basis for liability where it is a knowing part of a broader distribution scheme[207] or violates other laws.[208]

---

[204] *Cf.* 15 U.S.C. §§ 1058–59 (providing for successive renewals of trademark registrations so long as there is continued use in commerce). The longer term of copyright protection, in contrast, is intended to incentivize the creation of original works in order to promote progress.

[205] Even a short postmortem term could benefit from something like the voluntary registry INTA proposes in order to clarify the status of postmortem rights and facilitate licensing. Among states with right of publicity laws, California, Oklahoma, and Texas all have registration regimes through which descendants of rightsholders may publicly register their rights with the government. CAL. CIV. CODE § 3344.1 (West 2024) (registration is required in order to recover damages); OKLA. STAT. ANN. tit. 12, § 1448(F)(2) (2024) (same); TEX. PROP. CODE ANN. § 26.006 (West 2023) (registration is prima facie evidence of a valid claim to a property right).

[206] In the copyright context, see, *e.g.*, *Chapman v. Maraj*, No. 2:18-cv-9088, 2020 WL 6260021, at *10 (C.D. Cal. Sep. 16, 2020) (holding that Nicki Minaj's unauthorized creation of a derivative work based on a Tracy Chapman song for experimentation was a fair use).

[207] In these circumstances, principles of secondary liability could apply. *See infra* Section II.B.3.d.iii.

[208] For example, digital replicas used to create child sexual abuse material ("CSAM") or nonconsensual pornography would still be subject to criminal law penalties. *See, e.g.*, 18 U.S.C. § 2251(a) (criminalizing, among other acts, using a minor to produce CSAM with materials transported interstate, such as a computer); *United States v. Tatum*, No. 3:22-cr-157, 2023 WL 3185795, at *2 (W.D.N.C. May 1, 2023) (involving an indictment alleging, in part, production of sexually explicit content under 18 U.S.C. § 2251(a) for using a website to produce deepfake nude images); TEX. PENAL CODE ANN. § 21.165(b) (West 2023) ("A person commits an offense if, without the effective consent of the person appearing to be depicted, the person knowingly produces or distributes by electronic means a deep fake video that appears to depict the person with the person's intimate parts exposed or engaged in sexual conduct.").

(i) Commercial Nature of Use

As discussed above, state rights of publicity typically cover only commercial uses.[209] While some commenters suggested that any federal right be similarly limited,[210] others urged that it should cover both commercial and non-commercial uses,[211] noting the range of harms that can arise from unauthorized replicas.[212] As the Brooklyn Law Incubator & Policy Clinic ("BLIP") and National Public Radio ("NPR") pointed out, the creators of deepfakes do not always act for financial gain,[213] and deception can be harmful regardless of commercialization.[214] Moreover, distinguishing between commercial and non-commercial contexts can be challenging, especially in today's online environment. For example, social media posts that appear to be an individual's personal expression may be part of a paid influencer campaign in support of a commercial interest.[215]

The Office agrees that harmful uses of digital replicas are not limited to those that are commercial in nature. In fact, the commercial use requirement in many state laws is frequently cited as a major reason why these laws are unable to provide adequate protection.[216] We recommend that a federal digital replica law should encompass both non-commercial and

---

[209] *See supra* Section II.A.1.

[210] *See, e.g.*, INTA Initial Comments at 10 ("To be actionable, the use at issue should be for commercial purposes, and a direct connection between the use and the commercial purpose must exist."); Law Office of Seth Polansky Initial Comment at 37.

[211] *See, e.g.*, BLIP, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 32 (Oct. 30, 2023) ("BLIP Initial Comments") ("[C]onsidering the breadth of content on which AI systems can be trained, when it comes to infringement of the right of publicity through the use of AI systems, commercial use should not be a required element."); David Newhoff Initial Comments at 5.

[212] *E.g.*, Artist Rights Alliance & Future of Music Coalition Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Dec. 6, 2023) ("[T]he harms go deeper still. Just last week, reports surfaced of a generative AI engine used to create non-consensual pornographic images, including those depicting 'several multiplatinum Grammy Award-winning singer-songwriters and Academy Award-winning actresses' among others. And reports of other forms of deepfake harms such as cyber bullying, impersonation scams, and revenge porn are well-known.").

[213] *E.g.*, BLIP Initial Comments at 32 ("[S]ometimes the end-users do not produce [deepfakes] for commercial purposes, but for creativity or maliciousness.").

[214] *E.g.*, NPR, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 9 (Oct. 30, 2023) ("NPR Initial Comments").

[215] *See* Stacey M. Lantagne, *Famous on the Internet: The Spectrum of Internet Memes and the Legal Challenge of Evolving Methods of Communication*, 52 U. RICH. L. REV. 387, 416–17 (2018) ("[C]ommercial use on the internet—especially on social media—can be a complicated question. . . . In fact, everything on social media is advertising at some level—a level that has become increasingly difficult to determine." (footnotes omitted)).

[216] *See, e.g.*, WME Initial Comments at 5 ("Here too, there is often limited recourse. Right-of-publicity laws are generally limited to commercial uses, leaving it unclear whether they apply to fan-generated deepfakes that were not created for profit or for commercial distribution.").

commercial uses. In this respect, the law would incorporate aspects of the right of privacy, which typically guards against non-economic damage.[217]

(ii) Knowledge Standard

Although the Office did not receive many comments on this issue, we recommend adoption of an actual knowledge standard for direct liability.

Under the actual knowledge standard, liability would attach only where the distributor, publisher, or displayer acted with actual knowledge both that the representation in question was a digital replica of a real person, and that it was unauthorized. An objective or "should have known" standard might ensnare unsuspecting or technologically unsophisticated defendants. Given the volume of potential outputs produced by current technologies, and the number of individuals who could be targeted, there are likely to be cases where a user passes along an image or audio recording without realizing that it is a replica of someone's voice or likeness.[218] And even where the user recognizes the subject of a digital replica, they may not be aware that the replica is inauthentic or unauthorized.

Some commenters proposed the stricter standard of intent to deceive. NPR, for instance, argued that "Congress should adopt a narrow law that creates liability in instances where someone's name, image, likeness, or voice is used with intent to deceive the audience to believe that false, faked, or AI-generated content is true or represents actual facts or event[s]. . . . Liability would turn on intentional deceptiveness—that something is fake but is trying to persuade someone otherwise."[219] The Office notes, however, that there may be no intent to deceive in some situations where liability should attach, such as where an unauthorized digital replica is used to harass or ridicule an individual, or to profit from the replica of a popular

[217] *See supra* Section II.A.1.a.

[218] *E.g.*, Public Knowledge, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 20 (Oct. 30, 2023) ("[A]ny universally-available right needs to adequately address the 'digital doppelganger' problem—namely, ways of dealing with situations in which an AI-generated work, by pure mathematical chance, looks or sounds like an otherwise unknown individual. Such instances should not give rise to liability, or trigger a rabbit hole of provenance questions about the training data of the GAI system that generated the accidental lookalike.").

[219] NPR Initial Comments at 9–10.

performer's voice or image.[220] Proof of subjective intent is also a high barrier to meet when seeking to prevent damaging distributions of unauthorized replicas.[221]

(iii) Secondary Liability

Since digital replicas are generally distributed and displayed online through the services of various intermediaries, the treatment of secondary liability will be an important element of a new federal law.

Traditional secondary liability principles from copyright law may be drawn on here. Pursuant to these principles, a defendant may be contributorily liable if it "with knowledge of the infringing activity, induces, causes or materially contributes to the infringing conduct of another."[222] Vicarious liability may be found if the defendant "profits directly from the infringement and has a right and ability to supervise the direct infringer."[223] And a defendant may be liable for inducing infringement where it "distributes a device with the object of promoting its use to infringe copyright, as shown by clear expression or other affirmative steps taken to foster infringement."[224]

While these principles could apply in a variety of cases,[225] most of the comments on this topic related to online service providers ("OSPs") that transmit, cache, host, or link to user content. In several areas, Congress has provided OSPs with special safe harbors against liability for unlawful conduct by their subscribers. The most far-reaching example is Section 230 of the Communications Decency Act of 1996, which immunizes online platforms from civil liability for

---

[220] In some recent instances where popular artists' voices (both living and deceased) have been cloned without authorization, the rights owners' objections do not appear to depend on whether the unauthorized use was intended to deceive. *See, e.g.*, Vicky Wong & Bonnie McLaren, *Drake: AI Tupac track gone from rapper's Instagram after legal row*, BBC News (Apr. 26, 2024), https://www.bbc.com/news/newsbeat-68904385.

[221] *See* David Crump, *What Does Intent Mean?*, 38 HOFSTRA L. REV. 1059, 1071–72 (2010) ("[I]ntent, of course, cannot be seen directly by witnesses. It eludes all five senses. It is known only to the actor, and even here, only sometimes, because some definitions of intent allow the actor to readily believe that there is no intent, even when there is. . . . [T]he law evaluates intent by what the actor does, which means that the law evaluates intent by circumstantial evidence. At the same time, intent is easily denied or rebutted, even when it exists, and sometimes the denial is accompanied by convincing belief on the part of the actor."); *cf.* Thomas O. Depperschmidt, *Bankruptcy for Gamblers: The Questions of Fraudulent Intent, Dischargeability, and Remedial Policy in Credit Card Cash Advance Cases*, 13 BANKR. DEV. J. 389, 403–04 (1997) (concerning the difficulty of establishing intent in fraud cases).

[222] *Co-Star Grp. Inc. v. LoopNet, Inc.*, 373 F.3d 544, 550 (4th Cir. 2004).

[223] *Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*, 545 U.S. 913, 930 n.9 (2005).

[224] *Id.* at 919, 936–37.

[225] For example, the trafficking in devices tailored to create digital replicas might be addressed through secondary liability. *Cf.* No AI FRAUD Act, H.R. 6943, 118th Cong. § 3(c)(1)(A), (C) (2024) (establishing liability for distributing, and for facilitating the distribution of, services for creating digital replicas).

many types of illegal third-party content.[226] It provides that an OSP shall not be treated as the "publisher or speaker" of content provided by others, and that neither OSPs nor their users shall be liable for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[227]

Section 230 includes a significant carve-out; it does not "limit or expand any law pertaining to intellectual property."[228] Commenters had differing views on whether a federal digital replica law would constitute a "law pertaining to intellectual property" carved out from section 230.[229] Some favored including the law within the scope of immunity in order to support online platforms' ability to make moderation decisions and avoid chilling protected speech.[230] Electronic Frontier Foundation ("EFF"), for example, argued that "Congress should clarify that the right of publicity sounds in privacy and is not 'intellectual property' for purposes of Section 230" because "when platforms must fend off expensive lawsuits to protect user speech, they are likely to cave to censorious demands."[231]

Others asserted that including digital replica protection in the intellectual property carve-out is necessary to incentivize platforms to remove infringing material.[232] Several pointed to experiences under current law with OSPs refusing requests to remove AI-generated content violating state rights of publicity, citing section 230.[233]

---

[226] 47 U.S.C. § 230; *see Hepp v. Facebook*, 14 F.4th 204, 209 (3d Cir. 2021) (Section 230 "encourages [internet] companies to host and moderate third-party content by immunizing them from certain moderation decisions.").

[227] 47 U.S.C. § 230.

[228] 47 U.S.C. § 230(e)(2).

[229] *See, e.g.*, DiMA, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 30, 2023) ("DiMA Initial Comments"); A2IM-Recording Academy-RIAA Joint Reply Comments at 18; Jennifer Rothman Initial Comments at 4–5.

[230] *See, e.g.*, EFF, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023) ("EFF Initial Comments"); *see also* DiMA Initial Comments at 6.

[231] *See, e.g.*, EFF Initial Comments at 7; *see also* DiMA Initial Comments at 6 (DiMA "strongly believes that any such right should not be deemed a form of 'intellectual property . . . .").

[232] *See, e.g.*, UMG Initial Comments at 94; SAG-AFTRA Initial Comments at 7; A2IM-Recording Academy-RIAA Joint Reply Comments at 18.

[233] *See, e.g.*, UMG Initial Comments at 94; *see also* SAG-AFTRA Initial Comments at 7; *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Robert Kyncl, Chief Executive Officer, Warner Music Group, in response to QFRs from Sen. Thom Tillis) ("Some platforms have responded to our requests for removal and some have resisted. . . . Some have argued that under Section 230 . . . they are not required to remove them.").

The Copyright Office believes that exclusion from section 230 is advisable to encourage prompt removal of unauthorized digital replicas from online platforms. In many circumstances, OSPs are best positioned to prevent the continuing harm from the availability of such replicas.[234] For example, the disseminators may be anonymous or unreachable,[235] making it impossible to take direct action against them, either informally or through court action. OSPs should be incentivized to assist in removing the replicas once they know they are unauthorized and protected from liability when they do so.

Open AI advocated for "a form of safe harbor . . . for technology providers that do not induce users to create non-consensual digital replicas and take proactive steps to monitor and mitigate harmful uses."[236] Some commenters proposed a notice-and-takedown framework similar in concept to section 512 of the Digital Millennium Copyright Act.[237] This provision encourages copyright owners and OSPs to cooperate "to detect and deal with copyright infringement"[238] by providing qualifying OSPs with immunity from monetary liability for copyright infringement committed by their users.[239] The safe harbors for hosting or linking to infringing content are conditioned upon (among other things) a requirement that the OSP act expeditiously to remove allegedly infringing content upon receiving a valid notification or otherwise becoming aware of the infringing activity.[240]

A number of commenters suggested a safe harbor that differs from section 512 in various respects. DiMA, for instance, argued that if a digital replica law allows for secondary liability, then a safe harbor ought to provide "complete immunity when a service removes

---

[234] Recently, YouTube announced a program allowing individuals to demand a takedown of image and voice deepfakes. Dylan Smith, *YouTube Unveils New AI Likeness Protections—Covering Soundalike Audio and More—for 'Uniquely Identifiable' First Parties*, DIGI. MUSIC NEWS (July 2, 2024), https://www.digitalmusicnews.com/2024/07/02/youtube-ai-protections/.

[235] The hurdles rightsholders face to identify infringers was the subject of a number of comments in the Office's Section 512 Study. *See* U.S. COPYRIGHT OFFICE, SECTION 512 OF TITLE 17 164 (2020) ("SECTION 512 REPORT"), https://www.copyright.gov/policy/section512/section-512-full-report.pdf.

[236] *See* Letter from OpenAI, Summary of *Ex Parte* Meeting on May 28, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 4 (June 4, 2024).

[237] *See, e.g.*, *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Lisa P. Ramsey, Professor of Law, University of San Diego School of Law, in response to QFRs from Sen. Tillis).

[238] H.R. REP. NO. 105-551, pt. 2, at 49 (1998).

[239] U.S. COPYRIGHT OFFICE, SECTION 512 OF TITLE 17 13–21 (2020) ("SECTION 512 REPORT"), https://www.copyright.gov/policy/section512/section-512-full-report.pdf.

[240] *See* 17 U.S.C. § 512(c)–(d).

specifically identified content upon notice."[241] Warner Music Group, by contrast, requested a framework that conditions safe harbor on platforms' not only taking content down, but ensuring that it stays down.[242]

The Office agrees that a notice and takedown system, combined with an appropriate safe harbor, could be an effective approach. Such a system need not duplicate every element of section 512. In our 2020 report on section 512, we observed that some of its provisions were not working as Congress had intended;[243] the experience gained in that context could inform the design of a safe harbor here. The Office recommends conditioning its availability on the OSP expeditiously removing the digital replicas when it has actual knowledge or has received a sufficiently reliable notification that the replica is infringing. We would not, however, import the DMCA's "red flag" knowledge standard, given its interpretive problems in the copyright context.[244]

e) Licensing and Assignment

A digital replica law should address whether rights can be transferred either by assignment or licensing.[245] While the Office did not receive many comments on these issues, they were discussed in congressional hearings, and we have considered that testimony as part of our analysis.[246]

---

[241] *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Graham Davies, President & Chief Executive Officer, DiMA, in response to QFRs from Sen. Thom Tillis).

[242] *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Robert Kyncl, Chief Executive Officer, Warner Music Group, in response to QFRs from Sen. Thom Tillis); *see also id.* (statement of Duncan Crabtree-Ireland, National Executive Director, SAG-AFTRA, in response to QFRs from Sen. Thom Tillis).

[243] *See* SECTION 512 REPORT at 2–6.

[244] *See id.* at 113–20.

[245] The Office uses the terms assignment and licensing here in the same sense as in the copyright context. An assignment is an outright sale or transfer of all rights to another party, who then controls the use and distribution of those rights going forward. A license is a contractual permission, either exclusive or nonexclusive, for the use of a digital replica, which may include limitations such as the duration of the license and the uses licensed.

[246] *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop.*, 118th Cong. (2024) (statement of Robert Kyncl, Chief Executive Officer, Warner Music Group); *id.* (statement of Lisa P. Ramsey, Professor of Law, University of San Diego School of Law); *Artificial Intelligence and Intellectual Property: Part II—Identity in the Age of AI: Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Dana Rao, Executive Vice President, General Counsel, & Chief Trust Officer, Adobe, Inc.); *Artificial Intelligence and Intellectual Property: Part II—Identity in the Age of AI: Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Jeff Harleston, General Counsel & Executive Vice President, UMG).

Commenters in favor of assignability asserted that a digital replica right should be treated no differently than other intellectual property rights, such as copyrights, trademarks, and patents. UMG, for instance, stated that "it is important that . . . as with all forms of intellectual property, [a digital replica] right should be eligible for assignment or licensing either in whole or in part, so that enforcement may be delegated."[247] INTA likewise argued that digital replica "rights should be freely transferable, licensable and descendible property rights."[248]

Others raised concerns about the abuses that could occur if individuals were permitted to fully assign their rights, thereby permanently losing control over how their image is used. Professor Jennifer Rothman stated that "[a]llowing another person or entity to own a living human being's name, likeness, voice, or other indicia of a person's identity in perpetuity poses a significant threat to a person's fundamental rights and liberty, and should be prohibited."[249]

Most commenters favored the ability to license digital replica rights, but with different views on whether there should be any limitations and what the limitations should be. The MPA supported broad freedom to contract, including through licensing.[250] Professor Lisa Ramsey warned, however, that "[i]f digital replica right licenses are not limited in significant ways, this will undermine the objectives of [a digital replica law], which include preventing public deception and protecting the ability of people to control uses of their identity."[251] Duncan Crabtree-Ireland, National Executive Director, SAG-AFTRA, argued that licensing guardrails are needed and described those negotiated by the guilds in 2023 as "model provisions for protection against abuse."[252]

---

[247] UMG Initial Comments at 95.

[248] INTA Initial Comments at 11.

[249] Jennifer Rothman Initial Comments at 5; *see also The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Lisa P. Ramsey, Professor of Law, University of San Diego School of Law); *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary,* 118th Cong. (2024) (statement of FKA twigs, Singer, Songwriter, Producer, Dancer, and Actor, in response to QFRs from Sen. Thom Tillis) ("[T] he importance of limiting licensing agreements in time subject to reasonable renewable contractual terms and conditions must be made clear and provided for. There should be no suggestion of or opportunity for licensed rights being given in perpetuity.").

[250] *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Benjamin Sheffner, Senior Vice President & Associate General Counsel, MPA, in response to QFRs from Sen. Thom Tillis).

[251] *Id.* (statement of Lisa P. Ramsey, Professor of Law, University of San Diego School of Law, in response to QFRs from Sen. Thom Tillis); *id.* (statement of FKA twigs, Singer, Songwriter, Producer, Dancer, and Actor, in response to QFRs from Sen. Thom Tillis).

[252] *See id.* (statement of Duncan Crabtree-Ireland, National Executive Director, SAG-AFTRA, in response to QFRs from Sen. Thom Tillis).

Having considered these views, the Office recommends that individuals be able to license their images and voices for use in digital replicas but not to fully assign all rights. Licensing can facilitate the creation, distribution, and use of creative works, products, and services. It enables individuals who choose to do so to monetize and profit from their own personas.

At the same time, the Office acknowledges the potential for abuse. Given unequal contracting power or knowledge, particularly in the context of employment or talent contracts, individuals may lose control over their own personas for long periods of time or under broad terms, based on a decision made early in their career. Although assignments are common in other areas of intellectual property, digital replica rights are most appropriately viewed as a hybrid of privacy interests and a form of property. Unlike publicity rights, privacy rights, almost without exception, are waivable or licensable, but cannot be assigned outright.[253] Accordingly, we recommend a ban on outright assignments, and the inclusion of appropriate guardrails for licensing, such as limitations in duration and protection for minors.

(i) Duration

To avoid the effective result of an outright assignment, the Office suggests limiting licenses (other than those collectively bargained) to a relatively short term, such as five or ten years.[254] Parties that wish to continue the licensing arrangement could subsequently renegotiate it—allowing for consideration of changed circumstances, including bargaining power.

In adopting a time limitation, care should be taken not to block the ongoing use of content produced lawfully during the period of the license. In creative industries, an owner of digital replica rights is often not the owner of the copyrighted works that incorporate that replica. Similarly, the licensed digital replica may be incorporated into a product that the licensee has invested in with a reasonable expectation of continued distribution, such as packaging or labeling of consumer goods.[255]

Accordingly, we believe that the lapse of a digital replica license should bar only new uses after the expiration of the license period. In other words, if a singer consents to a record label using a digital replica of her voice for a period of years, when that period ends, the label would be prohibited from making new recordings using a digital replica (absent a subsequent

---

[253] *See* 2 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY §§ 10:1, 10:2 (2d ed. 2024).

[254] Jennifer Rothman Initial Comments at 6; *see, e.g.*, CAL. LAB. CODE § 2855(a) (West 2024) ("a contract to render personal service, other than a contract of apprenticeship . . . may not be enforced against the employee beyond seven years from the commencement of service under it").

[255] For example, a company's trademark or product packaging may include the image of an individual, such as Gerber's use of an image of a baby on its baby food packaging, or General Mills' use of images of athletes on its Wheaties cereal boxes.

agreement). It could, however, continue distribution of recordings that were prepared during the contractual period.[256]

(ii) Informed Consent

Reflecting the personal aspects of a digital replica right and the risks of public confusion, the Office believes that a federal statute should ensure that individuals licensing their rights are doing so with adequate knowledge and full disclosure of the intended uses. SAG-AFTRA's collective bargaining agreement provides an example: it requires contract terms regarding digital replica rights to be "clear and conspicuous" and agreed to in a separate contract or rider, or in some other form that stands out prominently.[257] We also note that other federal and state laws offer examples of similar terms required in contracts where personal rights are at issue.[258]

(iii) Contracts with Minors

A number of commenters emphasized the unequal bargaining power of minors.[259] They proposed restrictions on such contracts, including requiring that licenses involving minors would automatically expire when they reach the age of 18, and be subject to procedural safeguards such as court review and holding income in a trust.[260] The Office agrees that such safeguards are advisable.

---

[256] This approach is similar to how copyright law treats derivative works after termination of a grant, allowing those already prepared to continue to be utilized. *See* 17 U.S.C. §§ 203, 304(c), 304(d) (2022). As the Office has explained, a "derivative work prepared under authority of the grant before its termination may continue to be utilized under the terms of the grant after its termination, but this privilege does not extend to the preparation after the termination of other derivative works based upon the copyrighted work covered by the terminated grant." *See id.* § 203; *Notices of Termination*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/recordation/termination.html (last visited July 21, 2024).

[257] In its FAQs on digital replica rights, SAG-AFTRA explains that "this could mean [a digital replica provision] is in a separate rider or it could be in your contract as long as it clearly stands out, such as in a larger font, ALL CAPS or **bold**. . . . A bolded paragraph granting consent, alone, is not sufficient." *FAQs on AI*, SAG-AFTRA, https://www.sagaftra.org/files/sa_documents/AIFAQs.pdf (last visited July 21, 2024).

[258] For instance, the Age Discrimination in Employment Act requires a range of contractual safeguards for employees entering into certain severance and settlement agreements. *See* Older Workers Benefit Protection Act, Pub. L. No. 101-433, 104 Stat. 978 (codified as amended in scattered sections of 29 U.S.C.A.). Under this statute, agreements must explicitly advise the contracting party to consult an attorney before signing and must be written in plain, clear language that the contracting party can understand and not exaggerate any benefits.

[259] *See, e.g.*, Jennifer Rothman Initial Comments at 6 (stating that "[t]he most egregious licenses are likely to involve those with the least bargaining power," and highlighting children as a particularly vulnerable group currently lacking control over these rights).

[260] *E.g.*, *id.* For example, California requires that a court confirm entertainment or sports personal services contracts involving minors. CAL. FAM. CODE §§ 6750, 6751 (West 2024) ("A contract, otherwise valid, of a type described in

f) First Amendment Concerns

Digital replicas may be used in the context of constitutionally protected speech, including news reporting, artistic works, parody, and political opinion, in ways that may be unauthorized and objectionable. Federal legislation in this area will need to take into account the speech interests protected by the First Amendment.[261]

The First Amendment, however, does not protect all speech equally. While the Supreme Court has acknowledged that "some false statements are inevitable if there is to be an open and vigorous expression of views in public and private conversation, expression the First Amendment seeks to guarantee,"[262] it has permitted restrictions in cases of defamation,[263] fraud,[264] and commercially misleading speech[265]—each of which could be implicated by certain uses of unauthorized digital replicas.

In applying state rights of publicity, courts have acknowledged the tension between an individual's right to control their persona and a third party's free speech rights.[266] However, the outcomes in these cases are not consistent, leading to a lack of predictability. The application of First Amendment principles in right of publicity cases has been described by scholars as "a confusing morass of inconsistent, incomplete, or mutually exclusive approaches, tests, and

---

Section 6750, entered into during minority, cannot be disaffirmed on that ground either during the minority of the person entering into the contract, or at any time thereafter, *if the contract has been approved by the superior court* in any county in which the minor resides or is employed or in which any party to the contract has its principal office in this state for the transaction of business." (emphasis added)).

[261] *See, e.g.*, *The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Benjamin Sheffner, Senior Vice President & Associate General Counsel, MPA) ("[C]reation of a new right that would apply in expressive works raises serious First Amendment concerns and risks interfering with core creative freedoms.").

[262] *United States v. Alvarez*, 567 U.S. 709, 718 (2012).

[263] *United States v. Stevens*, 559 U.S. 460, 468 (2010) (citations omitted). Defamation claims involving public officials and public figures must meet a higher standard, however. *See N.Y. Times v. Sullivan*, 376 U.S. 254 (1964) (defamatory statements about public officials are protected by the First Amendment unless they are made with actual malice); *Curtis Publ'g Co. v. Butts*, 388 U.S. 130, 155 (1967) (applying *Sullivan* to public figures).

[264] *Stevens*, 559 U.S. at 468 (citations omitted).

[265] *Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n of N.Y.*, 447 U.S. 557, 566 (1980) (false or misleading commercial speech falls outside of the First Amendment).

[266] *See, e.g., Zacchini v. Scripps-Howard Broad. Co.*, 433 U.S. 562, 574–75 (1977) (acknowledging the tension between a news broadcaster's First Amendment rights and a performer's right of publicity, holding that First and Fourteenth Amendments "do not immunize the media when they broadcast a performer's entire act without his consent . . . "); *see also Comedy II Productions, Inc. v. Gary Saderup*, 25 Cal.4th 387, 397 (Cal. 2001) ("The right of publicity derived from public prominence does not confer a shield to ward off caricature, parody and satire. Rather, prominence invites creative comment.") (quoting *Guglielmi v. Spelling-Goldberg Prods.*, 603 P.2d 454, 460 (Cal. 1979) (Bird, C.J., concurring)).

standards"[267] or, more succinctly, a "dumpster fire."[268] At least five different balancing tests are in use by courts in different states across the country,[269] producing "conflicting outcomes in cases with similar facts."[270]

Our NOI questions on this issue elicited strong reactions from commenters. Many recognized that a federal law prohibiting unauthorized digital replicas must leave room for First Amendment-protected activity.[271] Commenters, however, disagreed on exactly how a statute should accommodate free speech rights. Some supported specific exceptions similar to those in some state right of publicity statutes, primarily for news reporting, various types of expressive works, and sports broadcasting, as well as parody, comment, and criticism.[272]

---

[267] Gloria Franke, *The Right of Publicity vs. the First Amendment: Will One Test Ever Capture the Starring Role?*, 79 S. CAL. L. REV. 945, 946 (2006).

[268] William McGeveran, *Selfmarks*, 56 HOUS. L. REV. 333, 362 (2018).

[269] Roberta Rosenthal Kwall, *A Perspective on Human Dignity, the First Amendment, and the Right of Publicity*, 50 B.C. L. REV. 1345, 1346 (2009); *see also* Robert C. Post & Jennifer E. Rothman, *The First Amendment and the Right(s) of Publicity*, 130 YALE L. J. 86, 127 n.167 (2020).

[270] JENNIFER E. ROTHMAN, THE RIGHT OF PUBLICITY 145, 147 (2018).

[271] *See, e.g.*, Senator Marsha Blackburn Initial Comments at 3 ("This liability must be balanced, of course, by significant protections for any applicable First Amendment rights."); International Center for Law & Economics, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 26 (Oct. 30, 2023) ("If Congress chose to enact a federal 'right of privacy' statute, several key issues would need to be addressed regarding . . . First Amendment limitations."); Digital Media Licensing Association, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 20–21 (Oct. 30, 2023) ("The right to create content for newsworthy and expressive purposes that are guaranteed under the First Amendment must be considered and balanced with the concerns of the public including actors and other public figures."); Internet Archive Initial Comments at 11–12 ("[G]ranting a new right of publicity along the lines of existing state laws . . . come with serious First Amendment concerns."); Pamela Samuelson et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 37 (Oct. 30, 2023) ("Pamela Samuelson Initial Comments") (advocating for "thoughtful self-regulation in [addressing the issue of deepfakes]," but expressing skepticism about the feasibility of imposing rules due to potential conflicts with the First Amendment).

[272] *See, e.g.*, Getty Images, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 28 (Oct. 30, 2023) ("Getty Images Initial Comments") ("[A]ny new federal right of publicity should be carefully considered so that constitutionally protected expression is not unduly limited. Accordingly, legislation should include explicit exemptions for First Amendment-protected expression."); MPA, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 33 (Dec. 6, 2023) ("MPA Reply Comments") ("At minimum, a bill establishing a federal digital-replica right must include exemptions where the use is in a work of political, public interest, educational, or newsworthy value, including comment, criticism, or parody, or similar works, such as documentaries, docudramas, or historical or biographical works, or a representation of an individual as himself or herself, regardless of the degree of fictionalization, and for uses that are de minimis or incidental."); INTA Initial Comments at 11 (proposing exceptions for "a. News, public affairs and sports reporting or commentary; b. Dramatic, literary, artistic, or musical works, so long as the use has artistic relevance to the work and does not explicitly mislead as to endorsement or approval by the individual; c. Works that parody, criticize, satirize or comment upon the individual; d. Advertising and promotion for (a)-(c); and (e). Any other noncommercial use, including, but not limited to, education and research").

Protection for expressive works was a principal area of focus.[273] The MPA offered several examples of expressive uses as deserving of protection, including documentaries using digital replicas "to re-create scenes from history where no actual footage exists" and late-night comedians "using digital replicas to poke fun at celebrities, politicians, athletes."[274] It asserted that categorical exceptions "are crucial to giving filmmakers clarity so they know what uses are allowed, or not allowed" before they undertake expensive projects, and proposed a list of exceptions, some subject to the caveat that the use was not intended to and did not create a false impression of authenticity.[275] Donaldson Callif Perez LLP stated that "the implementation of a right that does not explicitly exempt expressive works would have immediate negative consequences."[276]

Other commenters argued that categorical exceptions are unnecessary and could undermine effective protection. RIAA asserted that "categorical exclusions for certain speech-oriented uses are not constitutionally required and, in fact, risk overprotection of speech interests at the expense of important publicity interests."[277] Instead, "the First Amendment calls for a case-specific balancing of the right of publicity against whatever First Amendment

---

[273] *See* MPA, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 71 (Oct. 30, 2023) ("MPA Initial Comments") ("The right of publicity does not—and, to be consistent with the First Amendment, may not—regulate uses of or references to individuals' [name, image and likeness] in 'expressive works,' such as books, plays, news articles and broadcasts, songs, and movies and television programs. Such expressive works are non-commercial speech fully protected by the First Amendment, regardless of whether those works are sold for a profit."); David Newhoff Initial Comments at 5 ("If ROP law is expanded, it should . . . not restrict expressive uses of AI-generated likeness for purposes (e.g., biographical films) that fall within the scope of protected speech.").

[274] MPA Reply Comments at 32.

[275] Letter from MPA, Summary of *Ex Parte* Meeting on May 13, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 4 (May 20, 2024) ("Use of a digital replica would not constitute a violation where: 1. the digital replica is used to depict the individual in a documentary, docudrama, or historical or biographical work, or any other representation of the individual as such individual, regardless of the degree of fictionalization, unless use of the digital replica is intended to create, and does create, the false impression that the work is an authentic recording in which the individual participated; 2. the digital replica is used for purposes of a news, public affairs, or sports broadcast or report, or for a purpose that has political, public interest, educational, or newsworthy value, unless use of the digital replica is intended to create, and does create, the false impression that the work is an authentic recording in which the individual participated; 3. the use of the digital replica is for purposes of comment, criticism, scholarship, satire, or parody; 4. the use of the digital replica is de minimis, incidental, or fleeting; 5. the use of the digital replica is addressed by a collective bargaining agreement; or 6. the digital replica is used in an advertisement or commercial announcement for a work that includes the use of the digital replica as described in [1 through 5].").

[276] Donaldson Callif Perez, LLP, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10 (Oct. 30, 2023).

[277] *E.g.*, Letter from RIAA, Summary of *Ex Parte* Meeting on April 23, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (Apr. 29, 2024).

interests may be presented in the given case."[278] Professor Jennifer Rothman suggested that a new law be specific regarding uses that would not be Constitutionally privileged: "Current common areas of agreement are that unauthorized uses of a person's identity are not protected by the First Amendment if a person's 'entire act' or performance is used, or the uses are in commercial products or advertising not related to an authorized underlying work."[279]

Each of these approaches has advantages and disadvantages. Enumerated exceptions provide greater certainty: users can more easily determine from the face of the statute whether the use they are considering is lawful. In addition, the existing bodies of state precedent are available to be drawn upon for purposes of interpretation. But such exceptions may be both over- and under-inclusive depending on the facts.[280] For example, an exception for expressive works, if not appropriately cabined, could render the law toothless against common uses of digital replicas—such as voice clones in music or deepfake pornography. Indeed, many of the problematic uses reported have been in expressive or political contexts, such as the "Fake Drake" song and the President Biden robocalls described above.[281]

The Office stresses the importance of explicitly addressing First Amendment concerns. While acknowledging the benefits of predictability, we believe that in light of the unique and evolving nature of the threat to an individual's identity and reputation, a balancing framework is preferable. Although the potential overbreadth of categorical exceptions can be cabined by conditions like those proposed by MPA, this introduces a high level of complexity. In addition, we note that in today's online environment, traditional categories such as "news" or "public affairs" are often difficult to define.[282] The result may be to exempt conduct that legislators intended to prohibit.

In our view, a balancing framework permits greater flexibility to assess whether a particular unauthorized use is protected by the First Amendment. Rather than checking a box marked "news" or "musical work," courts can assess the full range of factors relevant to the First Amendment analysis. These could include the purpose of the use, including whether it is

---

[278] A2IM-Recording Academy-RIAA Joint Reply Comments at 16–17; *see also id.* at 17 ("[W]e disagree with those commenters who argue that any federal right of an individual to control uses of their voice or likeness must contain express, categorical exclusions for all uses of a certain type, such as unauthorized uses of an individual's voice or likeness in any 'expressive works,' regardless of the particular facts and circumstances of the use.").

[279] Jennifer Rothman Initial Comments at 6.

[280] *See* Joshua Matz, RIGHT OF PUBLICITY AND THE FIRST AMENDMENT (2024), humanartistrycampaign.com/rop-first-amendment; Letter from RIAA, Summary of *Ex Parte* Meeting on April 23, 2024 Regarding the Office's AI Study, to U.S. Copyright Office (Apr. 29, 2024) (endorsing Matz's view).

[281] *See supra* Section I.

[282] Indeed, the meanings of categories like "news" or "public affairs" are themselves contested. *See* Richard L. Hansen, *From Bloggers in Pajamas to The Gateway Pundit: How Government Entities Do and Should Identify Professional Journalists for Access and Protection* (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4776774; Sonja R. West, *Favoring the Press*, 106 CAL. L. REV. 91 (2018).

commercial; its expressive or political nature; the relevance of the digital replica to the purpose of the use; whether the use is intentionally deceptive; whether the replica was labeled; the extent of the harm caused; and the good faith of the user.[283] We believe that this approach would leave room for the types of expressive works that many commenters identified as a priority.[284]

g) Remedies

A federal right must offer effective remedies through which individuals can seek redress. Several commenters proposed a range of appropriate remedies, including monetary and injunctive relief.[285]

The Office agrees that a digital replica law should provide both monetary and injunctive relief. Damages should include compensation for loss of income, damage to reputation, or emotional distress. As INTA commented, "[t]he commercial value of a persona may have an impact on any damage amount claimed in a dispute."[286] Injunctive relief is essential to prevent ongoing unauthorized use of an individual's likeness or to prevent future violations.

We note that some individuals may have difficulty proving actual damages, particularly market-based injuries, in court. To ensure that protection is both accessible and effective for all, the Office recommends inclusion of special damages enabling recovery by those who may not be able to show economic harm or afford the cost of an attorney.[287] As in the copyright system,

---

[283] *See* 1 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY §§ 8:23, 8:71–3 (2d ed. 2024); A2IM-Recording Academy-RIAA Joint Reply Comments at 16–17 ("[T]he First Amendment calls for a case-specific balancing of the right of publicity against whatever First Amendment interests may be presented in the given case.").

[284] Importantly, application of these factors should permit a movie to use unauthorized digital replicas of deceased individuals where those individuals are portrayed in objectively unrealistic, fictionalized contexts. As an example, a remake of the movie *Bill & Ted's Excellent Adventure*, in which the protagonists use a time machine to interact with and transport historical figures (portrayed by digital replicas) into modern times, should not require authorization. *See Synopsis, Bill & Ted's Excellent Adventure*, IMDB, https://www.imdb.com/title/tt0096928/plotsummary/#synopsis (last visited July 21, 2024).

[285] *See, e.g.*, Law Office of Seth Polansky Initial Comments at 37 ("Enforcement mechanisms would also need to be specified, through both civil remedies and criminal penalties. Given the rapid speed at which AI-generated material can be created and distributed, it is crucial for enforcement measures to be timely and effective."); SAG-AFTRA Initial Comments at 7–8 ("Monetary relief . . . might include lost wages and reputational damage. Injunctive relief must also be available, particularly in the context of AI-generated content that might impact one's reputation (such as AI-generated voice or likeness used to sell shoddy merchandise or questionable services).").

[286] INTA Initial Comments at 11.

[287] Jennifer Rothman Initial Comments at 5 ("Any legislation should include statutory damages to protect people who may not otherwise be able to establish market-based injuries. A number of states have included statutory damages in publicity legislation with the express purpose of protecting ordinary people.").

without the potential for statutory damages or attorney's fee awards to the prevailing party, litigation costs may be a barrier to the filing of meritorious claims.

Several commenters urged the inclusion of criminal penalties,[288] particularly in connection with nonconsensual intimate material.[289] Criminal liability would recognize the seriousness of the harm caused by such actions and the need for accountability; appropriate penalties could deter bad actors and provide justice for victims. The Office agrees that there are specific unauthorized uses that should incur criminal liability, including sexual deepfakes and other particularly harmful or abusive imagery. We do not take a position, however, on whether criminal penalties should be included in a federal digital replica right as opposed to stand-alone criminal legislation, such as the bills pending in this Congress.[290]

h) Preemption

An overarching question is whether, and to what extent, a federal digital replica law should preempt state laws.[291] Commenters were divided on this issue, with some urging preemption and others seeking to preserve state flexibility.

The benefit of preempting state laws as they pertain to digital replicas would be to establish a uniform nationwide standard, entirely replacing the patchwork of existing coverage. Commenters who supported this approach asserted that it would provide clarity for creators, businesses, and consumers alike.[292] DiMA, for example, stated that "Congress should ensure that content related matters have consistent standards by preempting state and common laws where doing so would ensure consistency in application and reduce operational challenges and improve the customer experience."[293] MPA also urged preemption to "provide national uniformity."[294]

---

[288] *See, e.g.*, NPR Initial Comments at 10 ("Criminal penalties may be necessary."); Walker Wambsganss et al. Initial Comments at 5.

[289] *E.g.*, Anonymous AI Technical Writer, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 17 (Dec. 6, 2023) ("At minimum the right to not have one's face used in generative AI, with criminal penalties for nonconsensual pornography and child pornography generated with AI.").

[290] For example, the TAKE IT DOWN Act, introduced in the Senate in June of 2024, would criminalize publishing or threatening to publish non-consensual intimate imagery. TAKE IT DOWN Act of 2024, S.4569, 118th Cong. (2024).

[291] *See generally* CONG. RSCH. SERV., ARTIFICIAL INTELLIGENCE PROMPTS RENEWED CONSIDERATION OF A FEDERAL RIGHT OF PUBLICITY (2024), https://crsreports.congress.gov/product/pdf/LSB/LSB11052.

[292] *See, e.g.*, DiMA Initial Comments at 5; INTA Initial Comments at 9; MPA Reply Comments at 34; BLIP Initial Comments at 32.

[293] DiMA Initial Comments at 5.

[294] MPA Reply Comments at 34.

Commenters opposing preemption stressed the desirability of preserving the extensive body of state law precedent developed over many decades.[295] SONA, Black Music Artists Coalition, and MAC stated that "there is nothing that should warrant preemption of laws that have been thoroughly considered and vetted in our state."[296]

Many proposed that federal law should set a "floor," permitting states to offer broader protection than federal law and displacing only those state laws that fell beneath the federal standard.[297] In testimony before Congress, Warner Music Group advocated for this approach.[298] ImageRights International also urged that the federal right "should set a baseline (floor) for protections, allowing states to provide additional protections if they choose."[299] Dina LaPolt, an entertainment lawyer, stated, "It is important that any potential federal right protecting voice and likeness set a 'floor' of fundamental rights such that states can be individualized in their approach to cater to the potentially more stringent wishes of their residents."[300]

---

[295] *See, e.g.*, SAG-AFTRA Initial Comments at 8 (A "federal right should not preempt state right of publicity laws unless it provides individuals greater protections over their name, image, voice, or likeness than existing state law. Further, it is critical that any federal law relating to AI-generated image and/or voice not supersede, whether intentional or inadvertent, existing state law relating to sexually explicit digital replicas."); A2IM-Recording Academy-RIAA Joint Reply Comments at 18 ("[A] federal right should not preempt state law to deprive individuals of rights that have been carefully developed over decades of legislation and litigation."); SONA-MAC-BMAC Joint Initial Comments at 10–11.

[296] SONA-MAC-BMAC Joint Initial Comments at 10–11.

[297] *See, e.g.*, SAG-AFTRA Initial Comments at 8 ("[A]ny federal law should set a floor for state law protections, allowing states to provide greater protection to individuals residing in their state."); Dina LaPolt Initial Comments at 10. Commenters did not use a uniform term for this form of preemption. Some cases and scholars use the term "partial preemption" to describe statutes that create a regulatory floor that state laws can exceed.

[298] *See The NO FAKES Act: Protecting Americans from Unauthorized Digital Replicas: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Robert Kyncl, Chief Executive Officer, Warner Music Group, in response to QFRs from Sen. Thom Tillis) ("[L]imited preemption would be appropriate to the extent of the scope of a new federal right. But state laws that provide enhanced protections and that are supported by decades of helpful jurisprudence regrading protection of voice and likeness should be allowed to stand to the extent they do not conflict with federal law.").

[299] ImageRights International, Inc., Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Dec. 6, 2023).

[300] Dina LaPolt Initial Comments at 10. To the extent that states have developed diverse approaches, this may reflect differences in state industries and interests. For example, California, the domicile of many celebrities, has a retroactive postmortem right that arose in large part due to a ruling denying the descendability of Marilyn Monroe's right of publicity. *See* Laurie Henderson, *Protecting A Celebrity's Legacy: Living in California or New York Becomes the Deciding Factor*, 3 J. BUS. ENTREPRENEURSHIP & L. 165, 171, 177–83 (2009). Similarly, some have observed that Tennessee's potentially perpetual right of publicity may be traceable to the influence of native Tennessean Elvis Presley. Annie T. Christoff, *Long Live the King: The Influence of Elvis Presley on the Right of Publicity in Tennessee*, 41 U. MEM. L. REV. 667, 699 (2011). This is not to say that state variations in right of publicity laws always neatly track local or regional interests. JENNIFER ROTHMAN, THE RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR NEW YORK?, 592–93 (2018) (noting variation in postmortem rights in the tri-state area).

Alternatively, a non-preemptive law would leave existing state protections in place, regardless of whether those state protections exceeded or fell short of the federal protection. This approach would be similar to the Lanham Act, which leaves state trademark and unfair competition laws coexisting with federal protections.[301]

Although there are reasonable arguments for each approach, the Office recommends against preempting state laws for several reasons. Most importantly, as commenters pointed out, extensive state law in this area has developed over many decades, creating settled expectations.[302] Full preemption would reduce existing protections for individuals in states that currently provide broader rights, causing discrepancies between protection for digital replicas and other imitations of their personas. For example, in a state that provides for postmortem rights of publicity, a preemptive federal law without postmortem rights would result in a deceased individual's beneficiaries having longer-lasting rights in the non-digital context. And there may be advantages in preserving policy flexibility at the state level to respond to rapidly changing conditions, without the need to achieve consensus at the federal level.

Moreover, a non-preemptive federal right can achieve some of the benefits of uniformity, but without imposing a one-size-fits-all solution on all states. Although it would not fully harmonize the varied state approaches, it would fill in the gaps by ensuring the availability of effective national protection against unauthorized uses of digital replicas. Everyone, whatever the status of their own state's law, would have recourse to the same federal claim.

Finally, a non-preemptive law has the advantage of greater clarity. Either full or partial preemption raises the specter of extensive litigation over its scope and the question of which state-level protections remain available. This uncertainty could be minimized by specifying that the federal digital replica law supplements rather than preempts a state's protections.

4. Relationship to Section 114(b) of the Copyright Act

The Office's NOI inquired about the relationship between section 114(b) of the Copyright Act and state law protections against unauthorized digital replicas of voices in sound recordings.[303] Section 114(b) states that the copyright owner's reproduction and derivative work rights in a sound recording are limited to uses that appropriate "the actual sounds fixed in the recording," and "do not extend to the making or duplication of another sound recording that consists entirely of an independent fixation of other sounds, even though such sounds imitate or

[301] *See* J. THOMAS MCCARTHY, 3 MCCARTHY ON TRADEMARKS AND UNFAIR COMPETITION §§ 1:18, 22:1 (5th ed. 2024).

[302] *See supra* notes 295–96.

[303] NOI at 59948.

simulate those in the copyrighted sound recording."[304] The provision was intended to clarify that "mere imitation" of a copyrighted sound recording does not constitute infringement.[305]

Because section 114(b) permits imitation or simulation of sounds (including an individual's voice) in the context of a sound recording copyright, some have questioned whether it might preempt state laws that prohibit an unauthorized replica of someone's voice used in a sound recording. Several commenters noted this issue, with one stating that "legislative attention might be necessary to address potential conflicts and gaps in the law" to clarify these relationships.[306]

Only a few courts have analyzed whether section 114(b) preempts a state right of publicity claim based on the imitation of an individual's voice, with varying results. While the Ninth Circuit has twice rejected preemption in cases involving claims for voice misappropriation under California law,[307] a federal court in Michigan concluded a state right of publicity claim was preempted.[308] The uncertainty regarding section 114(b)'s impact appears to be having real-world consequences as state legislatures debate and enact laws with provisions on digital replicas. Presumably to avoid possible inconsistencies with the Copyright Act, both New York and Louisiana included in their recent laws language which mirrors section 114(b) and limits the scope of conduct prohibited by these state laws. These are right of publicity laws that encompass digital replicas in certain instances and not specifically "digital replica laws." Louisiana's statute, for example, exempts from liability "the making or duplication of another recording that consists entirely of an independent fixation of other sounds, even though the

---

[304] 17 U.S.C. § 114(b).

[305] H.R. REP. NO. 94-1476, at 106 (1976).

[306] *See, e.g.*, Rightsify Group LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 13 (Oct. 30, 2023); *see also* UMG Initial Comments at 98 ("[W]hile UMG maintains that rights of publicity as applied to AI-generated soundalikes are not preempted, the potential for disagreement further counsels in favor of a federal right of publicity that will eliminate debate on this issue."); A2IM and RIAA, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 43–44 (Oct. 30, 2023) ("A2IM-RIAA Joint Initial Comments") ("Case law concerning copyright preemption of rights of publicity is still developing, and we are not aware of any cases addressing that issue in the specific context of AI. Whether courts properly recognize the distinction between Section 114(b) and rights of publicity . . . with respect to generative AI is an issue that warrants attention. Legislative clarification is not clearly necessary but may prove to be helpful.").

[307] *Midler*, 849 F.2d at 462 ("Midler does not seek damages for Ford's use of [a licensed song], and thus her claim is not preempted by federal copyright law. Copyright protects 'original works of authorship fixed in any tangible medium of expression.' A voice is not copyrightable. The sounds are not 'fixed.'"); *Waits*, 978 F.2d at 1100, *abrogated on other grounds by Lexmark Int'l, Inc. v. Static Control Components, Inc.*, 572 U.S. 118 (2014) (finding that the elements of Tom Waits's voice misappropriation claim "are 'different in kind' from those in a copyright infringement case challenging the unauthorized use of a song or recording," and that the claim is not preempted by copyright).

[308] *Romantics v. Activision Publ'g, Inc.*, 574 F. Supp. 2d 758, 764 (E.D. Mich. 2008) (finding that the Copyright Act preempted a right of publicity claim based on use of a band's "distinctive sound" in a videogame, where the "distinctive sound" at issue was that of a copyrighted song rather than the band more generally).

sounds imitate or simulate the voice of the professional performer," and New York's law uses similar language.[309]

In the Office's view, these concerns are misplaced, and section 114(b) does not preempt state laws prohibiting unauthorized voice replicas. The Copyright Act does not preempt state laws with respect to "subject matter that does not come within the subject matter of copyright" or "activities violating legal or equitable rights that are not equivalent to any of the exclusive rights within the general scope of copyright."[310] The legislative history of the Act's preemption provision explains that "[t]he evolving common law rights of 'privacy,' 'publicity,' and trade secrets . . . remain unaffected as long as the causes of action contain elements, such as an invasion of personal rights or a breach of trust or confidentiality, that are different in kind from copyright infringement."[311]

The Office believes that digital replica rights in an individual's voice satisfy this test. An individual's voice, unlike a particular sound recording that may capture it, "does not come within the subject matter of copyright."[312] It is the product of biology, nature, environment, and, in the case of performers, training, skill, and talent. It is not an "original work[] of authorship,"[313] or "fixed in any tangible medium of expression . . . from which [it] can be perceived, reproduced, or otherwise communicated."[314]

Copyright and digital replica rights serve different policy goals; they should not be conflated. Section 114(b) shields vocal imitations and other soundalike recordings against claims of copyright infringement. But nothing indicates that Congress intended for this limitation on copyright to deprive individuals of rights in their unique voices, whether under state right of publicity laws or a new federal statute.[315] To avoid unnecessary confusion or carve outs like the limitations in the New York and Louisiana laws discussed above, the Office recommends that Congress clarify in express terms that section 114(b) does not preempt state laws or affect a new federal right protecting an individual's voice.

---

[309] *See* LA. STAT. ANN. § 51:470.2(4) (2024); N.Y. CIV. RIGHTS LAW § 50-f(1)(c) (McKinney 2024) ("A digital replica does not include . . . the making or duplication of another recording that consists entirely of the independent fixation of other sounds, even if such sounds imitate or simulate the voice of the individual.").

[310] 17 U.S.C. § 301(b)(1), (3).

[311] H.R. REP. NO. 94-1476, at 132 (1976).

[312] *See* 17 U.S.C. § 301(b)(1).

[313] *See id.* § 102(a). *Cf.* U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES § 906.7 (3d Ed. 2021) ("Because human authorship is required for copyright protection, the U.S. Copyright Office will not register naturally occurring objects or materials that are discovered in nature.").

[314] 17 U.S.C. § 102(a).

[315] *See* H.R. REP. NO. 94-1476, at 132 (1976).

III. PROTECTION OF ARTISTIC STYLE

The Office received many comments seeking protection against AI "outputs that imitate the artistic style of a human creator."[316] Commenters voiced concern over the ability of an AI system, in response to a text prompt asking for an output "in the style of artist X," to quickly produce a virtually unlimited supply of material evoking the work of a particular author, visual artist, or musician.[317] They asserted that these outputs can harm, and in some cases have already harmed, the market for that creator's works.[318]

For example, the Center for AI and Digital Policy warned that "if AI can replicate [artists'] signature style en masse, it might undermine the market value of their creations, unjustly depriving them of economic benefits."[319] The Authors Guild described "authors, who, after years of developing their unique voice and style, are finding AI appropriating a part of their personality and mimicries of their work being sold in the market."[320] In addition, commenters argued that, while in the past the impact of human imitators was limited by the demands of time and labor, AI systems present a challenge exponentially greater given their speed and scale.[321] An anonymous artist offered the following example:

---

[316] *See, e.g.*, The Authors Guild, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10–12 (Oct. 30, 2023) ("The Authors Guild Initial Comments") ("[W]e will need to find a way to prevent authors' body of work or recognizable style from being exploited by others without permission.").

[317] Some commenters described similar results where an image prompt uses copies of the artist's work. *E.g.*, Katherine Lee et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 44–46 n.229 (Oct. 29, 2023) (providing as a hypothetical text-to-image prompt, "a big dog facing left wearing a spacesuit in a bleak lunar landscape with the earth rising in the background as an oil painting in the style of Paul Cezanne high-resolution aesthetic trending on artstation"); *see also id.* at 20 n.66, 46 (describing image-to-image prompts).

[318] *See, e.g.*, *Artificial Intelligence and Intellectual Property: Part II—Identity in the Age of AI: Hearing Before the Subcomm. on Cts., Intell. Prop., & the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024) (statement of Karla Ortiz, Concept Artist, Illustrator, and Fine Artist) ("Artists and creators who have spent a lifetime honing and refining a skill can now have facsimiles of their hard work reproduced in an instant by a Generative AI model that has been trained on their work . . . .").

[319] Center for AI and Digital Policy, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 15–16 (Oct. 30, 2023).

[320] The Authors Guild, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Oct. 30, 2023) ("The Authors Guild Initial Comments") ("Suddenly, we see people using generative AI to generate texts in the style of authors. . . . [W]e have already seen someone write the last two novels in George R.R. Martin's *A Song of Ice and Fire (Game of Thrones)* series.").

[321] *See, e.g.*, Letter from The Authors Guild, Summary of *Ex Parte* Meeting on May 6, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (May 10, 2024) (noting "that the need for [style] protection has increased in light of the ease with which materials lacking proper attribution can be created and disseminated by AI, as well as the recent explosion of false attribution for AI generated works").

> [A]s of November 29th, 2023, the top result when googling American artist Kelly McKernan is an AI-generated imitation of her style. Not only does this demonstrate the ability of AI forgeries to quickly propagate and pollute search engines and the wider internet but this digital impersonation has a chilling effect on artists' agency and ability to control their online identity. Artists are essentially competing with a distorted version of themselves.[322]

The Office acknowledges the seriousness of these concerns and believes that appropriate remedies should be available for this type of harm.

Copyright law's application in this area is limited, as it does not protect artistic style as a separate element of a work.[323] As noted by several commenters, copyright protection for style would be inconsistent with section 102(b)'s idea/expression dichotomy.[324] Moreover, in most cases the elements of an artist's style cannot easily be delineated and defined separately from a particular underlying work.[325] Google and EFF both stressed that, as a policy matter, stylistic

---

[322] Anonymous Artist, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1 (Dec. 5, 2023).

[323] 17 U.S.C. § 102(b) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery."); *see also Douglas v. Osteen*, 317 F. App'x. 97, 99 (3d Cir. 2009) ("[T]he use of a particular writing style or literary method is not protected by the Copyright Act."); *Whitehead v. CBS/Viacom, Inc.*, 315 F. Supp. 2d 1, 11 (D.D.C. 2004) ("While similar writing styles may contribute to similarity between works' total concept and feel, a particular writing style or method of expression standing alone is not protected by the Copyright Act."); *Tangle Inc. v. Aritzia, Inc.*, No. 23-cv-1196, 2023 U.S. Dist. LEXIS 187348, at *7 (N.D. Cal.) ("Style, no matter how creative, is an idea, and is not protectible by copyright."). *But see* Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38.1 HARV. J.L. & TECH. (forthcoming 2024), https://www.bensobel.org/files/articles/Sobel_Elements-of-Style_Public-Draft-May-18-2024.pdf ("[A]n honest application of copyright law requires us to acknowledge that some of what we call style is copyrightable some of the time, and that in some legal contexts courts regularly protect emergent copyright interests that span multiple works.").

[324] *See, e.g.*, Pamela Samuelson Initial Comments at 36–37 ("[A]ny concept of style that can only be identified by considering several works collectively is far too abstract to merit copyright protection consistent with the idea/expression distinction and Section 102(b). Even if proposed copyright protection for 'style' were focused on stylistic features of individual works, it is difficult to see how copyright protection for style or artistic technique could be reconciled with the idea/expression distinction and Section 102(b)."); MPA Initial Comments at 74 ("However, the law does not grant individuals exclusive rights over artistic style. . . . This conclusion flows ineluctably from one of copyright's most fundamental precepts: that it protects expression, not ideas."). *Cf.* 17 U.S.C. § 102(b) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.").

[325] *See* 2 PATRY ON COPYRIGHT § 4:14 (2024) ("If an . . . artist claimed broad protection for a style not associated with a particular work . . . it would be difficult, if not impossible, to determine the scope of protection.").

aspects of expressive content should remain freely available for later creators to develop and build on.[326]

The Copyright Act may, however, provide a remedy where the output of an "in the style of" request ends up replicating not just the artist's style but protectible elements of a particular work. Additionally, as future Parts of this Report will discuss, there may be situations where the use of an artist's own works to train AI systems to produce material imitating their style can support an infringement claim.

Numerous commenters pointed out that meaningful protections against imitations of style may be found in other legal frameworks,[327] including the Lanham Act's prohibitions on passing off and unfair competition.[328] In its comments, the FTC stated:

> [M]imicking the creator's writing style . . . may also constitute an unfair method of competition or an unfair or deceptive practice, especially when the copyright violation deceives consumers, exploits a creator's reputation or diminishes the value of her existing or future works, reveals private information, or otherwise causes substantial injury to consumers.[329]

---

[326] *See* EFF Initial Comments at 7 ("A greater degree of restriction on the public's permissible range of speech, particularly one as elusive as a 'style,' would undermine the cultural advancement at the core of copyright's goals."); Google LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 16 (Oct. 30, 2023) ("[A] law protective of artistic style would inevitably inhibit creativity and impoverish the public domain, which is the cultural commons from which all artists can freely take inspiration.").

[327] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 41 ("To the extent that AI systems are based on, or derive their value from, a particular artist's identity, that artist should be protected by laws governing the use of an individual's brand or identity (such as the individual's voice or likeness), including the Lanham Act and laws regarding rights of publicity and unfair competition."); SAG-AFTRA Initial Comments at 8 ("To the extent an AI system is based on, or derives value from, the artist's brand or identity, that artist should have legal recourse. Laws such as the right of publicity or Lanham Act that protect an individual's persona may be implicated, including when an AI program generates output by using the name of a specific artist as a prompt."); Pamela Samuelson Initial Comments at 37 ("The tendency of users of text-to-image generators to invoke the names of living artists in prompts has caused considerable consternation . . . it can occasionally result in the names of particular artists being publicly associated with works they did not author, to an extent that dwarfs their own substantial artistic contributions . . . . This seems like harm that trademark law and right of publicity could address more easily than copyright law.").

[328] *See* Andrew J. Noreuil, *Nice Tie: Trade Dress Protection for Visual Artistic Style When Competitors Offer Artist-Inspired Products*, 67 FORDHAM L. REV. 3403, 3427 (1999) (discussing *Walt Disney Co. v. Goodtimes Home Video Corp.*, 830 F. Supp. 762 (S.D.N.Y. 1993)).

[329] FTC Initial Comments at 5–6. Adobe has proposed that Congress enact legislation, titled the Federal Anti-Impersonation Right (FAIR) Act, which would provide a right of action to artists whose unique personal style or likeness is intentionally imitated using AI tools for commercial gain. Adobe Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 8 (Oct. 30, 2023).

Although state right of publicity statutes do not explicitly refer to style, where a particular style is closely identified with an individual performer, it may be protected.[330] Protection may also be available under the common law.[331] Although the law in this area is not fully developed, that may be because the means of easy and near-perfect stylistic impersonation have not been widely available until recently,[332] and the advent of generative AI may result in an increase in such claims. Meanwhile, some AI developers have reportedly placed guardrails in their systems blocking requests to generate images in the style of living artists.[333]

In sum, there are several sources of protection under existing laws that may be effective against unfair or deceptive copying of artistic style. Given these resources, as well as the policy reasons not to extend property-like rights to style in itself, the Office does not recommend including style as protected subject matter under a federal digital replica law at this time.[334] If existing protections prove inadequate, this conclusion may be revisited.

---

[330] *See* 2 J. THOMAS MCCARTHY & ROGER E. SCHECHTER, THE RIGHTS OF PUBLICITY AND PRIVACY §§ 8:110, 4:75 (2d. ed. 2024); *see also Muzikowski v. Paramount Pictures Corp.*, No. 01-cv-6721, 2003 U.S. Dist. LEXIS 21766, at *17–18 (N.D. Ill. Dec. 2, 2003) (holding that the "Illinois Right of Publicity Act defines a person's 'identity' broadly . . . . While the term 'persona' is not contained in the [statute's] list of examples [of attributes that serve to identify an individual], the breadth of the . . . statute supports the conclusion that Illinois courts would use an expansive approach in determining what kinds of attributes are protected under the statute."). *But see Burck v. Mars, Inc.*, 571 F. Supp. 2d 446, 452 (S.D.N.Y. 2008) (holding that the New York right of publicity statute should be "strictly construed and is not to be applied so as to prohibit the portrayal of an individual's personality or style of performance").

[331] The Ninth Circuit has recognized athletic play style as an element of "likeness" under California common law, *Keller v. Elec. Arts, Inc.*, 724 F.3d 1268, 1272 (9th Cir. 2013), and the Northern District of Texas has recognized "performing style" as a protectible aspect of an individual's persona under Texas common law, *Henley v. Dillard Dep't Stores*, 46 F.Supp. 2d 587, 591 (N.D. Tex. 1999) (citing *Elvis Presley Enters. V. Capece*, 950 F.Supp. 783, 801 (S.D. Tex. 1996)).

[332] *See, e.g.*, Authors Guild Initial Comments at 10.

[333] *See Index: Creative Control*, OPENAI, https://openai.com/index/dall-e-3/ ("DALL·E 3 is designed to decline requests that ask for an image in the style of a living artist.") (last visited July 21, 2024); *Frequently Asked Questions*, COPILOT, https://www.bing.com/images/create/help ("We allow living artists, celebrities, and organizations to make requests to limit the creation of images associated with their names and brands.") (last visited July 21, 2024).

[334] In addition to artistic style, some commenters identified other subject matter—specifically name or attribution—they would like to have covered by a digital replica law. They seek to bar the unauthorized use of an individual's name on or in connection with AI-generated material or creative works in general—a different type of harm from the use of realistic image or voice replicas produced by AI, addressed in this Report. The Authors Guild would prohibit unauthorized use of an author's name in connection with AI-generated material. Letter from Authors Guild, Summary of *Ex Parte* Meeting on May 6, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (May 10, 2024). This would prevent unauthorized users from labeling such material with the name of the author the AI system was prompted to imitate. The Directors Guild of America would add moral rights of attribution and integrity, *see supra* note 41, in order to prevent the "harm to a Director's reputation when his/her creative work is altered without their involvement and when their name is falsely attributed to, or deleted from, a creative work." Letter from

IV. CONCLUSION

The Copyright Office agrees with the numerous commenters that have asserted an urgent need for new protection at the federal level. The widespread availability of generative AI tools that make it easy to create digital replicas of individuals' images and voices has highlighted gaps in existing laws and raised concerns about the harms that can be inflicted by unauthorized uses.

We recommend that Congress establish a federal right that protects all individuals during their lifetimes from the knowing distribution of unauthorized digital replicas. The right should be licensable, subject to guardrails, but not assignable, with effective remedies including monetary damages and injunctive relief. Traditional rules of secondary liability should apply, but with an appropriately conditioned safe harbor for OSPs. The law should contain explicit First Amendment accommodations. Finally, in recognition of well-developed state rights of publicity, we recommend against full preemption of state laws.

The Office remains available as a resource to Congress, the courts, and the executive branch in considering the recommendations in this Report and future developments.

Directors Guild, Summary of *Ex Parte* Meetings on May 22 and May 29, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 1–2 (June 4, 2024). In this area, too, it is important to note that several bodies of existing law protect against the unauthorized and confusing use of an individual's name. *See, e.g.*, CAL. CIV. CODE § 3344 (West 2024) (protecting "name, voice, signature, photograph, or likeness"); FLA. STAT. § 540.08 (2024) (protecting "name, portrait, photograph, or other likeness"); VA. CODE ANN. § 8.01-40 (2024) (protecting "name, portrait, or picture"); 15 U.S.C. § 1125 (providing a cause of action for using, among other things, a name in connection with goods and services that is likely to cause confusion).

ACKNOWLEDGEMENTS

Part 1 of the Copyright Office report on *Copyright and Artificial Intelligence,* addressing the topic of Digital Replicas, is the product of a large and talented team.

I would particularly like to acknowledge the Office's senior leaders who guided the project from the beginning and contributed their deep expertise. Suzy Wilson, General Counsel and Associate Register of Copyrights, and Andrew Foglia, Deputy Director for Policy and International Affairs, provided insightful analyses of novel and complex issues. Emily Chapuis, Deputy General Counsel, and Maria Strong, Associate Register of Copyrights and Director of Policy and International Affairs, played an important role in oversight, research, and drafting.

Thanks also go to the lawyers in the Office of Policy and International Affairs (PIA) and the Office of General Counsel (OGC) who were responsible for the foundational research and writing for this Part. This included Chris Weston, Caitlin Costello, and Danielle Johnson from PIA, with Brandy Karl, Heather Walters, and Gabriela Rojas-Luna from OGC. Additional research and input were provided by Jason Sloan from OGC, as well as PIA law clerks Max Alter, Matt Blaszczyk, and Justin Ross. Benjamin Brady and Isaac Klipstein from PIA, Heather Walters, Gabriela Rojas-Luna and Rhea Efthimiadis from OGC, and Jessica Chinnadurai from the Office of Public Information and Education (PIE) all assisted in finalizing the text and citations.

PIE's production and communications team ably shepherded the document's preparation and public release. Led by Associate Register for Public Information and Education Miriam Lord and Deputy Director George Thuronyi, the team included Steve Andreadis, Nicole Chen, Lisa Marflak, Alison Hall, Stanley Murgolo, Anjana Padmanabhan, Nora Scheland, and Naomi Wulansari.

Finally, I thank our colleagues in the U.S. Patent and Trademark Office, the U.S. Department of Justice, and the Federal Trade Commission for the open lines of communication and responsiveness in providing helpful information.

Shira Perlmutter
Register of Copyrights and Director
U.S. Copyright Office
July 31, 2024

To Follow

Further Parts of the U.S. Copyright Office's Copyright and Artificial Intelligence Report will be published in 2024. Visit www.copyright.gov/AI for more information and to sign up for updates.

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 2: Copyrightability

A REPORT OF THE REGISTER OF COPYRIGHTS

JANUARY 2025

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 2: Copyrightability

A REPORT OF THE REGISTER OF COPYRIGHTS

JANUARY 2025

ABOUT THIS REPORT

This Report by the U.S. Copyright Office addresses the legal and policy issues related to artificial intelligence ("AI") and copyright, as outlined in the Office's August 2023 Notice of Inquiry ("NOI").

The Report will be published in several Parts, each one addressing a different topic. This Part addresses the copyrightability of works created using generative AI. The first Part, published in 2024, addresses the topic of digital replicas—the use of digital technology to realistically replicate an individual's voice or appearance. A subsequent part will turn to the training of AI models on copyrighted works, licensing considerations, and allocation of any liability. To learn more, visit www.copyright.gov/ai.

ABOUT THE U.S. COPYRIGHT OFFICE

The U.S. Copyright Office is the federal agency charged by statute with the administration of U.S. copyright law. The Register of Copyrights advises Congress, provides information and assistance to courts and executive branch agencies, and conducts studies on national and international issues relating to copyright, other matters arising under Title 17, and related matters. The Copyright Office is housed in the Library of Congress. Its mission is to promote "creativity and free expression by administering the nation's copyright laws and by providing impartial, expert advice on copyright law and policy for the benefit of all." For more information, visit www.copyright.gov.

PREFACE

In early 2023, the U.S. Copyright Office announced a broad initiative to explore the intersection of copyright and artificial intelligence.

In March of that year, the Office released a policy statement with registration guidance for works incorporating AI-generated content. Over the spring and summer, we hosted a series of online listening sessions, presented educational webinars, and engaged with numerous stakeholders to enhance our understanding of the technology and how it is used, the copyright implications, and the potential impact on businesses and individuals.

These activities culminated in an August 2023 Notice of Inquiry, formally seeking public input on the full range of copyright issues that had been raised. In response, we received more than 10,000 comments representing a broad range of perspectives, including from authors and composers, performers and artists, publishers and producers, lawyers and academics, technology companies, libraries, sports leagues, trade groups and public interest organizations, and even a class of middle school students. Comments came from all 50 states and from 67 countries. That valuable and extensive input, supplemented by additional Office research and information received from other agencies, forms the basis for the discussion and recommendations in this Report.

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence
Part 2: Copyrightability

A REPORT OF THE REGISTER OF COPYRIGHTS

JANUARY 2025

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This second Part of the Copyright Office's Report on Copyright and Artificial Intelligence ("AI") addresses the copyrightability of outputs generated by AI systems. It analyzes the type and level of human contribution sufficient to bring these outputs within the scope of copyright protection in the United States.

Of the more than 10,000 comments the Office received in response to its Notice of Inquiry ("NOI"), approximately half addressed copyrightability. The vast majority of commenters agreed that existing law is adequate in this area and that material generated wholly by AI is not copyrightable.

Commenters differed, however, as to protection for generative AI outputs that involve some form of human contribution. They expressed divergent views on what types and amounts of contribution could constitute authorship under existing law. Many also stressed the desirability of greater clarity in this area, including with respect to the use of AI as a tool in the creative process.

As a matter of policy, some argued that extending protection to materials created by generative AI would encourage the creation of more works of authorship, furthering progress in culture and knowledge to the benefit of the public. The Office also heard concerns that an increased proliferation of AI-generated outputs would undermine incentives for humans to create.

While recognizing that copyrightability is determined on a case-by-case basis, in this Part the Office sets out the legal principles that govern the analysis and assesses their application to AI-generated content.

Section I identifies the copyrightability issues raised by AI technologies. It outlines the history of adapting copyright law to new technological developments and describes the Office's ongoing AI initiative.

Section II provides a brief background on the technologies involved. It then summarizes the existing legal framework, particularly the human authorship requirement, the idea/expression dichotomy, and the originality standard for copyright protection. After discussing the use of AI to assist authors in the process of creating works of authorship, it analyzes how the law may apply to various types of human contributions to AI-generated outputs: prompting, the inclusion of human-authored expressive inputs, and the modification or arrangement of AI-generated outputs.

Section III reports on the international landscape. It describes how other countries are approaching questions of copyrightability within their own legal systems.

Section IV addresses the policy implications of providing additional legal protection to AI-generated material and evaluates the arguments for and against legislative change.

Based on an analysis of copyright law and policy, informed by the many thoughtful comments in response to our NOI, the Office makes the following conclusions and recommendations:

- Questions of copyrightability and AI can be resolved pursuant to existing law, without the need for legislative change.
- The use of AI tools to assist rather than stand in for human creativity does not affect the availability of copyright protection for the output.
- Copyright protects the original expression in a work created by a human author, even if the work also includes AI-generated material.
- Copyright does not extend to purely AI-generated material, or material where there is insufficient human control over the expressive elements.
- Whether human contributions to AI-generated outputs are sufficient to constitute authorship must be analyzed on a case-by-case basis.
- Based on the functioning of current generally available technology, prompts do not alone provide sufficient control.
- Human authors are entitled to copyright in their works of authorship that are perceptible in AI-generated outputs, as well as the creative selection, coordination, or arrangement of material in the outputs, or creative modifications of the outputs.
- The case has not been made for additional copyright or *sui generis* protection for AI-generated content.

The Office will continue to monitor technological and legal developments to determine whether any of these conclusions should be revisited. It will also provide ongoing assistance to the public, including through additional registration guidance and an update to the *Compendium of U.S. Copyright Office Practices*.[1]

[1] U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES (3d ed. 2021) ("COMPENDIUM (THIRD)").

I. INTRODUCTION

This Part of the Copyright Office's Report on Copyright and Artificial Intelligence addresses the use of AI systems to produce outputs that would be copyrightable if created by a human author.

The use of technology in the production of works of authorship is not new. Authors have used computer-assisted technology for decades to enhance, modify, and add to their creations—expanding their range of expression and advancing the goals of the copyright system. And today they are leveraging advancements in technology to push the boundaries of creativity in exciting ways. Neither the use of AI as an assistive tool nor the incorporation of AI-generated content into a larger copyrightable work affects the availability of copyright protection for the work as a whole. But the capabilities of the latest generative AI technologies[2] raise challenging questions about the nature and scope of human authorship.

These technologies now permit the creation of textual, visual, and sound outputs that resemble the creative works traditionally protected by copyright. Should these outputs also enjoy copyright protection? The answer will turn on the nature and extent of a human's contribution, and whether it qualifies as authorship of expressive elements contained in the output. Finally, to the extent that protection is not available under existing copyright principles, should the law be changed? If so, how?

A. Technology and Copyright

As stated in the legislative history of the 1976 Copyright Act, "[t]he history of copyright law has been one of gradual expansion in the types of works accorded protection."[3]

Over the years, copyright has proven flexible enough to respond to new technologies and mediums as they emerge. Protection has been extended to photographs, motion pictures, video games, and computer programs—to name just a few.[4] At the same time, courts have been called on to explore and analyze the nature of authorship. As authors have increasingly used

---

[2] "Generative AI" refers to "application[s] of AI used to generate outputs in the form of expressive material such as text, images, audio, or video." Artificial Intelligence Study: Notice of Inquiry, 88 Fed. Reg. 59942, 59948–49 (Aug. 30, 2023) ("NOI").

[3] H.R. REP. NO. 94-1496, at 51 (1976), *reprinted in* 1976 U.S.C.C.A.N. 5659, 5664.

[4] When Congress extended copyright protection to architecture, it explained that these types of works would be governed by "the general standards of originality applicable for all other copyrightable subject matter." H.R. REP. NO. 101-735, at 21 (1990), *reprinted in* 1990 U.S.C.C.A.N. 6935, 6952. Courts have also applied those standards to claims involving new technology in numerous cases. *See, e.g.*, *Meshwerks, Inc. v. Toyota Motor Sales U.S.A.*, 528 F.3d 1258, 1264–65 (10th Cir. 2008) (then-judge Neil Gorsuch stating "we do not doubt for an instant that the digital medium before us, like photography before it, can be employed to create vivid new expressions fully protectable in copyright"); *Stern Elecs., Inc. v. Kaufman*, 669 F.2d 852, 856–67 (2d Cir. 1982) (audiovisual work); *M. Kramer Mfg. Co. v. Andrews*, 783 F.2d 421, 436 (4th Cir. 1986) (video games); *Tandy Corp. v. Personal Micro Computs., Inc.*, 524 F. Supp. 171, 173 (N.D. Cal. 1981) (computer program and silicon chip).

technology in the process of creation, the relative roles of human and machine can be central to the analysis of copyrightability.

Given its role in registering claims to copyright,[5] the Copyright Office has considerable experience addressing technological developments related to the creation of works of authorship. As early as 1965, developments in computer technology began to raise "difficult questions of authorship," including whether material created using technology is "'written' by computers" or authored by human creators.[6] As then-Register of Copyrights Abraham Kaminstein observed, there is no one-size-fits-all answer:

> The crucial question appears to be whether the "work" is basically one of human authorship, with the computer merely being an assisting instrument, or whether the traditional elements of authorship in the work (literary, artistic, or musical expression or elements of selection, arrangement, etc.) were actually conceived and executed not by man but by a machine.[7]

Because the answer depends on the circumstances of a work's creation, Barbara Ringer (then-Chief of the Examining Division and future Register of Copyrights) noted that the Office could not "take the categorical position that registration will be denied merely because a computer may have been used in some manner in creating the work."[8]

The same analysis applies in the context of AI technology. For a work created using AI, like those created without it, a determination of copyrightability requires fact-specific consideration of the work and the circumstances of its creation. Where AI merely assists an author in the creative process, its use does not change the copyrightability of the output. At the other extreme, if content is entirely generated by AI, it cannot be protected by copyright.[9] Between these boundaries, various forms and combinations of human contributions can be involved in producing AI outputs.

While few bright-line rules are possible in assessing copyrightability, this Part of the Report seeks to shed more light on the relevant considerations.

---

[5] The Register of Copyrights is responsible for administering the copyright system, including examining claims for copyright registration. 17 U.S.C. §§ 410(a), 701(a). Although copyright vests automatically in an original work of authorship when fixed in a tangible medium, registration (or its refusal) provides a number of practical and legal benefits, including enabling U.S. copyright owners to enforce their exclusive rights in court. *See generally id.* §§ 106, 408(a), 410(c), 412, 411(a); U.S. Copyright Office, Circular 1: Copyright Basics (Sept. 2021), https://copyright.gov/circs/circ01.pdf.

[6] U.S. COPYRIGHT OFFICE, SIXTY-EIGHTH ANNUAL REPORT OF THE REGISTER OF COPYRIGHTS FOR THE FISCAL YEAR ENDING JUNE 30, 1965, at 5 (1966), https://www.copyright.gov/reports/annual/archive/ar-1965.pdf.

[7] *Id.*

[8] U.S. COPYRIGHT OFFICE, ANNUAL REPORT OF THE Examining Division, Copyright Office, for the Fiscal Year 1965, at 4 (1965), https://copyright.gov/reports/annual/archive/ar-examining1965.pdf.

[9] *See Thaler v. Perlmutter*, 687 F. Supp. 3d 140, 149–50 (D.D.C. 2023).

B. The Copyright Office's AI Initiative

In February 2022, the Copyright Office's Review Board issued a final decision affirming the refusal to register a work claimed to be generated with no human involvement.[10] A year later, the Office issued a registration for a comic book incorporating AI-generated material.[11]

In early 2023, the Office announced the launch of a broad AI Initiative and issued a statement of policy providing guidance on the registration of works incorporating AI-generated material (the "Guidance" or "AI Registration Guidance").[12] The Guidance reiterated the Office's longstanding position that human authorship is an essential requirement for copyright protection in the United States.[13] It explained that if a work contains more than a *de minimis* amount of AI-generated material, the applicant should disclose that information and provide a brief statement describing the human author's contribution.[14]

Since the Guidance was issued, the Office has registered hundreds of works that incorporate AI-generated material, with the registration covering the human author's contribution to the work.[15]

In August 2023, the Office issued a Notice of Inquiry seeking comments on a wide range of copyright law and policy issues arising from the development and use of generative AI.[16] The NOI asked five questions related to the copyrightability of material generated using AI systems:

(1) Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material?

---

[10] U.S. Copyright Office Review Board, *Decision Affirming Refusal of Registration of A Recent Entrance to Paradise* (Feb. 14, 2022), https://copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf.

[11] U.S. Copyright Office, *Cancellation Decision re: Zarya of the Dawn* (VAu001480196) at 5 (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf (explaining that registration covered the work's human-authored text as well as the human-authored selection, coordination, and arrangement of the work's written and visual elements, but not images generated by Midjourney that were not the product of human authorship).

[12] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190 (Mar. 16, 2023) ("AI Registration Guidance"). A copy of the guidance is available on the Office's website. U.S. COPYRIGHT OFFICE, COPYRIGHT REGISTRATION GUIDANCE: WORKS CONTAINING MATERIAL GENERATED BY ARTIFICIAL INTELLIGENCE (2023), https://copyright.gov/ai/ai_policy_guidance.pdf.

[13] AI Registration Guidance at 16191–92; *see also Thaler*, 687 F. Supp. 3d at 149–50.

[14] AI Registration Guidance at 16193; *see also* Registration Guidance for Works Containing AI-Generated Content Tr. (June 28, 2023), https://www.copyright.gov/events/ai-application-process/Registration-of-Works-with-AI-Transcript.pdf (webinar on registration of works incorporating AI-generated material).

[15] Registration records are searchable in the Office's public record, including by using keywords and filters to search the Copyright Public Record System. *Copyright Public Records System - Pilot*, U.S. COPYRIGHT OFFICE, https://publicrecords.copyright.gov/ (last visited Jan. 17, 2025).

[16] NOI.

(2) Under copyright law, are there circumstances when a human using a generative AI system should be considered the "author" of the material produced by the system?

(3) Is legal protection for AI-generated material desirable as a policy matter?

(4) If so, should it be a form of copyright or a separate *sui generis* right?

(5) Are any revisions to the Copyright Act necessary to clarify the human authorship requirement?[17]

Approximately fifty percent of the more than 10,000 comments received in response to the NOI addressed one or more of these questions. The Office refers to these comments throughout the discussion below.

[17] *Id.* at 59947–48.

II. AUTHORSHIP AND ARTIFICIAL INTELLIGENCE

A. Technological Background

In the NOI, the Office defined an AI system as a "software product or service that substantially incorporates one or more AI models and is designed for use by an end-user."[18] As components to larger systems, AI models consist of computer code and numerical values (or "weights") designed to accomplish certain tasks, like generating text or images.[19]

Many of today's publicly available AI systems allow for the generation of an output from one or more inputs, such as text, images, audio, video, or a combination of mediums. A "prompt" is a common type of input, often in the form of text, that communicates the desired features of the output.[20] The AI system responds to these inputs by generating an output in the requested format (text, image, audio, video). Prompts typically describe a topic, theme, and/or subject that the user seeks to evoke, and may include the overall style, tone, and/or visual technique. Some are short and simple, such as a request for a "cartoon spaceship." Others are more detailed, requesting a litany of elements. Users may enter a prompt a single time or iteratively, refining it until the system generates an acceptable output.[21]

The practice of crafting prompts that are optimized to elicit a desired result is sometimes called "prompt engineering."[22] Prompts can also be automatically optimized by a generative AI

---

[18] NOI at 59948; *see also* James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, Pub. L. 117–263, § 7223(4)(A), 136 Stat. 2395, 3669 (2022) (defining "artificial intelligence system" as "any data system, software, application, tool, or utility that operates in whole or in part using dynamic or static machine learning algorithms or other forms of artificial intelligence").

[19] NOI at 59948–49; *see* ZHANG ET AL., DIVE INTO DEEP LEARNING, ch. 1 (2023), https://d2l.ai/chapter_introduction/index.html (ebook); GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON, at 404–05 (2023), https://www.statlearning.com/ (ebook) (explaining that the parameters of a neural network are sometimes referred to as "weights").

[20] *See, e.g.*, Leonardo Banh & Gero Strobel, *Generative Artificial Intelligence*, 33:63 ELEC. MKTS. 1, 3 (2023), https://doi.org/10.1007/s12525-023-00680-1 ("*Prompting* . . . enables end users using natural language to engage with and instruct [generative AI] application (e.g., LLMs) to create desired output such as text, images, or other types."); *Prompt*, GENLAW GLOSSARY, https://blog.genlaw.org/glossary.html#prompt ("Most generative-AI systems take [an] input (currently, this is often some text), which is then used to condition the output. This input is called the prompt.") (last visited Jan. 17, 2025); *Image Prompts*, MIDJOURNEY, https://docs.midjourney.com/docs/image-prompts ("You can use images as part of a prompt to influence a Job's composition, style, and colors.") (last visited Jan. 17, 2025); Sander Schulhoff et al., *The Prompt Report: A Systematic Survey of Prompting Techniques* at 5, ARXIV (Dec. 30, 2024), https://arxiv.org/abs/2406.06608 ("A prompt is an input to a Generative AI model, that is used to guide its output.").

[21] Other strategies are more complex, such as "prompt chaining" where a complex prompt is divided into a sequence of intermediate subtasks with a prompt for each step. Robert Clariso & Jordi Cabot, *Model-Driven Prompt Engineering*, IEEE *XPLORE*, 2023, at 48, DOI: 10.1109/MODELS58315.2023.00020.

[22] *See, e.g.*, *id.* at 47; Sander Schulhoff et al., *The Prompt Report: A Systematic Survey of Prompting Techniques* at 7, ARXIV (Dec. 30, 2024), https://arxiv.org/abs/2406.06608.

system that revises or expands them in order to improve the quality of outputs.[23] For example, ChatGPT "automatically generate[s] tailored, detailed prompts for [OpenAI's text-to-image model] DALL·E 3."[24]

As described below,[25] however, the output of current generative AI systems may include content that was not specified and exclude content that was. Although AI technology continues to advance, uncertainty around how a particular prompt or other input will influence the output may be inherent in complex AI systems built on models with billions of parameters.[26] Some observers describe AI as a "black box,"[27] and even expert researchers are limited in their ability to understand or predict the behavior of specific models.[28]

---

[23] *See, e.g.*, Siddhartha Datta et al., *Prompt Expansion for Adaptive Text-to-Image Generation* at 4, 14, ARXIV (Dec. 27, 2023), https://arxiv.org/abs/2312.16720 (describing a model that "takes a text prompt as input, . . . and outputs a set of *N* expanded text prompts that include specialized keywords (to improve image quality) and interesting additional details (to add diversity to the generated images"); PROMPTPERFECT, https://promptperfect.jina.ai/ (last visited Jan. 17, 2025); PROMPTIST, https://foundr.ai/product/promptist (last visited Jan. 17, 2025).

[24] *DALL·E 3*, OPENAI, https://openai.com/index/dall-e-3/ (last visited Jan. 17, 2025).

[25] *See infra* notes 84–87 and pp. 24–25.

[26] *See, e.g.*, GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON, at 23–25 (2023), https://www.statlearning.com/ (ebook) (discussing the fundamental tradeoff between the flexibility and interpretability of statistical learning models, with neural networks as an example of highly flexible and difficult to interpret models); Christian Szegedy et al., *Intriguing properties of neural networks* at 1, ARXIV (Feb. 19, 2024), https://arxiv.org/abs/1312.6199 ("Neural networks achieve high performance because they can express arbitrary computation that consists of a modest number of massively parallel nonlinear steps. But as the resulting computation is automatically discovered[,] . . . it can be difficult to interpret and can have counter-intuitive properties."); Pantelis Linardatos et al., *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 ENTROPY 1, 1 (Dec. 25, 2020), https://dx.doi.org/10.3390/e23010018 (The "increasing complexity combined with the fact that vast amounts of data are used to train and develop such complex systems, while, in most cases, boost[ing] the systems' predictive power, inherently reduc[es] the[] ability to explain their inner workings and mechanisms. As a consequence, the rationale behind their decisions becomes quite hard to understand and, therefore, their predictions hard to interpret.").

[27] Steven Levy, *AI Is a Black Box. Anthropic Figured Out a Way to Look Inside*, WIRED (May 24, 2024), https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features/ ("When I asked the researchers whether they were claiming to have *solved* the black box problem, their response was an instant and unanimous no."); Lou Blouin, *AI's mysterious 'black box' problem, explained*, UMDEARBORN.EDU NEWS (Mar. 6, 2023), https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained. *See also infra* notes 84–87.

[28] *See, e.g.*, Trenton Bricken et al., *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*, TRANSFORMER CIRCUITS THREAD (Oct. 4, 2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html ("Mechanistic interpretability seeks to understand neural networks by breaking them into components that are more easily understood than the whole. By understanding the function of each component, and how they interact, we hope to be able to reason about the behavior of the entire network."); Adly Templeton et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*, TRANSFORMER CIRCUITS THREAD (May 21, 2024), https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html ("Our work has many limitations. Some of these are superficial limitations relating to this work being early, but others are deeply fundamental challenges that require novel research to address.").

In addition, many popular AI systems are unpredictable in the sense that their outputs may vary from request to request, even with an identical prompt.[29] Some systems allow users to control this behavior and generate consistent results by setting a "seed" value, which is a number used to initialize the output generation process.[30] For example, Midjourney users can set a seed (*e.g.*, "123") and receive nearly identical images when repeating the same prompt.[31] Even these systems, however, are not always able to guarantee perfect consistency.[32]

B. Legal Framework

As the Office affirmed in the Guidance, copyright protection in the United States requires human authorship. This foundational principle is based on the Copyright Clause in the Constitution and the language of the Copyright Act as interpreted by the courts. The Copyright Clause grants Congress the authority to "secur[e] for limited times to authors . . . the exclusive right to their . . . writings."[33] As the Supreme Court has explained, "the author [of a copyrighted work] is . . . *the person* who translates an idea into a fixed, tangible expression entitled to copyright protection."[34]

No court has recognized copyright in material created by non-humans, and those that have spoken on this issue have rejected the possibility. In two well-known cases, the Ninth

[29] *See, e.g.*, *Reproducible Outputs*, OPENAI, https://platform.openai.com/docs/advanced-usage/reproducible-outputs (last visited Jan. 17, 2025); Shuyin Ouyang et al., *LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation*, ARXIV (Oct. 17, 2024), https://arxiv.org/abs/2308.02828.

[30] *See, e.g.*, *Reproducible Outputs*, OPENAI, https://platform.openai.com/docs/advanced-usage/reproducible-outputs (last visited Jan. 17, 2025); *Seeds*, MIDJOURNEY, https://docs.midjourney.com/docs/seeds ("The Midjourney bot uses a seed number to create a field of visual noise, like television static, as a starting point to generate the initial image grids. Seed numbers are generated randomly for each image but can be specified with the --seed parameter. If you use the same seed number and prompt, you will get similar final images.") (last visited Jan. 17, 2025).

[31] *Seeds*, MIDJOURNEY, https://docs.midjourney.com/docs/seeds (last visited Jan. 17, 2025).

[32] *See* Alexander Schlögl et al., *Causes and Effects of Unanticipated Numerical Deviations in Neural Network Inference Framework*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 36 (A. Oh et al. eds., 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/af076c3bdbf935b81d808e37c5ede463-Abstract-Conference.html; *Reproducible Outputs*, OPENAI, https://platform.openai.com/docs/advanced-usage/reproducible-outputs (explaining that users can obtain "mostly" deterministic outputs by setting the same seed value) (last visited Jan. 17, 2025); *Seeds*, MIDJOURNEY, https://docs.midjourney.com/docs/seeds ("Identical --seed values [for certain model versions] will produce *nearly* identical images.") (emphasis added) (last visited Jan. 17, 2025).

[33] U.S. Const. art. I, § 8, cl. 8.

[34] *Cmty. for Creative Non-Violence v. Reid* ("*CCNV*"), 490 U.S. 730, 737 (1989) (emphasis added).

Circuit held that text purportedly "authored by non-human spiritual beings"[35] and photographs that a monkey captured with a camera could not be protected by copyright.[36]

In 2023, the U.S. District Court for the District of Columbia became the first court to specifically address the copyrightability of AI-generated outputs.[37] The plaintiff challenged the Office's refusal to register an image that was described in his application as "autonomously created by a computer algorithm running on a machine."[38] Affirming the Office's refusal, the court stated that "copyright law protects only works of human creation," and that "human authorship is a bedrock requirement of copyright."[39] It found that "copyright has never stretched so far [as] . . . to protect works generated by new forms of technology operating absent any guiding human hand."[40] Because, by his own representation, the "plaintiff played no role in using the AI to generate the work," the court held that it did not meet the human authorship requirement.[41] The decision has been appealed.[42]

In most cases, however, humans will be involved in the creation process, and the work will be copyrightable to the extent that their contributions qualify as authorship. It is axiomatic that ideas or facts themselves are not protectible by copyright law,[43] and the Supreme Court has made clear that originality is required, not just time and effort. In *Feist Publications, Inc. v. Rural Telephone Service Co*., the Court rejected the theory that "sweat of the brow" alone could be sufficient for copyright protection.[44] "To be sure," the Court further explained, "the requisite level of creativity is extremely low; even a slight amount will suffice. The vast majority of

---

[35] *Urantia Found. v. Kristen Maaherra*, 114 F.3d 955, 957–59 (9th Cir. 1997) (holding that "some element of human creativity must have occurred in order for the Book to be copyrightable" because "it is not creations of divine beings that the copyright laws were intended to protect"). While the compilation of the book was entitled to copyright, the alleged "divine messages" were not. *Id.*

[36] *Naruto v. Slater*, No. 15-cv-04324, 2016 U.S. Dist. LEXIS 11041, at *10 (N.D. Cal. Jan. 28, 2016) ("[Monkey] is not an 'author' within the meaning of the Copyright Act"), *aff'd*, 888 F.3d 418 (9th Cir. 2018) (finding that monkey cannot sue for copyright infringement).

[37] *Thaler*, 687 F. Supp. 3d 140. A second case challenging the Office's refusal to register an AI-generated output was recently filed. Compl., Allen v. Perlmutter, No. 1:24-cv-2665 (D. Colo. Sept. 26, 2024), Doc. No. 1.

[38] *Thaler*, 687 F. Supp. 3d at 142–43.

[39] *Id.* at 146.

[40] *Id.*

[41] *Id.* at 149–50.

[42] Notice of Appeal, Thaler v. Perlmutter, No. 23-5233 (D.C. Cir. Oct 18, 2023). Oral argument was heard on September 19, 2024.

[43] *See* 17 U.S.C. § 102(b); *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 344-45 (1991) (explaining that "[t]he most fundamental axiom of copyright law is that 'no author may copyright his ideas or the facts he narrates'" (quoting *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 556 (1985))).

[44] 499 U.S. at 352–61.

works make the grade quite easily, as they possess some creative spark, 'no matter how crude, humble or obvious' it might be."[45]

More than a century ago, the Court analyzed the nature of authorship in a case involving the then-new technology of the camera. In *Burrow-Giles Lithographic Co. v. Sarony*, the Court considered a constitutional challenge to Congress's extension of copyright protection to photographs.[46] The defendant argued that photographs were not copyrightable because they lacked human authorship; instead, they were the product of a machine.[47]

The Court began its analysis by defining an "author" as "he to whom anything owes its origin; originator; maker; one who completes a work of science or literature."[48] It described copyright as "the exclusive right of a man to the production of his own genius or intellect."[49] Applying this framework, it identified numerous creative contributions made by the photographer, including "posing the [subject] in front of the camera, selecting and arranging the costume, draperies, and other various accessories," "arranging the subject so as to present graceful outlines," and "evoking the desired expression."[50] In sum, the use of a machine as a tool does not negate copyright protection, but the resulting work is copyrightable only if it contains sufficient human-authored expressive elements.

More recently, in cases involving more than one human contributor, courts have grappled with the nature of the contribution necessary to qualify as authorship. The Supreme Court provided additional guidance in the context of a commissioned sculpture. The parties in *Community for Creative Non-Violence v. Reid* ("*CCNV*") disputed who the author of the sculpture was: the nonprofit organization that conceived of it or the artist asked to make it. The Court concluded that the artist's contributions, which included sketching the design and executing his creative vision in a tangible medium of expression, made him an author.[51] In a remand to the trial court to determine whether the organization could be a joint author of the sculpture, the D.C. Circuit made clear that commissioning the sculpture and providing detailed suggestions and directions were insufficient, as such contributions constitute unprotectible ideas.[52]

The Third Circuit engaged in a similar analysis in *Andrien v. Southern Ocean County Chamber of Commerce*. *Andrien* involved an authorship claim by a plaintiff who had asked a

---

[45] *Id.* at 345.

[46] 111 U.S. 53, 55–57 (1884).

[47] *Id.* at 56, 59–60.

[48] *Id.* at 57–58 (internal quotation marks omitted).

[49] *Id.* at 58.

[50] *Id.* at 60 (internal quotation marks omitted).

[51] *CCNV*, 490 U.S. at 751–53.

[52] *Cmty. for Creative Non-Violence v. Reid*, 846 F.2d 1485, 1497 (D.C. Cir. 1988).

printer to rescale and print a collection of maps.[53] The plaintiff had "expressly directed the copy's preparation in specific detail," so that the "compilation needed only simple transcription to achieve final tangible form."[54] Because the printer "did not change the substance of [plaintiff's] original expression," the court held that the plaintiff was the author.[55] Applying *CCNV*, it stated that the author is the "party who actually creates the work, that is, the person who translates an idea into an expression that is embodied in a copy by himself or herself, or who authorizes another to embody the expression in a copy."[56]

Although an AI-generated output cannot be considered a joint work with respect to the user and AI system,[57] joint authorship provides a helpful analogy in assessing whether a party contributed sufficient expression to be considered an author.[58] To be a joint author, one must make a copyrightable contribution.[59] "A person who merely describes to an author what the

---

[53] 927 F.2d 132, 133 (3d Cir. 1991) (Under plaintiff's direction, the printer's work "included coordinating the scales, relettering the street names and adding designations for the diving sites as well as for local points of interest.").

[54] *Id.* at 135.

[55] *Id.* at 135–36. *Cf. S.O.S., Inc. v. Payday, Inc.*, 886 F.2d 1081, 1086–87 (9th Cir. 1989) (rejecting an authorship claim from a party who commissioned software noting that "[t]he supplier of an idea is no more an 'author' of a program than is the supplier of the disk on which the program is stored"); *M.G.B. Homes, Inc. v. Ameron Homes Inc.*, 903 F.2d 1486, 1493 (11th Cir. 1990) (providing sketches and ideas did not render client an "author" of the finished expression); *Geshwind v. Garrick*, 734 F. Supp. 644 (S.D.N.Y. 1990) (producer was not the author where he "wanted changes in details and aspects of the [animation clip] and even made suggestions," but did not materially constrain the animator's expression or otherwise influence how the animator executed the instructions), *vacated in part on other grounds*, 738 F. Supp. 792 (S.D.N.Y. 1990), *and aff'd*, 927 F.2d 594 (2d Cir. 1991); *Whelan Assocs. v. Jaslow Dental Lab'y*, 609 F. Supp. 1307, 1318–19 (E.D. Pa. 1985) ("general assistance and contributions to the fund of knowledge" do not make one "a creator of any original work"), *amended*, 609 F. Supp. 1325 (E.D. Pa. 1985), *aff'd*, 797 F.2d 1222 (3d Cir. 1986), *and cert. denied*, 479 U.S. 1031 (1987).

[56] *Andrien*, 927 F.2d at 134–35 ("When one authorizes embodiment, that process must be rote or mechanical transcription that does not require intellectual modification or highly technical enhancement.").

[57] A "joint work" is "a work prepared by two or more authors with the intention that their contributions be merged into inseparable or interdependent parts of a unitary whole." 17 U.S.C. § 101 (defining "joint work"). Because an AI system is not a human being, it cannot be considered an "author" in collaboration with a user. *See* Kernochan Center for Law, Media and the Arts ("Kernochan Center"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6–9 (Oct. 30, 2023) ("Kernochan Center Initial Comments") (noting that machines are not "authors" within the meaning of the Copyright Act, nor are they capable of forming an intention to merge their output with the contributions from the user that interacts with these systems).

[58] *See* The Authors Guild, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 33 (Oct. 30, 2023) ("The Authors Guild Initial Comments") ("Areas of the law that will instruct courts in how to determine what is copyrightable in an AI-assisted human-created work or human-assisted AI-generated material include . . . joint work cases where the issue of whether a secondary creator contributed a sufficient amount to rise to the level of an author . . . .").

[59] *Brownstein v. Lindsay*, 742 F.3d 55, 64 (3d Cir. 2014) ("For two or more people to become co-authors, each author must contribute some non-trivial amount of creative, original, or intellectual expression to the work and both must intend that their contributions be combined."); *Ashton-Tate Corp. v. Ross*, 728 F. Supp. 597 (N.D. Cal. 1989) (finding that a contribution to a joint work must be protectable in itself and that only expressions of ideas, not ideas themselves, give rise to protected interest), *aff'd*, 916 F.2d 516, 521 (9th Cir. 1990).

commissioned work should do or look like is not a joint author for purposes of the Copyright Act."[60]

The following sections apply these legal principles in the context of generative AI systems. After describing uses of computer-assisted tools in the creation process, we discuss the following three kinds of human contribution to AI-generated outputs: (1) prompts that instruct an AI system to generate an output; (2) expressive inputs that can be perceived in AI-generated outputs; and (3) modifications or arrangements of AI-generated outputs.[61]

C. Assistive Uses of AI Systems

Many commenters expressed concern about continuing the longstanding and growing use of computer-assisted tools in the creation process.[62] They pointed to various tasks that have been performed in creative fields for years, some of which now incorporate recent developments in AI, such as "aging" or "de-aging" actors, identifying chord progressions, detecting errors in software code, and removing unwanted objects or crowds from a scene.[63]

---

[60] *Payday*, 886 F.2d at 1087; *see also Sullivan v. Flora, Inc.*, 936 F.3d 562 (7th Cir. 2019) (upholding jury finding that plaintiff and defendant were not joint authors of illustrations because defendant merely offered suggestions on color, style, and text and rough outlines and sketches to guide the plaintiff's work, while the plaintiff used digital design software to create the illustrations, sometime incorporating defendant's suggestions and other times not); *BancTraining Video Sys. v. First American Corp.*, No. 91-cv-5340, 1992 U.S. App. LEXIS 3677, at *12 (6th Cir. 1992) ("Providing sketches, ideas or supervision over copyrightable material is not sufficient to make one a joint author.").

[61] Of course, many cases may involve a combination of two or more of these types of contributions. For example, a user could make creative modifications to an output generated using their own expressive input and multiple prompts.

[62] Commenters from the music industry noted that musicians and sound engineers have used such tools for many years, citing Autotune as one example. Songwriters of North America, et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 30, 2023); *see also* Recording Academy, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1 (Oct. 30, 2023) ("Recording Academy Initial Comments"). In the software industry, programmers and computer engineers use automated tools to modify software code, such as to perform refactoring and translate from one programming language into another. Apple Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3–4 (Oct. 30, 2023) ("Apple Initial Comments").

[63] For example, commenters reported that musicians are beginning to use AI systems for developing beats or mixing a track. *See* Recording Academy Initial Comments at 3; *see also* Universal Music Group ("UMG"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5–7 (Oct. 30, 2023) ("UMG Initial Comments"); Dina LaPolt, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023) ("Dina LaPolt Initial Comments"). Motion picture companies use AI tools as part of their creative process, particularly in the context of visual effects and post-production. For example, these tools may be used for color correction, detail sharpening, or de-blurring. Motion Picture Association ("MPA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 37–38 (Oct. 30, 2023) ("MPA Initial Comments"); *see also* Holton Lemaster, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Aug. 31, 2023) ("AI as a support tool for artists who choose to use them in their creation pipeline is fine. Crowd removal from photos, video stabilization tools, and ray tracing are all tools that really shine when enhanced by AI."). AI tools are frequently used in a process called rotoscoping, a time-consuming task that involves "altering individual frames within a single shot to align live-action and computer-generated images." MPA Initial Comments at 6, 37–38.

Commenters argued that these types of uses of AI should not affect the availability of copyright protection for the output.[64]

The Office agrees that there is an important distinction between using AI as a tool to assist in the creation of works and using AI as a stand-in for human creativity. While assistive uses that enhance human expression do not limit copyright protection, uses where an AI system makes expressive choices require further analysis. This distinction depends on how the system is being used, not on its inherent characteristics.[65]

Commenters also identified situations where creators have begun to experiment with using AI as a brainstorming tool. The Recording Academy, for instance, stated that "[m]any Academy members already use generative AI as a tool to assist them in creating new music," including through song ideation.[66] Another stakeholder noted that AI can be used to structure or create a preliminary outline for literary works.[67] In these cases, the user appears to be prompting a generative AI system and referencing, but not incorporating, the output in the development of her own work of authorship. Using AI in this way should not affect the copyrightability of the resulting human-authored work.[68]

D. Prompts

1. Commenters' Views

Many of the comments received in response to the NOI focused on the legal implications of creating outputs by providing prompts to an AI system. At the outset, as several

[64] *See, e.g.*, Intellectual Property Owners Association ("IPO"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6–7 (Oct. 30, 2023) ("IPO Initial Comments") ("[I]t is desirable to provide copyright protection for works resulting from a human using an AI system as a tool of creativity and where that human activity satisfies the traditional requirements of human authorship. A lack of this protection would be detrimental to rights holders and creators alike.").

[65] One commenter urged the Office to adopt a distinction based on the type of AI platform a user employs. Scenario, Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 18, 2023) ("Scenario Initial Comments") (arguing that output generated by a multimodal generative AI platform should presumptively be deemed copyrightable, while output generated by a unimodal generative AI platform should presumptively be deemed uncopyrightable).

[66] Recording Academy Initial Comments at 10.

[67] Literary Works Listening Session Tr. at 31:18–23 (Apr. 19, 2023) (statement by Mary Rasenberger, The Authors Guild).

[68] Other examples of such uses provided by commenters include digital and copy editing and other uses that "are intended to assist, not displace, human creativity." Recording Academy Initial Comments at 3; Lori Wilde, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Oct. 24, 2023); IPO Initial Comments at 2; Authors Alliance, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Oct. 30, 2023) ("Authors Alliance Initial Comments").

commenters noted, prompts themselves, if sufficiently creative, may be copyrightable.[69] The copyright status of the output generated, however, is a separate question.[70]

Most commenters agreed that inputting simple prompts is insufficient to make a user the author of the AI-generated output.[71] Several described prompts as unprotectible ideas or instructions.[72] The American Society of Composers, Authors and Publishers ("ASCAP"), a performing rights organization, asserted that "[w]here a human's involvement is limited to the simple generation of minimal queries and prompts for an AI tool, the resulting material is not entitled to copyright protection."[73] The Brooklyn Law Incubator & Policy Clinic asserted that a simple, general prompt lacks "enough human creativity for the output to qualify for copyright protection."[74] Universal Music Group ("UMG") stated: "The prompting user is no more an

---

[69] *See* AI Registration Guidance at 16192 n.27; The Authors Guild Initial Comments at 32 n.39 (arguing that the creator of a prompt "has a copyright in the prompt assuming it has sufficient original expression"); American Association of Independent Music ("A2IM") and the Recording Industry Association of America, Inc. ("RIAA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 34 (Oct. 30, 2023) ("A2IM-RIAA Joint Initial Comments"); Daniel Gervais, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 30, 2023) ("Daniel Gervais Initial Comments").

[70] *See generally* A2IM-RIAA Joint Initial Comments at 34 ("While the text of those prompts may be independently copyrightable if sufficiently expressive, that does not confer upon the author of the prompt any copyright in the output generated by the AI system."); Johan Brandstedt, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 29 (Oct. 29, 2023) ("Johan Brandstedt Initial Comments").

[71] Commenters used "simple" with varying degrees of specificity, generally referring to prompts that contain only generic descriptions or a short number of words. *See, e.g.*, Donaldson Callif Perez, LLP, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Oct. 30, 2023) ("Donaldson Callif Perez Initial Comments") ("[W]e agree that simple prompts by humans that result in a complex, creative work should not be granted copyright protection."); Dina LaPolt Initial Comments at 7 (stating that "a user inputting a simple generic prompt" should not be able to claim copyright protection); Edward Lee, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Oct. 30, 2023) (stating that "a simple one- or two-word prompt" is unlikely to satisfy the minimum standard for copyright protection in the output); Peer Music and Boomy, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Oct. 30, 2023) ("Peer Music-Boomy Joint Initial Comments") (finding it difficult to imagine how a single prompt that produces a complex output could provide a basis for claiming copyright protection in the output).

[72] *See, e.g.*, Adobe Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5–6 (Oct. 30, 2023) ("Adobe Initial Comments") ("[A] prompt is not copyrightable because the prompt represents the idea."); Johan Brandstedt Initial Comments at 19 (stating that "prompts express *ideas*, image and text generators provide stored *expression*"); European Writers' Council ("EWC"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 16 (Oct. 30, 2023) ("EWC Initial Comments") (stating that "the person formulating the prompts [cannot] claim any rights with respect to the results on the basis of the prompts alone, because the mere formulation of the task and the choice between several results proposed by the AI system is not a creative or protectable act").

[73] ASCAP, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 48–49 (Oct. 30, 2023) ("ASCAP Initial Comments").

[74] Brooklyn Law Incubator & Policy Clinic ("BLIP"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 20 (Oct. 30, 2023) ("BLIP Initial Comments"); *see also* Qualcomm Incorporated, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Dec. 6, 2023) ("Qualcomm Reply Comments") (stating that output "based on a single, general prompt with *de minimis* creativity" lacks "requisite human expression").

author than someone who tells a musician friend to 'write me a pretty love song in a major key' and then falsely claims co-ownership."[75]

By contrast, other commenters disputed the notion that prompts merely "influence" the AI system and do not provide "specific instructions to create a particular expressive result."[76] For example, the Intellectual Property Owners Association stated that "[i]f a user prompts Midjourney to produce an image or series of images of a city scape under water, the user is going to get a city scape under water."[77]

Commenters' views on more detailed prompts, including those that are revised and repeated, varied. Some viewed highly detailed prompts as sufficient to make some AI-generated outputs copyrightable.[78] Professors Pamela Samuelson, Christopher Jon Sprigman, and Matthew Sag stated that "[s]ophisticated prompts that specify details of an image should be sufficient to meet the [human authorship] requirement," and that "[a] person who instructs a Generative AI with enough detail, such that model output reflects that person's original

[75] Letter from UMG, Summary of *Ex Parte* Meeting on Apr. 22, 2024, Regarding the Office's AI Study, to U.S. Copyright Office 11 (Dec. 3, 2024) (arguing that "users prompting [music generative AI companies] to generate audio files are not composing or writing anything, much less 'their own, original music'" and instead are "simply supply[ing] an uncopyrightable idea in a text prompt . . . and the software itself generates an audio track based on its own predictive algorithms").

[76] IPO Initial Comments at 5; Van Lindberg, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 41 (Oct. 30, 2023) ("Van Lindberg Initial Comments") ("Randomness is part of the generative process—but *the output of an AI model is not random*. A human using the AI system typically describes what should be generated and/or provides other inputs that are used to initialize and guide the generative process."); Ashley Greenwald, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Oct. 30, 2023) (arguing that interior designers initiate generative AI systems by "giv[ing] certain prompts and instructions," refining and modifying interim results, and "mak[ing] the *final determination* whether and how the output co-created with the help of generative AI tools should be utilized"); Christa Laser, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Oct. 30, 2023) ("Christa Laser Initial Comments") ("A few uses of generative AI employ random strings and undirected outcomes, but a more significant role of generative AI is to implement a human's extensive creativity, direction, and selection towards an outcome pre-dreamed in the human mind.").

[77] IPO Initial Comments at 5.

[78] *See* BLIP Initial Comments at 23 (stating that users "may provide very detailed and extensive prompts to an AI-system to ensure that its output is as close as possible to what they anticipated" and such outputs should be copyrightable if "they provided sufficient input and prompts to control the output of an AI system"); Van Lindberg Initial Comments at 42 (stating that "the more information that is given within the prompt, the more control is exerted over the output"); Law Office of Seth Polansky LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 26 (Oct. 12, 2023) ("Seth Polansky Initial Comments") ("[A] human who closely guides the output of a generative AI system through curated training or by providing detailed prompts may be able to claim some form of 'joint authorship' with the machine."); Donaldson Callif Perez Initial Comments at 2 ("[I]f someone spends a significant amount of effort creating very specific and detailed prompts to create a complex work, perhaps there should be some copyright protection for that work.").

conception of the work, should be regarded as the author of the resulting work."[79] Another commenter asserted that, with detailed prompts, users "can achieve remarkable control over the expressive elements of the work, such as lighting, pose, style, expressions, and setting."[80]

In contrast, the Authors Guild argued that the unpredictability of the prompt-to-output generation process may make it "difficult to show that there was sufficient control and consequently a sufficient closeness between 'conception and execution.'"[81] Others agreed.[82] Adobe, for instance, stated that "[w]hen you submit a prompt (or idea), you then receive an output based solely on the AI's interpretation of that prompt," and the "AI's expression of [that] idea is not copyrightable."[83]

Several commenters described AI systems as black boxes,[84] meaning that not only do users in most cases not know what "will inform the [AI's] response" to prompts,[85] but that even developers of AI systems cannot generally predict outputs or explain why they include certain

---

[79] Pamela Samuelson et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Oct. 30, 2023) ("Pamela Samuelson et al. Initial Comments"); *see also* MPA Initial Comments at 47 (predicting that prompts could become "much more detailed" as technology improves to the point where "the inputs themselves may provide the substantive content for the output" and concluding that "[a] rule that prompts would never satisfy the human authorship requirement neglects those likely possibilities").

[80] Christa Laser Initial Comments at 5. Several other commenters cautioned that while there may "be cases where the prompts are so directive and detailed" that the user could be entitled to copyright protection for the output, this is likely to be rare. The Authors Guild Initial Comments at 32; *see also* Daniel Gervais Initial Comments at 6 (describing as "exceptional" cases "in which a detailed prompt . . . could contain expressions of specific ideas that reflect human creative choices directly perceptible in the machine's output").

[81] The Authors Guild Initial Comments at 31.

[82] *See* Association of Medical Illustrators, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 8–9 (Oct. 30, 2023); Kernochan Center Initial Comments at 5–6.

[83] Adobe Initial Comments at 5–6; *see also* Johan Brandstedt Initial Comments at 14, 29; EWC Initial Comments at 16.

[84] *See, e.g.*, Professional Photographers of America, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023); SeaQVN, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 63 (Sept. 13, 2023); IAC Inc. and Dotdash Media Inc., d/b/a Dotdash Meredith, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Oct. 30, 2023); Eric Bourdages, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1 (Nov. 26, 2023); James Horvath, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1 (Sep. 13, 2023); Cooper Reid, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Aug. 31, 2023).

[85] Kernochan Center Initial Comments at 5; *see also* Gabriel Moise, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Aug. 31, 2023); The Authors Guild Initial Comments at 31–32; Vikas Hassija et al., *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 COGNITIVE COMPUTATION 45, 47 (2024), https://link.springer.com/article/10.1007/s12559-023-10179-8 (noting that "the internal workings of [a black-box] model are not easily accessible or interpretable" and that this "lack of transparency" makes it difficult "to understand the model's behavior").

elements and not others.[86] Some provided examples of prompts containing detailed descriptions of what the user had in mind, where the output omitted some elements requested and inserted others.[87]

Commenters also noted that prompts are often entered into an AI system in one medium (such as text) and the output is generated in a different medium (such as a visual image, video, or audio clip). Several commenters asserted that moving from one medium to another requires interpretation, and where AI provides that interpretation, the user's control over the execution of their idea is indirect.[88] UMG highlighted one popular text-to-music generator that cautions users, "[n]o matter how detailed[,] text prompts cannot fully define an actual piece of music."[89]

Some stressed that generative AI systems can produce a seemingly limitless number of variations in response to the same prompt, no matter how many times that prompt is used.[90] The Kernochan Center argued that "[e]xtending the scope of copyright protection in the written prompts to cover the multiplicity of potential outputs" that may be generated by an AI system "comes uncomfortably close to conferring a copyright in a method of generating images (or other works)," which would be prohibited under section 102(b).[91]

---

[86] *See* EWC Initial Comments at 9 ("In computer science, PROCESSING (computation) is consistently described as a black box; not even the operators of AI systems know exactly what happens during the learning process—and they do not control it."); *see also supra* Section II.A.

[87] *See* Kernochan Center Initial Comments at 8–9 & n.13 (noting that "even highly elaborated prompts will . . . yield multiple outputs (not all of them fully or accurately responsive to the prompts)" and providing examples). *See also* Tonio Inverness, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Sept. 12, 2023) (demonstrating labor that goes into refining prompts after the results of initial prompt were "not at all what [commenter] had in mind"); UMG Initial Comments at 76–77.

[88] Johan Brandstedt Initial Comments at 14, 19 (stating that "anything started in writing ought not to merit copyright claims over an image"); Kernochan Center Initial Comments at 8 (stating that a textual description "would need to evince an extremely high degree of precision" in order to claim copyright in a pictorial work produced through the use of those instructions); The Authors Guild Initial Comments at 31 n.36 (stating that converting a "text instruction to images created from training data makes the output unpredictable").

[89] Letter from UMG, Summary of *Ex Parte* Meeting on Apr. 22, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 3 (Dec. 3, 2024) (internal citation omitted); *see also How do I make music with Udio?*, UDIO, https://www.udio.com/guide (last visited Jan. 17, 2025) (explaining that prompts cannot fully define an output because "the same text describes an infinite number of possible audio tracks").

[90] *See, e.g.*, Kernochan Center Initial Comments at 8–9; The Authors Guild Initial Comments at 32 n.39.

[91] Kernochan Center Initial Comments at 8–9; *see also* 17 U.S.C. § 102(b) (excluding from copyright protection "any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work").

A few commenters asserted that human-directed revisions to prompts may result in greater control over an output's expressive elements.[92] One technique entails submitting a prompt to the AI system, then revising the prompt, either by adding, removing, or replacing certain terms based on the initial output produced, to generate a new output. The user may revise and repeat upwards of hundreds of times.[93] Eventually the system may generate an output that meets the user's needs; if not, the user may decide to revise the prompt again or abandon the effort. Commenters noted that this process can require a significant amount of time and "demonstrable human effort."[94]

Some commenters advanced a theory of "authorship by adoption" (though few used that phrase).[95] They suggested that a user may exercise creative judgment when deciding to accept the output produced by a generative AI system. One suggested that a user who "repeatedly enters prompts until the output matches their desired expression" is no different than an "artist who continues to dab paint on the canvas until the image matches the painter's vision."[96] In contrast, the Authors Guild likened repetitive prompting to "spinning a roulette

---

[92] *See, e.g.*, Evangelical Christian Publishers Association ("ECPA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 8 (Oct. 30, 2023) ("ECPA Initial Comments") ("If the issue is one of control and predictability, fine-tuning repeatedly until the final expression is satisfactory demonstrates the author's ultimate control of the final work, even if each iteration leading up to the final expression may be subject to unpredictability."); SCA Robotics, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Sept. 29, 2023) ("SCA Robotics Initial Comments") (stating that authorship should depend on factors such as "the human user's control of the artistic expression outputted by the platform," including "the extent of the human party's discretion over accepting and/or modifying the outputted work"); International Center for Law & Economics, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 18 (Oct. 30, 2023) ("AI systems remain tools that require human direction and judgment. As such, when a person provides the initial prompt or framing, makes choices regarding the iterative development of the AI output, and decides that the result is satisfactory for inclusion in a final work, they are fundamentally engaging in creative decision making that constitutes authorship under copyright law.").

[93] *See* IPO Initial Comments at 5 (noting that "[t]he same user might iterate on dozens, even hundreds, of prompts of greater complexity and specificity before achieving a desired result").

[94] Superframe, LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Sept. 6, 2023); *see also* AI and Metaverse Task Force of the Trust over IP Foundation, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 4 (Oct. 30, 2023); Donaldson Callif Perez Initial Comments at 2.

[95] This theory would find authorship in the decision to adopt something unplanned or unexpected occurring in the course of creating a work. *See* Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343 (2019). It can be traced to *Alfred Bell & Co. v. Catalda Fine Arts*, which assessed the originality of mezzotint engravings that were based on paintings in the public domain. 191 F.2d 99, 104 (2d Cir. 1951). The defendant argued that the engravings were mere copies of preexisting paintings, and therefore not protected by copyright. *Id.* In finding that the engraver's versions were sufficiently different, the court speculated that "[a] copyist's bad eyesight or defective musculature, or a shock caused by a clap of thunder, may yield sufficiently distinguishable variations." *Id.* at 105. "Having hit upon such a variation unintentionally," the court held that "the 'author' may adopt it as his and copyright it." *Id.*

[96] ECPA Initial Comments at 7.

wheel with infinite possibilities."[97] It argued that "when a user [metaphorically] spins the wheel dozens of times until they land on an output they like," such activity should not give the user a right to claim ownership of that output.[98]

Discussing the authorship by adoption theory, Professors Jane Ginsburg and Luke Ali Budiardjo concluded that, "[w]ere post-execution adoption to substitute for any authorial participation, even indirect or inadvertent, in giving physical form to a work, then, in addition to [naming] the 'wrong' author, copyright law would effectively vest adopters with rights in ideas."[99] Professor Daniel Gervais made a similar point with the following analogy: "If I walk into a gallery or shop that specializes in African savanna paintings or pictures because I am looking for a specific idea (say, an elephant at sunset, with trees in the distance), I may find a painting or picture that fits my idea," but "[t]hat in no way makes me an author."[100]

2. Analysis

The Office concludes that, given current generally available technology, prompts alone do not provide sufficient human control to make users of an AI system the authors of the output. Prompts essentially function as instructions that convey unprotectible ideas. While highly detailed prompts could contain the user's desired expressive elements, at present they do not control how the AI system processes them in generating the output.

Cases regarding joint authorship support this conclusion. These cases address the amount of control that is necessary to claim authorship. The provision of detailed directions, without influence over how those directions are executed, is insufficient.[101] As the Third Circuit explained, when a person hires someone to execute their expression, "that process must be rote or mechanical transcription that does not require intellectual modification or highly technical enhancement" for the delegating party to claim copyright authorship in the final work.[102] Although entering prompts into a generative AI system can be seen as similar to providing instructions to an artist commissioned to create a work, there are key differences. In a human-to-human collaboration, the hiring party is able to oversee, direct, and understand the contributions of a commissioned human artist. Depending on the nature of each party's contributions, the artist may be the sole author, or the outcome may be a joint work or work

---

[97] The Authors Guild Initial Comments at 31–32.

[98] *Id.*; *see also* Kernochan Center Initial Comments at 9 (asserting that "selection of a single output is not itself a creative act"); Daniel Gervais Initial Comments at 6–7; Johan Brandstedt Initial Comments at 29.

[99] Ginsburg & Budiardjo, *supra* note 95, at 370.

[100] Daniel Gervais Initial Comments at 7.

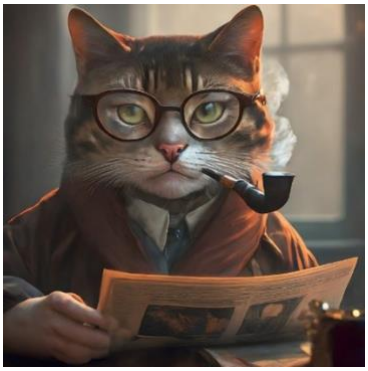
[101] *Payday*, 886 F.2d at 1087; *see, e.g.*, *CCNV*, 490 U.S. at 737 ("As a general rule, the author is the party who actually creates the work, that is, the person who translates an idea into a fixed, tangible expression entitled to copyright protection.").

[102] *Andrien*, 927 F.2d at 134–35.

made for hire.[103] In theory, AI systems could someday allow users to exert so much control over how their expression is reflected in an output that the system's contribution would become rote or mechanical.[104] The evidence as to the operation of today's AI systems indicates that this is not currently the case. Prompts do not appear to adequately determine the expressive elements produced, or control how the system translates them into an output.[105]

The gaps between prompts and resulting outputs demonstrate that the user lacks control over the conversion of their ideas into fixed expression, and the system is largely responsible for determining the expressive elements in the output. In other words, prompts may reflect a user's mental conception or idea, but they do not control the way that idea is expressed. This is even clearer in the case of generative AI systems that modify or rewrite prompts internally. That process recasts the human contribution—however detailed it may be—into a different form.

The following image, which the Office generated by entering a prompt into a popular commercially available AI system, illustrates this point:[106]

Prompt	***Output***
professional photo, bespectacled cat in a robe reading the Sunday newspaper and smoking a pipe, foggy, wet, stormy, 70mm, cinematic, highly detailed wood, cinematic lighting, intricate, sharp focus, medium shot, (centered image composition), (professionally color graded), ((bright soft diffused light)), volumetric fog, hdr 4k, 8k, realistic	

This prompt describes the subject matter of the desired output, the setting for the scene, the style of the image, and placement of the main subject. The resulting image reflects some of these instructions (*e.g.*, a bespectacled cat smoking a pipe), but not others (*e.g.*, a highly detailed wood environment). Where no instructions were provided, the AI system filled in the gaps.

---

[103] In contrast, AI systems cannot produce joint works or works made for hire because they are not "authors," they are not capable of forming an intention to merge their output with the user's contributions, and they cannot enter into binding contracts. *See* Kernochan Center Initial Comments at 7; Brief for Appellees, at 27, Thaler v. Perlmutter, No. 23-5233 (D.C. Cir. Mar. 6, 2024).

[104] This outcome would raise additional questions about the utility of AI in creative expression.

[105] *Cf. Geshwind*, 734 F. Supp. at 650–51 ("The fact that the agent, Geshwind, wanted changes in details and aspects of the portrait and even made suggestions, the compliance with which may or may not have improved the effect, does not make him the creator."); *M.G.B. Homes*, 903 F.2d at 1493; *Payday*, 886 F.2d at 1087.

[106] The Office used Google's generative AI chatbot Gemini to generate this image. GEMINI, https://gemini.google.com/ (last visited Jan. 17, 2025).

For instance, the prompt does not specify the cat's breed or coloring, size, pose, any attributes of its facial features or expression, or what clothes, if any, it should wear beneath the robe. Nothing in the prompt indicates that the newspaper should be held by an incongruous human hand.

The fact that identical prompts can generate multiple different outputs further indicates a lack of human control.[107] As one popular system explains on its website, "[n]o matter how detailed . . . the same text describes an infinite number of possible" outputs.[108] In these circumstances, the black box of the AI system is providing varying interpretations of the user's directions.

Repeatedly revising prompts does not change this analysis or provide a sufficient basis for claiming copyright in the output. First, the time, expense, or effort involved in creating a work by revising prompts is irrelevant, as copyright protects original authorship, not hard work or "sweat of the brow."[109] Second, inputting a revised prompt does not appear to be materially different in operation from inputting a single prompt. By revising and submitting prompts multiple times, the user is "re-rolling" the dice, causing the system to generate more outputs from which to select, but not altering the degree of control over the process.[110] No matter how many times a prompt is revised and resubmitted, the final output reflects the user's acceptance of the AI system's interpretation, rather than authorship of the expression it contains.

Some commenters drew analogies to a Jackson Pollock painting or to nature photography taken with a stationary camera, which may be eligible for copyright protection even if the author does not control where paint may hit the canvas or when a wild animal may step into the frame.[111] However, these works differ from AI-generated materials in that the human author is principally responsible for the execution of the idea and the determination of the expressive elements in the resulting work. Jackson Pollock's process of creation did not end with his vision of a work. He controlled the choice of colors, number of layers, depth of texture, placement of each addition to the overall composition—and used his own body movements to execute each of these choices. In the case of a nature photograph, any copyright protection is based primarily on the angle, location, speed, and exposure chosen by the photographer in

---

[107] *See supra* note 32. The Office re-ran the prompt above and received a much different image of a cat in a stormy setting.

[108] *How do I make music with Udio?*, UDIO, https://www.udio.com/guide (emphasis omitted) (last visited Jan. 17, 2025).

[109] *Feist*, 499 U.S. at 352.

[110] *See, e.g.*, Kernochan Center Initial Comments at 8 ("If each prompt newly rolls the dice, it is difficult to discern the dominance of will that 'direction' implies, and thus hard to classify it as meeting the requirement of an objective 'intent.'").

[111] *See, e.g.*, Tim Boucher, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 8 (Oct. 26, 2023); Christa Laser Initial Comments at 4; MPA Initial Comments at 47–50; Pamela Samuelson et al. Initial Comments at 4.

setting up the camera, and possibly post-production editing of the footage.[112] As one commenter explained, "some element of randomness does not eliminate authorship," but "the putative author must be able to constrain or channel the program's processing of the source material."[113] The issue is the degree of human control, rather than the predictability of the outcome.[114]

The Office also agrees that authorship by adoption does not in itself provide a basis for claiming copyright in AI-generated outputs. As commenters noted, providing instructions to a machine and selecting an output does not equate to authorship.[115] Selecting an AI-generated output among uncontrolled options is more analogous to curating a "living garden," than applying splattered paint.[116] As the Kernochan Center observed, "selection among the offered options" produced by such a system cannot be considered copyrightable authorship, because the "selection of a single output is not itself a creative act."[117]

There may come a time when prompts can sufficiently control expressive elements in AI-generated outputs to reflect human authorship. If further advances in technology provide users with increased control over those expressive elements, a different conclusion may be called for.[118] On the other hand, technological advancements that facilitate increased automation and optimization may bolster our current conclusions. For example, if generative

---

[112] Like other copyrighted works, nature photography must have a sufficient amount of creative expression to satisfy the originality standard.

[113] Kernochan Center Initial Comments at 5.

[114] *See* Digital Media Licensing Association ("DMLA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 16 (Oct. 30, 2023) ("DMLA Initial Comments") (stating that "the foreseeability of the AI's results may bear on authorship" in cases "where there is a limited range of specific expressive output that is objectively foreseeable as a result of a human user's prompt"); Kernochan Center Initial Comments at 5; MPA Initial Comments at 45–46 (acknowledging that evaluating "the elements of predictability and control may be appropriate in certain cases"); International Trademark Association ("INTA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Oct. 30, 2023) ("INTA Initial Comments") (acknowledging that if a program generated an image by simply populating "each pixel with a randomly-selected color, it seems obvious that the resulting work should not be considered a work of authorship"); The Authors Guild Initial Comments at 31.

[115] The Authors Guild Initial Comments at 31–32; Daniel Gervais Initial Comments at 6–7; Kernochan Center Initial Comments at 8.

[116] *Kelley v. Chicago Park Dist.*, 635 F.3d 290, 304 (7th Cir. 2011); *see also* COMPENDIUM (THIRD) § 306; *Thaler*, 687 F. Supp. 3d at 146 (holding that the "key" to copyright protection is "[h]uman involvement in, and ultimate creative control over, the work at issue").

[117] Kernochan Center Initial Comments at 9.

[118] S*ee* Authors Alliance Initial Comments at 19 ("[A]s both generative AI systems and the ways that creators use them change and evolve, the application of the human authorship requirement to content that is AI-generated or AI-assisted may also change. For example, if these tools developed in a way that would give creators more control over the outputs, works created with these tools could potentially be considered works of human authorship.").

AI systems integrate or further improve automated prompt optimization, users' control may be diminished.

E. Expressive Inputs

As discussed above, AI systems take inputs in the form of text, images, audio, video, or a combination of these mediums. Some systems—whether via tools, settings, or prompts—allow inputs to be substantially retained as part of the output. For example, one commenter noted that a human author may create an original illustration, input that work into an AI system, and instruct the system "to modify [the] color or layer portions of [the] existing image."[119] Another observed that an AI system may be used to modify or translate a copyrighted work,[120] such as uploading a story written in the first person and instructing the system to convert it to a third-person point of view.

These types of expressive inputs, while they may be seen as a form of prompts, are different from those that merely communicate desired outcomes. As commenters pointed out, where human-authored inputs are reflected in the output, they contribute more than just an intellectual conception. One explained that "a human author who inputs their own illustration or media file" into an AI system "may have a greater claim to authorship," because "there is a limited range of specific expressive output that is objectively foreseeable as a result of a human user's" contribution.[121] Another noted that when a user provides an input to an AI system such as "a traditional work created or designated by the user . . . the specified starting point constrains the 'autonomy' of the outputs" and thus may "present a more persuasive case of human intervention" than simply applying "prompts to an unknown starting point."[122]

---

[119] DMLA Initial Comments at 16.

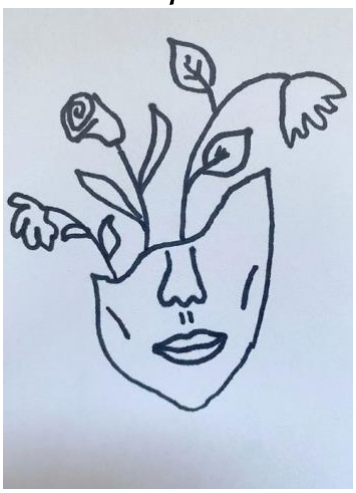
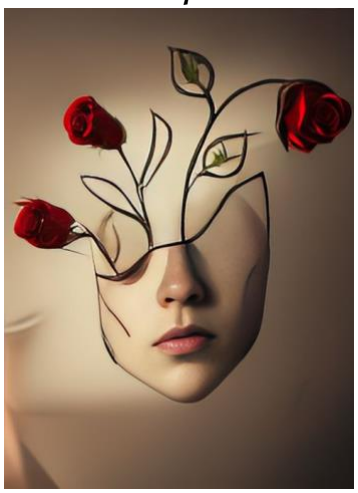
[120] Pearson, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023) ("Pearson Initial Comments").

[121] DMLA Initial Comments at 16; *see* Pearson Initial Comments at 7–8 (acknowledging that "copyright can only protect material that is the product of human activity" and stating that "further consideration should be given to whether a claim of authorship in output may exist where the input itself is a representation of the original intellectual conception of an author"); National Music Publishers' Association ("NMPA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 30 (Oct. 30, 2023) ("NMPA Initial Comments") ("Creators that use AI to refine, recast, or modify, or to create new derivative works based on their preexisting works may also have legitimate claims of authorship over the resulting work in some circumstances.").

[122] Kernochan Center Initial Comments at 5–6; *see* MPA Initial Comments at 50 (noting that "material human creators provide to the AI tool" such as "inputs, like a drawing or photo" can be considered "intellectual and creative contributions that are inseparable from the ultimate work").

As an example, in the following work submitted to the Office for registration, the author had created a hand-drawn illustration and used it as an input, along with the prompt shown below.[123]

The AI system produced this output:

Prompt	***Input***	***Output***
"a young cyborg woman (((roses))) flowers coming out of her head, photorealism, cinematic lighting, hyper realism, 8k, hyper detailed."		

The drawing itself is a copyrightable work, and its expressive elements are clearly perceptible in the output, including the outline of the mask, the position of the nose, mouth, and cheekbones relative to the shape of the mask, the arrangement of the stems and rosebuds, and the shape and placement of the four leaves.

The applicant disclaimed "any non-human expression" appearing in the final work, such as the realistic, three-dimensional representation of the nose, lips, and rosebuds, as well as the lighting and shadows in the background. After reviewing the information provided in the application, the Office registered the work with an annotation stating: "Registration limited to unaltered human pictorial authorship that is clearly perceptible in the deposit and separable from the non-human expression that is excluded from the claim."[124]

---

[123] *Rose Enigma*, VAu001528922 (Mar. 21, 2023). More about the artist Kris Kashtanova's creation of this work is available on their website. *Portfolio: Rose Enigma*, KRIS KASHTANOVA, https://www.kris.art/portfolio-2/rose-enigma (last visited Jan. 17, 2025).

[124] *Rose Enigma*, VAu001528922 (Mar. 21, 2023). By contrast, the Office's Review Board upheld a refusal to register an image produced by an AI system with a human author's photograph as an input. U.S. Copyright Office Review Board, *Decision Affirming Refusal of Registration of Suryast* at 1 (Dec. 11, 2023), https://copyright.gov/rulings-filings/review-board/docs/SURYAST.pdf. The applicant disclosed that the image was generated by "RAGHAV Artificial Intelligence Painting App" ("RAGHAV"), which had been trained on Vincent van Gogh's *The Starry Night*—with an instruction to apply the style of *The Starry Night* to the photograph. *Id*. at 2. The Board found that the resulting image did not "contain sufficient human authorship necessary to sustain a claim to copyright" because the applicant "exerted insufficient creative control over RAGHAV's" generation of the output. *Id*. at 3, 7–8. Unlike *Rose Enigma*, the output did not clearly show the copyrightable work input by the applicant. *See id.* at 7–8.

As illustrated in this example, where a human inputs their own copyrightable work and that work is perceptible in the output, they will be the author of at least that portion of the output. Their own creative expression will be protected by copyright, with a scope analogous to that in a derivative work. Just as derivative work protection is limited to the material added by the later author,[125] copyright in this type of AI-generated output would cover the perceptible human expression. It may also cover the selection, coordination, and arrangement of the human-authored and AI-generated material, even though it would not extend to the AI-generated elements standing alone.

F. Modifying or Arranging AI-Generated Content

Generating content with AI is often an initial or intermediate step, and human authorship may be added in the final product. As explained in the AI Registration Guidance, "a human may select or arrange AI-generated material in a sufficiently creative way that 'the resulting work as a whole constitutes an original work of authorship.'"[126] A human may also "modify material originally generated by AI technology to such a degree that the modifications meet the standard for copyright protection."[127]

As several commenters noted, human authors should be able to claim copyright if they select, coordinate, and arrange AI-generated material in a creative way.[128] This would provide protection for the output as a whole (although not the AI-generated material alone).[129] A relatively common scenario in registration applications is the combination of human-authored text with AI-generated images. In one early case, for instance, the Office found that the selection and arrangement of AI-generated images with human-authored text in a comic book were protectable as a compilation. We explained:

---

[125] *See* H.R. REP. NO. 94-1476, at 57; S. REP. NO. 94-473, at 55 ("[C]opyright in a 'new version' covers only the material added by the later author, and has no effect one way or the other on the copyright or public domain status of the preexisting material.").

[126] AI Registration Guidance at 16192.

[127] *Id.* at 16192–93.

[128] *See, e.g.*, BLIP Initial Comments at 20; Center for Art Law, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 9 (Oct. 26, 2023); Cisco Systems, Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023) ("Cisco Initial Comments"); IPO Initial Comments at 4–6; Peer Music-Boomy Joint Initial Comments at 12.

[129] *See Feist*, 499 U.S. at 348 (noting that copyright protection for a compilation "may extend only to those components of a work that are original to the author").

> [T]he Office finds that the compilation of these images and text throughout the Work contains sufficient creativity under *Feist* to be protected by copyright. Specifically, the Office finds the Work is the product of creative choices with respect to the selection of the images that make up the Work and the placement and arrangement of the images and text on each of the Work's pages. Copyright therefore protects [the applicant's] authorship of the overall selection, coordination, and arrangement of the text and visual elements that make up the Work.[130]

Multiple similar registrations have been made since then.[131]

A number of commenters also made the point that if a user edits, adapts, enhances, or modifies AI-generated output in a way that contributes new authorship, the output would be entitled to protection.[132] They argued that these modifications "should be assessed in the same way as . . . editorial or other changes to a pre-existing work."[133] Although such works would not technically qualify as "derivative works,"[134] derivative authorship provides a helpful analogy in identifying originality. Again, the copyright would extend to the material the human author contributed but would not extend to the underlying AI-generated content itself.[135]

---

[130] U.S. Copyright Office, *Cancellation Decision re: Zarya of the Dawn (VAu001480196)* at 5 (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf.

[131] *See supra* notes 15, 123.

[132] *See, e.g.*, Apple Initial Comments at 1; ASCAP Initial Comments at 49; The Authors Guild Initial Comments at 32; BLIP Initial Comments at 25; Cisco Initial Comments at 7; Graphic Artists Guild, Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 19 (Oct. 30, 2023) ("Graphic Artists Guild Initial Comments"); OpenAI, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 15 (Oct. 30, 2023).

[133] Kernochan Center Initial Comments at 6.

[134] A derivative work is "a work based upon one or more preexisting works." 17 U.S.C. § 101 (defining "derivative work"). Because entirely AI-generated outputs do not contain the human authorship required to be a "work of authorship," the modified versions cannot qualify under this definition. *See* H.R. REP. NO. 94-1476, at 57; S. REP. NO. 94-473, at 55 (noting that "the 'pre-existing work' must come within the general subject matter of copyright set forth in section 102, regardless of whether it is or was ever copyrighted").

[135] *See* H.R. REP. NO. 94-1476, at 57; S. REP. NO. 94-473, at 55.

Many popular AI platforms offer tools that encourage users to select, edit, and adapt AI-generated content in an iterative fashion. Midjourney, for instance, offers what it calls "Vary Region and Remix Prompting," which allow users to select and regenerate regions of an image with a modified prompt. In the "Getting Started" section of its website, Midjourney provides the following images to demonstrate how these tools work.[136]

(1) Generate Candidate Images with Prompt: *meadow trail lithograph*

(2) Select and Upscale Image

(3) Use Freehand Editing Tool to Select Region

(4) Generate Candidate Images with Prompt*:* *meadow stream lithograph*

(5) Select and Upscale Image

The image was further modified by repeating the editing process:

Other generative AI systems also offer tools that similarly allow users to exert control over the selection, arrangement, and content of the final output.[137]

[136] *Vary Region + Remix*, MIDJOURNEY, https://docs.midjourney.com/docs/vary-region-remix (last visited Jan. 17, 2025). Text descriptions below each image were added by the Office.

[137] OpenAI's ChatGPT, for instance, has a feature called "canvas," which provides an interactive interface for users to "collaborate" with the model while writing a document or code. Users can edit AI-generated text; highlight regions for the model to focus on; use built-in tools to request in-line suggestions, length adjustments, and changes to the reading level; and write instructions that detail particular edits to be made. *See Introducing Canvas*, OPENAI (Oct. 3, 2024), https://openai.com/index/introducing-canvas/.

Unlike prompts alone, these tools can enable the user to control the selection and placement of individual creative elements. Whether such modifications rise to the minimum standard of originality required under *Feist* will depend on a case-by-case determination.[138] In those cases where they do, the output should be copyrightable.

Similarly, the inclusion of elements of AI-generated content in a larger human-authored work does not affect the copyrightability of the larger human-authored work as a whole.[139] For example, a film that includes AI-generated special effects or background artwork is copyrightable, even if the AI effects and artwork separately are not.

[138] The selection, coordination, and arrangement of only two or three elements is not generally sufficient for copyright protection. *See* COMPENDIUM (THIRD) § 312.2 ("[T]he Office generally will not register a compilation containing only two or three elements, because the selection is necessarily *de minimis*." (citing H.R. REP. NO. 94-1476, at 122 (stating that a work does not qualify as a collective work "where relatively few separate elements have been brought together," as in the case of "a composition consisting of words and music, a work published with illustrations or front matter, or three one-act plays"))).

[139] *Cf.* AI Registration Guidance at 16192–93.

III. INTERNATIONAL APPROACHES

Other countries are also analyzing whether copyright protection should extend to works containing AI-generated material. Those that have addressed this issue so far have agreed that copyright requires human authorship.

The Korean Copyright Commission and the Ministry of Culture, Sports and Tourism issued *A Guide on Generative AI and Copyright* in 2023, in which it explained that "only a natural person can become an author"[140] and that "copyright registration for an AI output is impossible if a human did not contribute creatively to the expressive form."[141] The Korean guidance noted that "if a human had performed *additional work* on the AI output, such as modifying, or making additions or deletions, only the part that had undergone such change is copyrightable."[142] It also stated that an author can register a work as a compilation if he or she selected and rearranged the AI output creatively.[143]

In Japan, the Copyright Subdivision of the Cultural Council published a summary of its guidelines in May 2024 titled *General Understanding on AI and Copyright in Japan*.[144] The guidelines explained that the copyrightability of AI-generated content will be determined on a case-by-case basis, depending on the following factors: (1) the amount and content of the instructions and input prompts by the AI user; (2) the number of generation attempts; (3) the selection by the AI user from multiple output materials; and (4) any subsequent human additions and corrections to the AI-generated work.[145]

In the People's Republic of China, the Beijing Internet Court evaluated arguments in a copyright infringement case involving an AI-generated work in 2023, starting with the premise that human authorship was required for copyright protection.[146] It found that an image created

---

[140] Ministry of Culture, Sports and Tourism & Korea Copyright Comm'n, *A Guide on Generative AI and Copyright*, at 40 (2023), https://www.copyright.or.kr/eng/doc/etc_pdf/Guide_on_Generative_AI_and_Copyright.pdf.

[141] *Id.* at 41.

[142] *Id.*

[143] *Id.* It has been reported that a copyright registration was granted in December 2023 for an AI-generated film based on the "human editing of the AI[-]generated film and images." Edward Lee, *South Korea grants copyright to AI generated work, 'AI Suro's Wife' film as work edited by humans*, CHATGPT IS EATING THE WORLD (Jan. 8, 2024), https://chatgptiseatingtheworld.com/2024/01/08/south-korea-grants-copyright-to-ai-generated-work-ai-suros-wife-film-as-work-edited-by-humans/.

[144] Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* (May 2024), https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.

[145] *Id.* at 17.

[146] Li Mou Mou Su Liu Mou Mou Qin Hai Zuo Pin Shu Ming Quan, Xin Xi Wang Luo Chuan Bo Quan Jiu Fen An (李某某诉刘某某侵害作品署名权, 信息网络传播权纠纷案) [Li v. Liu, Dispute over Copyright Infringement of the Right of Attribution and Right of Information Network Distribution of Works], at 14 (Beijing Internet Ct. Nov. 27, 2023), https://english.bjinternetcourt.gov.cn/pdf/BeijingInternetCourtCivilJudgment112792023.pdf. Page numbers in this Report are based on the English translation released by the Beijing Internet Court online.

using Stable Diffusion was protected under China's copyright law,[147] and that the person who used AI to create the image was the author.[148] According to the court, the selection of over 150 prompts combined with subsequent adjustments and modifications demonstrated that the image was the result of the author's "intellectual achievements," reflecting his personalized expression.[149]

In the European Union, the majority of member states agreed, in response to a 2024 policy questionnaire on the relationship between generative AI and copyright, that current copyright principles adequately address the copyright eligibility of AI outputs and there is no need to provide new or additional protection.[150] Member states also shared the view that AI-generated content may be eligible for copyright "only if the *human input in* [*the*] *creative process was significant*."[151] This consensus extended to the understanding that purely AI-generated works cannot be protected by copyright, as only a natural person can be considered an author.[152] Based on similar reasoning, in 2024, a court in Czechia, also known as the Czech Republic, held that an AI tool cannot be the author of a copyrighted work.[153]

In the United Kingdom, a statute predating the development of generative AI technologies protects works "generated by computer in circumstances such that there is no human author of the work."[154] It designates the author as a "person by whom the arrangements

---

[147] *Id.* at 10–14; *see also* Copyright Law of the People's Republic of China (promulgated by the Standing Comm. Nat'l Cong., Feb. 26th, 2010, effective Apr. 1, 2010), art. 3.

[148] *See supra* note 146 at 14–15. While the ruling is not precedential under Chinese judicial practice, it may inform policies and practices about the copyrightability of AI-generated art under Chinese law. *Id.* at 11–12. China has recently considered statutory clarifications for when a work generated by AI is protected under copyright. A preliminary draft of China's proposed AI law states that when a work generated using AI meets the conditions under the Copyright Law then it can be protected under that law "based on the extent of the user's contribution to the final presentation of the content." Zhong Hua Ren Min Gong He Guo Ren Gong Zhi Neng Fa (Xue Zhe Jian Yi Gao) (中华人民共和国人工智能法 (学者建议稿)) [Law of the People's Republic of China on Artificial Intelligence (Scholar's Draft Proposal)], art. 36, Official WeChat account of the Digi. Rule of Law Inst. at East China Univ. of Political Sci. and L., *translated by* Center for Sec. and Emerging Tech., https://cset.georgetown.edu/wp-content/uploads/t0592_china_ai_law_draft_EN.pdf.

[149] *See supra* note 146 at 11–12.

[150] Council of the European Union, *Policy questionnaire on the relationship between generative Artificial Intelligence and copyright and related rights – Revised Presidency summary of the Member States contributions*, at 16–18 (Dec. 20, 2024), https://data.consilium.europa.eu/doc/document/ST-16710-2024-REV-1/en/pdf.

[151] *Id.* at 16.

[152] *Id.* at 15.

[153] *See* Tomáš Ščerba & Jaroslav Fořt, *The first Czech case on generative AI*, TECH.'S LEGAL EDGE (Apr. 4, 2024), https://www.technologyslegaledge.com/2024/04/the-first-czech-case-on-generative-ai/; Alessandro Cerri, *Czech court finds that AI tool DALL-E cannot be the author of a copyright work*, THE IPKAT (Apr. 15, 2024), https://ipkitten.blogspot.com/2024/04/czech-court-finds-that-ai-tool-dall-e.html.

[154] Copyright, Designs and Patents Act 1988, c. X, I, §§ 178, 9(3) (UK), https://www.legislation.gov.uk/ukpga/1988/48/data.pdf. Protection lasts for fifty years from the date the work is made. *Id.*, c. I, § 12(7).

necessary for the creation of the work are undertaken."[155] In 2021, the United Kingdom Intellectual Property Office ("UKIPO") sought public comment on whether to change this law, in light of advancements in generative AI. Based on the lack of any case law applying this provision to AI,[156] the UKIPO concluded that "[i]t is unclear whether removing [protection for computer-generated works] would either promote or discourage innovation and the use of AI for the public good."[157] It elected to leave the law in place but did not rule out future changes.[158] Since then, the UK government has initiated a new consultation on copyright and AI, including questions about prompts, computer-generated works, and outputs of AI models.[159]

Several other former and current commonwealth countries, such as Hong Kong,[160] India,[161] and New Zealand,[162] have enacted similar provisions, but there too it is unclear whether or how they will apply to AI-generated works.

In Canada, a 2021 review of the Copyright Act acknowledged a lack of clarity concerning the authorship of an AI-generated work.[163] While the Standing Committee on Industry, Science and Technology, which led the review, recommended that legislation should

---

[155] *Id.*, c. I, § 9(3).

[156] UKIPO, *Consultation outcome of the Intell. Prop. Office on Artificial Intelligence and Intellectual Property: copyright and patents*, ¶ 22 (June 28, 2022), https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation#copyright-in-computer-generated-works.

[157] *Id.* ¶ 29.

[158] *Id.* ¶¶ 29–30.

[159] *See* UKIPO, *Open Consultation of the Intell. Prop. Office on Copyright and Artificial Intelligence* (Dec. 17, 2024), https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence#bcopyright-and-artificial-intelligence.

[160] Section 11(3) of Hong Kong's Copyright Ordinance states: "In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author is taken to be the person by whom the arrangements necessary for the creation of the work are undertaken." Copyright Ordinance, (2019) Cap. 528, § 11(3) (H.K.).

[161] Section 2(d)(vi) of India's Copyright Act defines author as "in relation to any literary, dramatic, musical or artistic work which is computer-generated, the person who causes the work to be created." The Copyright Act, 1957, § 2(d)(vi). Without citing that section, in 2020 the Indian Copyright Office registered the AI-generated work described in note 124, listing the AI tool as a co-author, but a year later issued a notice of withdrawal of the registration. Sukanya Sarkar, *Exclusive: Indian Copyright Office issues withdrawal notice to AI co-author*, MANAGINGIP (Dec. 13, 2021), https://www.managingip.com/article/2a5d0jj2zjo7fajsjwwlc/exclusive-indian-copyright-office-issues-withdrawal-notice-to-ai-co-author.

[162] Section 5(2)(a) of New Zealand's copyright law defines author as "in the case of a literary, dramatic, musical, or artistic work that is computer-generated, the person by whom the arrangements necessary for the creation of the work are undertaken." Copyright Act 1994, s 5(2)(a).

[163] Innovation, Sci. and Econ. Dev. Canada ("ISED Canada"), *A Consultation on a Modern Copyright Framework for Artificial Intelligence and the Internet of Things*, at 12 (2021), https://ised-isde.canada.ca/site/strategic-policy-sector/sites/default/files/attachments/2022/ConsultationPaperAIEN.pdf.

provide greater clarity, the Canadian Parliament has not yet acted on the recommendation.[164] In 2023, Canada relaunched the consultation process, with a comment period that closed in January 2024.[165]

Similarly, in Australia, participants in 2024 consultations held by the Select Committee on Adopting Artificial Intelligence shared concerns over the lack of clarity in Australia's copyright laws regarding the "extent of copyright protection, if any, that is afforded to works created by humans with the assistance or augmentation of AI."[166] The Select Committee in its recommendations, however, did not specifically address this issue or suggest any action.

Although some level of consensus on the need for human authorship appears to be emerging, and most countries have so far continued to apply existing law, it is clear that views are still being formed. It remains to be seen how copyrightability standards will be interpreted and applied. The Office is closely monitoring developments abroad and evaluating how other countries' evolving approaches may ultimately overlap or differ from our own.

---

[164] *Id.* at 13.

[165] ISED Canada, *Consultation on Copyright in the Age of Generative Artificial Intelligence* (2021), https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-copyright-age-generative-artificial-intelligence.

[166] Select Committee on Adopting Artificial Intelligence, Parliament of Australia (Final Report, November 2024) ¶ 4.166, https://parlinfo.aph.gov.au/parlInfo/download/committees/reportsen/RB000470/toc_pdf/SelectCommitteeonAdoptingArtificialIntelligence(AI).pdf.

IV. THE ARGUMENTS FOR LEGAL CHANGE

A. Providing Incentives

Commenters generally stressed the value of incentives to produce new works of authorship.[167] They differed, however, in their interpretations of the Copyright Clause and their assessment of the impact of providing such incentives for AI-generated content.

Those supporting copyright protection for AI-generated material contended that it would encourage the creation of more works, furthering progress in culture and knowledge to the benefit of the public.[168] They took the position that the Copyright Clause should be read flexibly to encompass new technologies.[169] For instance, one commenter argued that this interpretation should "evolve with technological advancements" to ensure that "the spirit of this mandate continues to foster innovation and artistic expression in all its forms."[170]

Most commenters that opined on this issue, however, agreed with the Office's view that the Copyright Clause requires human authorship.[171] They supported the conclusion that AI-

---

[167] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 4 n.11 (quoting *Thaler v. Perlmutter*, No. 22-cv-1564, 2023 WL 5333236, at *4 (D.D.C. Aug. 18, 2023)); Copyright Alliance, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5 (Oct. 30, 2023) ("Copyright Alliance Initial Comments"); DMLA Initial Comments at 17–18; Graphic Artists Guild Initial Comments at 1; Internet Archive, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10–11 (Oct. 30, 2023) ("Internet Archive Initial Comments").

[168] *See, e.g.*, Dallas Joder, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Nov. 30, 2023) ("Dallas Joder Reply Comments"); Peer Music-Boomy Joint Initial Comments at 14.

[169] For example, AI company BigBear.ai asserted that the Constitution "does not prohibit protection of AI-generated material," and that the availability of copyright protection "should not depend on the method through which [it] was generated." BigBear.ai Holdings, Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 25 (Oct. 18, 2023); *see also* Ryan Abbott, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 21, 2023) ("Ryan Abbott Initial Comments") ("The history and purpose of the Constitution and the Copyright Act both weigh in favor of the protection of AI-generated works because the public interest trumps any direct benefit to authors."); Peer Music-Boomy Joint Initial Comments at 15 ("[W]e do not believe that placing limitations on creators by limiting the sort of output we incentivize furthers the constitutional aims of copyright."); BLIP Initial Comments at 25 ("The Copyright Act should be amended to include a new section that provides protection for AI-generated material.").

[170] Dallas Joder Reply Comments at 3; *see also* Duane Valz, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Oct. 18, 2023) (While "the authors of the Constitution may not have imagined that entities other than natural persons would ever qualify as authors or inventors. . . .[, t]his doesn't mean that new types or persons or entities cannot be made eligible as authors or owners of copyrights if Congress sees fit to deem them such.").

[171] *See* A2IM-RIAA Joint Initial Comments at 34–35; The Authors Guild Initial Comments at 34; Anonymous AI Technical Writer, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 15 (Dec. 6, 2023) ("Anonymous AI Technical Writer Reply Comments"); Copyright Alliance Initial Comments at 96–97; DMLA Initial Comments at 17–18; Graphic Artists Guild Initial Comments at 20; David Newhoff, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 3 (Oct. 17, 2023) ("David Newhoff Initial Comments"); UMG Initial Comments at 81–82.

generated material can only be protected where there is sufficient human involvement or where AI is used as a tool to enhance human expression.[172]

These commenters emphasized that the Copyright Clause refers to promoting progress specifically by providing *authors* with legal and economic incentives.[173] They noted that AI systems, by contrast, are inanimate objects that "do not need an incentive to create."[174] As one commenter stated, "AIs do the work they are programmed to do, without regard to incentives."[175]

---

[172] *See* American Bar Association, Intellectual Property Law Section ("ABA-IPL"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 14 (Oct. 30, 2023) ("ABA-IPL Initial Comments"); American Intellectual Property Law Association ("AIPLA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Oct. 30, 2023) ("AIPLA Initial Comments"); Johan Brandstedt Initial Comments at 30; ACT | The App Association ("App Association"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6–7 (Oct. 30, 2023) ("App Association Initial Comments"); Entertainment Software Association, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 7 (Oct. 30, 2023); IPO Initial Comments at 7; Recording Academy Initial Comments at 11; Scenario Initial Comments at 16–17.

[173] *See* A2IM-RIAA Joint Initial Comments at 4 (quoting *Thaler v. Perlmutter*, No. 22-cv-1564, 2023 WL 5333236, at *4 (D.D.C. Aug. 18, 2023)), 35; ASCAP Initial Comments at 50; Authors Alliance Initial Comments at 18–19; DMLA Initial Comments at 17–18; Graphic Artists Guild Initial Comments at 1, 20; Daniel Gervais Initial Comments at 7; Fight for the Future, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 8 (Oct. 30, 2023) ("Fight for the Future Initial Comments"); Internet Archive Initial Comments at 10–11; Kernochan Center Initial Comments at 10–11; David Newhoff Initial Comments at 3; NMPA Initial Comments at 29–30; Seth Polansky Initial Comments at 29; Copyright Alliance Initial Comments at 5.

[174] Google LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 12 (Oct. 30, 2023); *see also* Computer & Communications Industry Association ("CCIA"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 19 (Oct. 30, 2023) ("CCIA Initial Comments") ("Computers don't need incentives; only people do. And existing incentives—both legal, such as copyrights and patents, and non-legal, such as first-mover advantages and a desire to supply a commercial need—will suffice to ensure the development of generative AI technologies."); AIPLA Initial Comments at 11; NMPA Initial Comments at 29 ("As a policy matter, copyright law should never protect purely AI-generated content that does not represent human expression. Existing copyright law rightfully incentivizes human creativity by granting protection to the 'the fruits of intellectual labor' that 'are founded in the creative powers of the mind.'"); Xiyin Tang et al., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10–11 (Oct. 30, 2023) ("Xiyin Tang et al. Initial Comments") ("The artificial intelligence itself needs no incentive, as it is programmed to create, and needs only human prompting to generate works. The only other party that could need the incentive of copyright would be the users of AI systems. However, creation of works using AI technology requires substantially less time and effort than most human created works. Humans receive copyright protection for their works to balance against the cost of creating those works, and the risk in investing so much time and resources only for another party to copy the finished product. With AI-created works, 'both the fixed and variable costs of producing each copyrightable article are effectively zero, which allows producers to compete with imitators even absent legal protection.'" (citations omitted)).

[175] Pamela Samuelson et al. Initial Comments at 3. *See also* A2IM-RIAA Joint Initial Comments at 4 (quoting *Thaler v. Perlmutter*, No. 22-cv-1564, 2023 WL 5333236, at *4 (D.D.C. Aug. 18, 2023)); Association of American Publishers ("AAP"), Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 31–32 (Oct. 30, 2023) ("AAP Initial Comments"); CCIA Initial Comments at 19; Internet Archive Initial Comments at 10–11 ("The traditional policy foundations for extending copyright protection generally do not apply in the case of AI-generated material. There is no evidence that copyright law provides necessary incentives for the creation of AI-generated works, and regardless,

Several commenters asserted that there appear to be sufficient incentives for AI companies under existing law.[176] Some pointed out that the exponential growth of AI technologies—even in the absence of copyright protection—indicates that their developers do not need copyright incentives to produce these technologies.[177] "As machine learning practitioners," the AI company Hugging Face, stated: "[W]e find that very little to no innovation in generative AI is driven by the hope of obtaining copyright protection for model outputs. The incentives for innovation already exist without modifying copyright law."[178]

Finally, many expressed concern that providing legal protection to AI-generated content would discourage human authorship. Representatives of copyright owners maintained that the proliferation of legally protected AI-generated outputs would stifle creativity, leading to an overall decrease in human-authored works available to the public because humans will be disincentivized to create.[179] For example, the Copyright Alliance predicted that "[i]f . . .

---

the constitutional foundations of copyright make clear that its goal is to incentivize human authorship."). *But see* Dallas Joder Reply Comments at 4 (predicting that self-aware AI might someday "rationally respond to [intellectual property ("IP")] incentives just like humans," such that they should be "permitted to keep and profit from the fruits of their creativity").

[176] A2IM-RIAA Joint Initial Comments at 35; AAP Initial Comments at 31–32; AIPLA Initial Comments at 11 ("At this time, it does not appear that legal protection for AI-generated outputs is critical to incentivizing the creation of AI technologies and systems; and the copyrightability of the AI system itself is sufficient."); CCIA Initial Comments at 19; Copyright Alliance Initial Comments at 95–96. Commenters identified several incentives, separate from any potential legal protection in AI-generated outputs, that encourage the development of AI technologies. *See, e.g.*, R Street Institute, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10 (Oct. 30, 2023) ("R Street Initial Comments") ("Existing copyright protection for computer code does offer some incentives for the development of generative AI technologies."); Xiyin Tang et al. Initial Comments at 10–11 ("There are already incentives for the creation and development of AI technology through patent and copyright protection in the machinery and software, so the developers of AI have been sufficiently incentivized to create and improve their programs."); CCIA Initial Comments at 19 (discussing perceived commercial need and first-mover advantage); Anonymous AI Technical Writer Reply Comments at 15 (discussing the availability of venture capital and stock-market funding for AI development); DMLA Initial Comments at 17 (discussing patents and trade secrets); UMG Initial Comments at 81 (discussing AI as a tool or service).

[177] AIPLA Initial Comments at 11 (noting that AI systems were "generated and commercialized in the absence of any clear authority providing legal protection to the outputs, and the absence of such protections does not appear to have diminished the public's interest in consuming AI, nor service-providers' interest in providing it"); The Authors Guild Initial Comments at 33; Copyright Alliance Initial Comments at 95–96; Graphic Artists Guild Initial Comments at 19–20.

[178] Hugging Face, Inc., Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 13 (Oct. 30, 2023).

[179] Take Creative Control, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Oct. 18, 2023); Software Freedom Conservancy, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Dec. 6, 2023); Timothy Allen, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry (Sept. 6, 2023) ("Not only does it prevent people from being able to claim any kind of ownership to their designs, it also creates a great degree of consumer confusion as to which pieces are real and which are not, and could have a chilling effect on further creative fields (many of which are already deeply suffering economically)[.]"); Anonymous Artist, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1, 10 (Dec. 5, 2023); Letter from UMG, Summary of *Ex Parte* Meeting on Apr. 22, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 6, 14 (Dec. 3, 2024).

policymakers give incentives to generate AI content, the sheer volume and speed with which AI material is generated could obliterate the markets for much human creation."[180] It further asserted that "[o]ur popular culture will be overtaken by low quality, AI-generated works because the cost of human creation would be deemed too burdensome in comparison to using AI."[181] The Authors Guild cautioned that if "AI-generated works were entitled to the same protection as human-created works," the producers of this material would have an "unfair leverage in the marketplace" which "would further incentivize the distribution of AI-generated content to the public, crowding and diluting the marketplace to the point that copyright incentives no longer function as intended."[182] It expressed particular concern that "[t]he creative middle class professions . . . will be drowned out and decimated," and that "our literary works and arts will suffer tremendously as a result."[183]

Some commenters sought to achieve the perceived value of incentives outside of the copyright system, proposing that AI-generated works could be protected instead through the establishment of new *sui generis* rights. They suggested that a "specialized right could be tailored to address the unique aspects of AI creations, including the balance between human input and AI processing," the term of protection, and the identity of rightsholders, among others.[184]

Of the commenters who addressed *sui generis* rights specifically, most opposed the idea. They saw *sui generis* rights as raising similar concerns about incentives and the impact on

---

[180] Copyright Alliance Initial Comments at 95; *see also* David Newhoff Initial Comments at 2–3 (explaining that vesting copyrights in corporate production of AI-generated material "pos[es] a threat to the careers of creative professionals" and that "[b]eyond posing a threat to the careers of creative professionals (and to the cultural value of creative work), at a certain point, the application of copyright law itself may become irrelevant and/or unconstitutional"); The Authors Guild Initial Comments at 34 ("Few human creators will be able to earn enough to sustain a profession and the human quality of work produced by professionals . . . will disappear."); Fight for the Future Initial Comments at 6. *But see* Donaldson Callif Perez Initial Comments at 2 ("Critics of artificial intelligence worry that the technology will eradicate jobs and be used to replace artists at the expense of human stories. Its proponents say that it is the way of the future and should be treated like just another tool in an artist's toolbox. The truth likely lies somewhere in the middle."); UMG, Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 11 (Dec. 6, 2023); A2IM-RIAA Joint Initial Comments at 35.

[181] Copyright Alliance Initial Comments at 95.

[182] The Authors Guild Initial Comments at 34.

[183] *Id.*

[184] ImageRights International, Inc. ("ImageRights"), Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 9 (Dec. 6, 2023) ("ImageRights Reply Comments"); *see also* Seth Polansky Initial Comments at 29 (suggesting shorter term for AI-generated material and clearer definition of who owns rights in outputs); Public Knowledge, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 19 (Oct. 30, 2023) (arguing that benefits of a *sui generis* right "may include faster and cheaper registration, and a lowered standard of documentation to illustrate which parts are attributable to AI, and (potentially) provenance of the work's AI components"); Rightsify Group LLC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10 (Oct. 30, 2023). A few advocated for *sui generis* protection specifically for AI model weights. *See* BLIP Initial Comments at 25–26; Van Lindberg Initial Comments at 5.

human authors.[185] Some also characterized past experience with *sui generis* regimes as problematic in various respects.[186]

In the Office's view, the case has not been made for additional protection for AI-generated material beyond that provided by existing law. As an initial matter, because copyright requires human authorship, copyright law cannot be the basis of protection for works that do not satisfy that requirement. As most commenters recognized, the incentives authorized by the Copyright Clause are to be provided to human authors as the means to promote progress. While Congress could instead consider establishing *sui generis* rights,[187] we do not find the policy arguments for additional protection to be persuasive.

To begin with, it is not clear that new incentives are needed. The developers of AI models and systems already enjoy meaningful incentives under existing law (as indicated by the rapid development and adoption of those models and systems). These incentives include patent, copyright, and trade-secret protection for the machinery and software, as well as potential funding and first-mover advantages. Moreover, we are not convinced that providing further incentives would promote progress. We share the concerns expressed about the impact of AI-generated material on human authors and the value that their creative expression provides to society. If a flood of easily and rapidly AI-generated content drowns out human-authored works in the marketplace, additional legal protection would undermine rather than advance the goals of the copyright system. The availability of vastly more works to choose from could actually make it harder to find inspiring or enlightening content. Indeed, AI

---

[185] *See, e.g.*, The Authors Guild Initial Comments at 33 (arguing that *sui generis* rights "will dilute the market for human-created works and . . . does not serve the goals of copyright or the needs of society"); EWC Initial Comments at 17; AAP Initial Comments at 31–32; ABA-IPL Initial Comments at 13–14; ASCAP Initial Comments at 49; Authors Alliance Initial Comments at 18–19; Kernochan Center Initial Comments at 10; NMPA Initial Comments at 29; App Association Initial Comments at 7; Pamela Samuelson et al. Initial Comments at 4; AIPLA Initial Comments at 11; R Street Initial Comments at 10.

[186] Consumer Technology Association, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 6 (Oct. 30, 2023) ("The history of *sui generis* approaches has been that as technology advances, they either quickly become obsolete (e.g., Semiconductor Chip Protection Act of 1984), or may raise uncertainties and impediments pertaining to copyright.").

[187] *See, e.g.*, the Semiconductor Chip Protection Act of 1984, establishing *sui generis* rights in mask works. H.R. REP. NO. 98-781, at 7–8 (1984), *reprinted in* 1984 U.S.C.C.A.N. 5750, 5756–57; Trademark Clarification Act of 1984, Pub. L. No. 98-620, § 301, 98 Stat. 3335, 3347 (1984); 17 U.S.C. §§ 901–14. *See also* the Vessel Hull Design Protection Act, establishing *sui generis* protection for original designs of vessel hulls. Digital Millenium Copyright Act, Pub. L. No. 105-304, Title V, § 502, 112 Stat. 2860, 2905 (1998), *amended by* the IP and Communications Omnibus Act of 1999, Pub. L. No. 106-113, § 5005, 113 Stat. 1536, 1501A–593 (1999); 17 U.S.C. §§ 1301–32. These rights differ from copyright in terms of eligibility, ownership rights, registration procedures, term, and remedies. It is difficult, however, to extrapolate from these examples, since experience with their use is limited and the context of today's widely used AI technologies is quite different.

training itself is reportedly reliant on human-generated content, with synthetic data leading to lower-quality results.[188]

There are already indications that AI-generated content has impacted some creators' ability to be compensated for their work.[189] Musicians and songwriters, for instance, have been impacted by the proliferation of AI-generated content on streaming services. UMG reported that "content oversupply," produced by an estimated 170 million AI-generated music tracks, currently threatens to dilute human creators' royalties.[190] AI-generated works have also threatened to reduce the pool of royalties available to human creators through the Mechanical Licensing Collective.[191]

If authors cannot make a living from their craft, they are likely to produce fewer works. And in our view, society would be poorer if the sparks of human creativity become fewer or dimmer.

B. Empowering Creators with Disabilities

A number of commenters asserted that extending protection to AI-generated works would empower more individuals with physical and cognitive disabilities to create.[192] The

---

[188] Kristian Hammond et al., *Degenerative AI: The Risks of Training Systems on their own Data*, NORTHWESTERN UNIV. CENTER FOR ADVANCING SAFETY OF MACHINE INTELL. (Sept. 6, 2024), https://casmi.northwestern.edu/news/articles/2024/degenerative-ai-the-risks-of-training-systems-on-their-own-data.html; Aatish Bhatia, *When A.I.'s Output Is a Threat to A.I. Itself*, N.Y. TIMES (Aug. 25, 2024), https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html.

[189] Researchers are beginning to seek to quantify the impacts of AI on artists' livelihoods. *See, e.g.*, International Confederation of Societies of Authors and Composers ("CISAC"), STUDY ON THE ECONOMIC IMPACT OF GENERATIVE AI IN THE MUSIC AND AUDIOVISUAL INDUSTRIES (Nov. 2024), https://www.cisac.org/services/reports-and-research/cisacpmp-strategy-ai-study; Gaétan de Rassenfosse et al., *Intellectual Property and Creative Machines*, NAT'L BUREAU OF ECON. RSCH. WORKING PAPERS, July 2024, Working Paper No. 32698, https://www.nber.org/papers/w32698.

[190] UMG Initial Comments at 13.

[191] Under a blanket license established in Section 115 of the Copyright Act, royalties for digital phonorecord deliveries of nondramatic musical works are paid into a pool for the mechanical licensing collective to divide and distribute to copyright owners. Although the Office has clarified that musical works that lack human authorship are not eligible for the blanket license, parties have attempted to obtain royalties for streams of AI-generated content. Letter from Suzanne V. Wilson, Gen. Couns. and Assoc. Register of Copyrights, U.S. Copyright Office, to Kris Ahrend, Chief Exec. Officer, The Mechanical Licensing Collective (Apr. 20, 2023), https://copyright.gov/ai/USCO-Guidance-Letter-to-The-MLC-Letter-on-AI-Created-Works.pdf. Such conduct has even been the basis of a criminal indictment for fraud. Press Release, U.S. Attorney's Office, Southern District of New York, North Carolina Musician Charged with Music Streaming Fraud Aided by Artificial Intelligence (Sept. 4, 2024), https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence.

[192] *See, e.g.*, BLIP Initial Comments at 24; ECPA Initial Comments at 8; Tom Yonge, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 33–36 (Sept. 18, 2023). Some commenters illustrated how generative AI has helped them create despite their disabilities. *See* Elisa Rae Shupe, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry Initial Comments at 1 (Oct. 27, 2023); Michael Summey, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 1 (Oct. 30, 2023).

specific applications they identified, however, involve the use of AI as a tool to assist in creating works, rather than to generate output without human authorship. The Brooklyn Law Incubator & Policy Clinic, for instance, cited functionalities like text-to-speech, visual art generative algorithms, and improving the written communication of those with cognitive disabilities.[193] Discussing creators with disabilities, another noted that "AI acts as a tool in the hands of an author," rather than a source of expressive content.[194]

The Office strongly supports the empowerment of all creators, including those with disabilities. We stress that to the extent these functionalities are used as tools to recast, transform, or adapt an author's expression, copyright protection would be available for the resulting work.[195] For example, the Office recently considered an application to register a sound recording by GRAMMY-winning country artist Randy Travis, who has limited speech function following a stroke.[196] The track was created based on the recording of a human voice, using "[a] special-purpose AI vocal model . . . as a tool . . . to help realize the sounds that Mr. Travis and the other members of the human creative team desired."[197] The result, which would have been infeasible without this technology, was a new track appearing to be sung in Travis's legendary voice. Because the sound recording used AI as a tool, not to generate expression, the Office registered the work.

The distinction between assistive uses and generative ones applies equally to creators with disabilities and other human authors. Copyright protection remains available where AI functions as an assistive tool that allows human authors to express their creativity.

C. Countering International Competition

A few commenters raised concerns about international competition. One organization warned that without copyright protection in the United States, "the scientific and creative communities will not be able to exploit the economic value of [AI-generated works]," which "may contribute to the U.S. lagging in the development of generative AI technologies and

---

[193] BLIP Initial Comments at 24.

[194] *See* ECPA Initial Comments at 8 (discussing artists who are not able to hold a paintbrush and stating that creators with disabilities are "wielding [AI] to create intended expression").

[195] Registration Guidance for Works Containing AI-Generated Content Tr. at 4–5 (June 28, 2023), https://www.copyright.gov/events/ai-application-process/Registration-of-Works-with-AI-Transcript.pdf.

[196] *Where That Came From*, SR0001018989 (May 29, 2024).

[197] Letter from Steven Englund to U.S. Copyright Office (Oct. 28, 2024). In correspondence with the Office, the applicant further explained that the model "was developed specifically for th[e] project under Mr. Travis'[s] supervision using a curated set of vocal tracks from prior recordings by Mr. Travis" and that "the creative team [used the tool] to translate a sonically-tailored recording of James Dupré singing the composition 'Where That Came From' into a vocal track in Mr. Travis'[s] distinctive voice, while preserving the original cadence, phrasing, articulation, dynamics and other musical characteristics of Mr. Dupré's human performance." *Id.*

systems."[198] Another commenter similarly stated that if the U.S. does not adopt copyright protection for AI-generated outputs, "the global locus of cultural [intellectual property] generation will . . . shift to other nations with more AI-friendly policy environments."[199] This commenter further argued that excluding AI-generated works from copyright protection would not actually serve artists' interests,[200] as American artists instead "will be swept away by a public domain flood of [low-cost] foreign AI content" with which they cannot compete.[201]

Regardless of what other countries conclude, however, the United States is bound by our own Constitution and copyright principles. We should not abandon or distort those principles simply because other countries may not share them. Rather, we should make a persuasive case that a human-centered approach is good policy and inherent to copyright.

In any event, as described above, it remains to be seen how other jurisdictions' copyright laws will address generative AI. Commenters' concerns assume a substantial disparity in legal protection for AI-generated material, but no such disparity has yet clearly emerged. As a group of law professors acknowledged, while generative AI is likely to have widespread impact on human creativity, its effects on employment are difficult to predict.[202]

D. Providing Greater Clarity

Some commenters stressed the benefits of clarity and certainty. They posited that creators would be better off with certainty that their works produced using AI would be protected and available to license or sell. One commenter said that otherwise, the "commercial viability of the works made using AI tools is undermined [and] . . . [t]he adoption of these tools will also be impacted."[203] Some cautioned that, absent greater clarity, authors may question whether they own what they create using AI, whether they can license their content to other parties, whether they can register their works with the Office, and

---

[198] The Knot Worldwide Inc. ("TKWW"), Reply Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 2 (Dec. 6, 2023) ("TKWW Reply Comments").

[199] Dallas Joder Reply Comments at 2.

[200] *See id.*

[201] *Id.* This commenter further cautioned that American AI startups will expend more financial resources on IP litigation than competitors in other countries that offer more expansive legal protection but did not explain how the volume of litigation would hinge on the copyrightability of AI-generated works. *See id.*

[202] *See* Pamela Samuelson et al. Initial Comments at 5.

[203] Microsoft and Github, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10 (Oct. 30, 2023); *see also* IPO Initial Comments at 6–7 ("[I]f works created by humans using AI tools are not protected, that creates uncertainty for companies. Uncertainty leads to difficulty planning, developing, and investing, which could undermine the encouragement and promotion of arts and sciences."); ABA-IPL Initial Comments at 13–14; App Association Initial Comments at 6; ECPA Initial Comments at 7–8; Van Lindberg Initial Comments at 46; MPA Initial Comments at 59; TKWW Reply Comments at 2; SCA Robotics Initial Comments at 1.

whether their registration certificates will be entitled to a presumption of validity in an infringement action.[204]

A number of commenters urged the enactment of legislation to articulate the scope of protection through guidelines or standards.[205] One suggested establishing a legal presumption that an AI system's owner is the author of any output that the system may generate.[206] Another contended that the law should clarify that an "insignificant use of an AI tool that is otherwise substantially created by a human" does not make that work ineligible for copyright protection.[207]

The Office understands the desire for clarity around the copyrightability of AI-generated material. We do not believe, however, that legislation is necessary at this point. Much of the concern expressed focused on the assistive use of AI tools, and this Report seeks to provide assurances that such uses do not undermine protection. As to determining the copyrightability of AI outputs, the courts will provide further guidance on the human authorship requirement as it applies to specific uses of AI (including in reviewing the Office's registration decisions). Meanwhile, the analysis in this Part of the Report can help to shed light on how existing principles and policies apply.

Even if Congress were to consider addressing this issue through legislation, greater clarity would be difficult to achieve. Because the copyrightability inquiry requires analysis of each work and the context of its creation, statutory language would be limited in its ability to provide brighter lines. Unless and until future developments raise new problems, the Office does not recommend a change in the law.

---

[204] *See* Sandra Aistars, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 10–11 (Oct. 30, 2023); Graphic Artists Guild Initial Comments at 2–3; Qualcomm Reply Comments at 6.

[205] BLIP Initial Comments at 22; CISAC, Comments Submitted in Response to U.S. Copyright Office's Aug. 30, 2023, Notice of Inquiry at 5–6 (Oct. 30, 2023); ImageRights Reply Comments at 8–9; INTA Initial Comments at 4–5; Seth Polansky Initial Comments at 27–28.

[206] Ryan Abbott Initial Comments at 18.

[207] ASCAP Initial Comments at 49.

V. CONCLUSION

Based on the fundamental principles of copyright, the current state of fast-evolving technology, and the information received in response to the NOI, the Copyright Office concludes that existing legal doctrines are adequate and appropriate to resolve questions of copyrightability. Copyright law has long adapted to new technology and can enable case-by-case determinations as to whether AI-generated outputs reflect sufficient human contribution to warrant copyright protection. As described above, in many circumstances these outputs will be copyrightable in whole or in part—where AI is used as a tool, and where a human has been able to determine the expressive elements they contain. Prompts alone, however, at this stage are unlikely to satisfy those requirements. The Office continues to monitor technological and legal developments to evaluate any need for a different approach.

The Office will provide ongoing assistance to the public on the copyrightability issues related to generative AI, including by issuing additional registration guidance and updating the relevant sections of the *Compendium of U.S. Copyright Office Practices*. In doing so, we will rely on the comments received in response to the NOI, judicial developments, and other relevant input.

ACKNOWLEDGEMENTS

Part 2 of the Copyright Office report on *Copyright and Artificial Intelligence*, addressing the topic of copyrightability of outputs generated by AI systems, is the product of a large and talented team.

I would particularly like to acknowledge the Office's senior leaders who guided the project and contributed their deep expertise: Emily Chapuis, Deputy General Counsel, Erik Bertin, Deputy Director of Registration Policy and Practice, and Robert Kasunic, Associate Register of Copyrights and Director of Registration Policy and Practice. Suzy Wilson, General Counsel and Associate Register of Copyrights, Andrew Foglia, Deputy Director for Policy and International Affairs, and Maria Strong, Associate Register of Copyrights and Director of Policy and International Affairs, played an important role in oversight, research, and drafting.

Thanks also go to the lawyers in the Office of General Counsel ("OGC"), the Office of Policy and International Affairs ("PIA"), and the Office of Registration Policy and Practice ("RPP") who were responsible for the foundational research and writing for this Part. Jalyce Mangum from OGC captained this team and contributed her deft writing and analysis at every stage of this project. Emily Lanza from PIA, Aaron Watson and Frank Muller from RPP, and Laurie Ann Taylor from OGC were key contributors. Additional research and input were provided by Michael Druckman and Elizabeth Porter from OGC. Danielle Johnson from PIA, and Joanna Blatchly, Michael Druckman, Brittany Lamb, Brandy Karl, and Gabriela Luna from OGC all assisted in finalizing the text and citations.

The Office's production and communications team ably shepherded the document's preparation and public release. Led by Associate Register for Public Information and Education Miriam Lord and Deputy Director George Thuronyi, the team included Steve Andreadis, Nicole Chen, Alison Hall, Lisa Marflak, Stanley Murgolo, Anjana Padmanabhan, Nora Scheland, and Naomi Wulansari.

Shira Perlmutter
Register of Copyrights and Director
U.S. Copyright Office
January 29, 2025

The U.S. Copyright Office's Copyright and Artificial Intelligence Report and additional information about the Office's AI initiative are available on the Copyright Office's website. Visit www.copyright.gov/AI for more information and to sign up for updates.

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 3: Generative AI Training

PRE-PUBLICATION VERSION

A REPORT OF THE REGISTER OF COPYRIGHTS

MAY 2025

UNITED STATES COPYRIGHT OFFICE

Copyright and Artificial Intelligence

Part 3: Generative AI Training PRE-PUBLICATION VERSION

A REPORT OF THE REGISTER OF COPYRIGHTS MAY 2025

The Office is releasing this pre-publication version of Part 3 in response to congressional inquiries and expressions of interest from stakeholders. A final version will be published in the near future, without any substantive changes expected in the analysis or conclusions.

TABLE OF CONTENTS

I. INTRODUCTION

This Part of the Copyright Office's Report on Copyright and Artificial Intelligence addresses the use of copyrighted works in the development of generative AI systems. The groundbreaking technologies involved draw on massive troves of data,[1] including copyrighted works, to enable the extraordinary capabilities they now offer to the public. Do any of the acts involved require the copyright owners' consent or compensation? And to the extent they do, how can that feasibly be accomplished?

These issues are the subject of intense debate. Dozens of lawsuits are pending in the United States, focusing on the application of copyright's fair use doctrine. Legislators around the world have proposed or enacted laws regarding the use of copyrighted works in AI training, whether to remove barriers or impose restrictions.

The stakes are high, and the consequences are often described in existential terms. Some warn that requiring AI companies to license copyrighted works would throttle a transformative technology, because it is not practically possible to obtain licenses for the volume and diversity of content necessary to power cutting-edge systems. Others fear that unlicensed training will corrode the creative ecosystem, with artists' entire bodies of works used against their will to produce content that competes with them in the marketplace. The public interest requires striking an effective balance, allowing technological innovation to flourish while maintaining a thriving creative community.

Pursuant to the Register of Copyrights' statutory responsibility to "[c]onduct studies" and "[a]dvise Congress on national and international issues relating to copyright,"[2] the Office published a Notice of Inquiry (NOI) in August 2023 posing a series of questions about copyright and AI. These included technical questions about how copyrighted works are collected, curated and used in training AI models,[3] legal questions about the application of the fair use doctrine,[4] and factual questions about existing or potential licensing arrangements.[5]

---

[1] We use the terms "data" and "dataset" here as shorthand for all types of content used in generative AI training, including copyrighted works. It is important to stress, however, that the works are not merely "data" in the ordinary sense, as they embody creative expression constituting protected authorship.

[2] 17 U.S.C. § 701(b)(1), (b)(4). *See also* Letter from Shira Perlmutter, Reg. of Copyrights, and Kathi Vidal, Under Sec'y of Com. for Intell. Prop. and Dir., U.S. Pat. and Trademark Off., to Sen. Chris Coons, Chair, and Sen. Thom Tillis, Ranking Member, Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary (Dec. 12, 2022), https://www.copyright.gov/laws/hearings/Letter-to-USPTO-USCO-on-National-Commission-on-AI-1.pdf.

[3] Artificial Intelligence Study: Notice of Inquiry, 88 Fed. Reg. 59942, 59948–49 (Aug. 30, 2023) ("NOI").

[4] *Id.* at 59946 (questions 6–6.4).

[5] *Id.* at 59946–47 (questions 6.1, 6.2, 9.3, 10.1–14).

Of the more than 10,000 comments received in response to the NOI, the overwhelming majority addressed one or more of these questions. The Office refers to these comments throughout the discussion below.

This Part of the Report proceeds as follows: Section II provides a technical overview of how generative AI systems are developed and deployed, as relevant to the copyright analysis. Section III identifies points in the development of generative AI systems where copying or other acts implicating copyright rights may occur. Section IV analyzes how the fair use doctrine may apply to those acts. Section V examines the practicality and advisability of various licensing options. Without opining on specific cases, we provide an analytical framework for identifying relevant facts and policy considerations. In so doing, we draw on substantial experience advising Congress, the courts, and the public on the fair use doctrine.[6]

The Office's analysis is necessarily limited to current circumstances and publicly available information. We recognize that the technology and markets involved are rapidly evolving, and courts and policymakers are at early stages in their considerations. As with other Parts of this Report, we will continue to monitor developments to determine whether any conclusions should be revisited.

Finally, we note that other parts of the U.S. government are also engaged on these important issues. In addition to ongoing activities in the courts and Congress,[7] the White

[6] The Office has the statutory responsibility to evaluate, every three years, whether proposed exemptions to the anti-circumvention provision of the Digital Millennium Copyright Act are likely to be fair use, s*ee Rulemaking Proceedings Under Section 1201 of Title 17*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/1201/. In addition, we maintain a Fair Use Index as a public resource on the case law, *Fair Use Index*, U.S. COPYRIGHT OFFICE, https://www.copyright.gov/fair-use/, and often evaluate the scope of fair use in policy studies. *See, e.g.*, U.S. COPYRIGHT OFFICE, SECTION 108 OF TITLE 17 (2017), https://www.copyright.gov/policy/section108/discussion-document.pdf; U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION (2015), https://www.copyright.gov/orphan/reports/orphan-works2015.pdf; U.S. COPYRIGHT OFFICE, COPYRIGHT PROTECTIONS FOR PRESS PUBLISHERS (2022), https://www.copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf. The Office also contributes to formulating U.S. government positions in major fair use cases, *see, e.g.*, Br. of the United States as *Amicus Curiae* Supporting Resp'ts, Andy Warhol Found. for the Visual Arts v. Goldsmith, 143 S. Ct. 1258 (2023) (No. 21-869), https://www.copyright.gov/rulings-filings/briefs/andy-warhol-found-for-the-visual-arts-v-goldsmith-no.21-869-2022.pdf; Br. of the United States as *Amicus Curiae* Supporting Resp't, Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021) (No. 18-956), https://www.copyright.gov/rulings-filings/briefs/google-llc-v-oracleamerica-inc-no-18-956-2020.pdf.

[7] *See, e.g.*, AI TASK FORCE, BIPARTISAN HOUSE TASK FORCE REPORT ON ARTIFICIAL INTELLIGENCE (Dec. 2024); *Artificial Intelligence and Intellectual Property: Part 1 — Interoperability of AI and Copyright Law: Hearing Before the Subcomm. on Cts., Intell. Prop., and the Internet of the H. Comm. on the Judiciary*, 118th Cong. (2024), https://www.congress.gov/event/118th-congress/house-event/115951; *Artificial Intelligence and Intellectual Property—Part II: Copyright: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 118th Cong. (July 12, 2023), https://www.congress.gov/event/118th-congress/senate-event/334425.

House is developing an AI Action Plan to advance America's AI leadership and has received public comments, including on the subject of intellectual property.[8]

[8] Request for Information on the Development of an Artificial Intelligence (AI) Action Plan, 90 Fed. Reg. 9088 (Feb. 6, 2025) (requesting input on "any relevant AI policy topic," including "intellectual property," to develop an AI Action Plan "to establish U.S. policy for sustaining and enhancing America's AI dominance in order to promote human flourishing, economic competitiveness, and national security"). A number of submissions have addressed copyright issues. *See, e.g.*, APP Comments on OSTP AI Action Plan (Mar. 15, 2025), https://publishers.org/wp-content/uploads/2025/03/White-House-AI-Action-Plan-Association-of-American-Publishers.pdf; OpenAI Comments on OSTP AI Action Plan (Mar. 13, 2025), https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf; MPA Comments on OSTP AI Action (Mar. 14, 2025), https://www.motionpictures.org/wp-content/uploads/2025/03/MPA-OSTP-AI-Responses-FINAL-3.14.25-1.pdf; Google Comments on OSTP AI Action Plan (Mar. 13, 2025), https://static.googleusercontent.com/media/publicpolicy.google/en//resources/response_us_ai_action_plan.pdf.

II. TECHNICAL BACKGROUND

This section describes how and why copyrighted works are used in the development of generative AI models. We begin by explaining how machine learning is applied to create generative AI models, using language models as an example. We then turn to the data required to train generative models and the nature of its use by developers. We describe different phases of training and the relationship between trained models and their training data. Finally, we address the deployment of models in generative AI systems, which may have a variety of purposes and incorporate software or processes intended to augment or restrict their behavior.

A. Machine Learning

Machine learning is a field of artificial intelligence focused on designing computer systems that can automatically learn and improve based on data or experience, without relying on explicitly programmed rules.[9] The basic technique involves creating a statistical model using examples of inputs and expected outputs, called "training data," along with a metric of how well the model performs.[10]

For example, machine learning can model the relationship between a company's advertising expenditures and product sales.[11] The training examples would be past expenditure and sales data, while the performance metric would be the difference between predicted and actual sales.[12] By measuring its performance on training examples and using that as feedback to make adjustments, the model "learns" from the data.[13] The goal is to develop a model that does not simply memorize training data, but reflects patterns or inferences that extend to new, or unseen situations, a concept called "generalization."[14]

---

[9] *See* National Artificial Intelligence Initiative Act of 2020, 15 U.S.C. § 9401(11); FRANÇOIS CHOLLET, DEEP LEARNING WITH PYTHON 4 (2d ed. 2021) ("DEEP LEARNING WITH PYTHON").

[10] DEEP LEARNING WITH PYTHON at 5.

[11] This example is based on an example in GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON 15–22 (2023) ("AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON"), https://www.statlearning.com/ (ebook).

[12] *See* AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON at 71.

[13] DEEP LEARNING WITH PYTHON at 5.

[14] *See id.* at 122; ZHANG ET AL., DIVE INTO DEEP LEARNING, chs. 3.6, 5.5 (2023) ("DIVE INTO DEEP LEARNING"), https://d2l.ai/; Peter L. Bartlett et al., *Deep Learning: A Statistical Viewpoint* at 2, ARXIV (Mar. 16, 2021), https://arxiv.org/abs/2103.09177.

Generative AI relies on a subset of machine learning that builds models using neural networks.[15] Broadly speaking, neural networks are mathematical functions that map, or transform, input data to output data.[16] These functions are described by a general structure and large collections of numbers, called parameters, which define the mapping of inputs to outputs.[17] With billions of parameters, collectively referred to as the network's "weights,"[18] modern neural networks are capable of computing highly complex transformations,[19] such as the conversion of text to video.[20]

When a neural network is first created, its weights are assigned random numbers, and it will not convert inputs to meaningful outputs.[21] By repeatedly exposing the network to training examples, measuring its performance on those examples, and making small adjustments to the weights in a direction that improves performance—sometimes analogized to tweaking and turning "knobs and dials"—the network approximates or "learns" how to transform inputs into expected outputs.[22]

---

[15] *See What is generative AI?*, IBM, https://www.ibm.com/think/topics/generative-ai ("Generative AI relies on sophisticated machine learning models called deep learning models"); DIVE INTO DEEP LEARNING, ch. 1.7 ("Deep learning is the subset of machine learning concerned with models based on many-layered neural networks").

[16] DEEP LEARNING WITH PYTHON at 8 ("machine learning is about mapping inputs (such as images) to targets (such as the label "cat") . . . deep neural networks do this input-to-target mapping via a deep sequences of simple data transformations (layers) and . . . these data transformations are learned by exposure to examples."). For a visual introduction to neural networks, *see* 3Blue1Brown, *But what is a neural network? | Deep learning chapter 1*, YOUTUBE (Oct. 5, 2017), https://youtu.be/aircAruvnKk.

[17] *See* DEEP LEARNING WITH PYTHON at 8–9. As a highly simplified example, the mathematical function $2x + 5$, has a general structure ($ax + b$) and numerical parameters ($a = 2, b = 5$). Changing the values of the parameters (*e.g.*, $4x + 7$ rather than $2x + 5$) results in a different mapping of inputs to outputs. *See also* MATHWORLD, PARAMETER, WOLFRAM, https://mathworld.wolfram.com/Parameter.html.

[18] AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN PYTHON at 404.

[19] *See* DEEP LEARNING WITH PYTHON at 9 ("It's a simple idea—but, as it turns out, very simple mechanisms, sufficiently scaled, can end up looking like magic.").

[20] *See, e.g.*, Meta, *Movie Gen: A Cast of Media Foundation Models*, ARXIV (Oct. 4, 2024), https://arxiv.org/abs/2410.13720 (describing a 30 billion parameter neural model capable of generating high-quality video clips from text).

[21] DEEP LEARNING WITH PYTHON at 9–10, 48 ("Finding the correct values for all of them may seem like a daunting task, especially given that modifying the value of one parameter will affect the behavior of all the others!"); DIVE INTO DEEP LEARNING, ch. 1.1.

[22] *See* DEEP LEARNING WITH PYTHON at 5, 10, 48. To return to the model of advertising expenditures (*Adv*) and product sales (*Sales*), if we were building our model from scratch, our first guess at the relationship between the two might be no better than random. For example, we might initially configure the model as $Sales = 2 \cdot Adv + 5000$. Machine learning would then use training examples (past expenditure and sales data) and a performance metric (the average squared difference between predicted and true sales) and update the parameters to improve predictions, *e.g.*, $Sales = 50 \cdot Adv + 7500$. For a simple model like this, it is often possible to directly calculate the optimal parameters for a

Accordingly, while code defines the basic structure of a neural network, it is the weights that reflect patterns learned from the training data, and which are most likely to be treated as proprietary by developers or draw the scrutiny of copyright owners.[23] After training, some developers use weights directly in their own products, while others distribute them to the public for use or further training.[24]

B. Generative Language Models

Given the line: "[i]t was the best of times, it was the worst of times, it was the age of wisdom, it was the age of . . .," many would be able to guess that the next word is "foolishness." Even if one is not familiar with *A Tale of Two Cities*, the context indicates that the next word is likely to be an antonym of "wisdom." This is not an unusual task for humans—we can all sometimes finish another's sentences.

This task can also be mathematically modeled. A statistical model of language can be represented by the probability of the next word given all the preceding words or "context."[25] By using a model to select a probable next word based on context, and then repeating the process, an AI system can take a short prompt and generate a continuing stream of language.[26] As Professor Murry Shanahan noted:

> [W]e might give [a large language model] the prompt 'Twinkle twinkle,' to which it will most likely respond 'little star.' On one level, we are asking the model to remind us of the lyrics of a well-known nursery rhyme. But in an important sense what we are really doing is asking it the following question: Given the statistical

given training dataset. However, this not feasible for neural networks. Instead, training is an iterative process that involves "modify[ing] the parameters little by little based on the [model's performance] on a random batch of data." DEEP LEARNING WITH PYTHON at 52–55.

[23] Although Meta provides freely viewable source code for the model architecture on Github, the corresponding weights are "gated" on Hugging Face, requiring an account and agreeing to a license agreement to access. *See* Meta-llama/Llama-3.1.-405B, HUGGING FACE, https://huggingface.co/meta-llama/Llama-3.1-405B/tree/main. *See also Introducing Meta Llama 3: The most capable openly available LLM to date*, META, https://ai.meta.com/blog/meta-llama-3/ (describing the architecture of Llama 3 as "relatively standard"); Framework for Artificial Intelligence Diffusion, 90 Fed. Reg. 4544 (Jan. 15, 2025) (interim final rule adopting export controls on artificial intelligence model weights for certain advanced closed-weight dual-use AI models).

[24] *See infra* Sections III.B, III.D.

[25] Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RSCH. 1137, 1138 (2003), https://jmlr.csail.mit.edu/papers/volume3/bengio03a/bengio03a.pdf.

[26] More precisely, models generate a probability distribution (*i.e.*, a list of probabilities) over their entire vocabulary. Many systems then use some form of random sampling to choose from among the most probable candidates. *See, e.g.*, Ari Holtzman et al., *The Curious Case of Neural Text Degeneration*, ICLR (2020), https://openreview.net/pdf?id=rygGQyrFvH; Alexandra DeLucia et al., *Decoding Methods for Neural Narrative Generative*, ASS'N. COMPUTATIONAL LINGUISTICS (2021), https://aclanthology.org/2021.gem-1.16.pdf.

> distribution of words in the public corpus, what words are most likely to follow the sequence 'Twinkle twinkle'? To which an accurate answer is 'little star.'[27]

In practice, models estimate probabilities for "tokens"[28] rather than words themselves. These are numbers that are pre-assigned or "indexed" to particular words, pieces of words, or punctuation marks.[29] Because neural networks are mathematical functions,[30] text must be converted to a numerical format for processing.[31] Tokens simply bridge the two formats, providing the unit of analysis for the model (*i.e.*, what it takes as an input and predicts as an output).

Example of text converted to a sequence of tokens prior to input.[32]

it	was	the	age	of	wisdom	,	it	was	the	age	of	foolish	-ness
480	673	290	5744	328	32646	11	480	673	290	5744	328	87785	2816

Currently, generative language models are typically trained with a technique called "generative pre-training."[33] During generative pre-training, text examples serve as both the

---

[27] Murray Shanahan, *Talking About Large Language Models* at 2, ARXIV (2023), https://arxiv.org/abs/2212.03551. Of course, whether a particular model predicts a high likelihood for "little" followed by "star," will depend on a variety of factors, including the frequency of that specific phrase in the training data, the frequency of other completions (*e.g.*, "twinkle twinkle song"), and the accuracy of the trained model.

[28] If we go back to the advertising and sales example, the model is a mathematical function that takes an input value (*Adv*) and maps it to an output value (*Sales*). For that function to work, the input must be in numerical form (e.g., dollars spent on television ads rather than clips of the ads themselves); *see also* BSA Initial Comments at 6 (Data is "transformed through 'tokenization,' which involves breaking down a piece of text or data into smaller units (or 'tokens') for purposes of computational analysis.").

[29] Many modern models use subword tokens because, among other reasons, they better accommodate rare words. *See, e.g.*, Sennrich et al., *Neural Machine Translation of Rare Words with Subword Units*, ARXIV (June 10, 2016), https://arxiv.org/abs/1508.07909.

[30] *See supra* notes 16–17 and accompanying text.

[31] *See* DEEP LEARNING WITH PYTHON at 311–12 (explaining that, as differentiable functions, deep learning models require numerical data, created through a process of tokenization and vectorizing).

[32] This example is based on the tokenizer demo from OpenAI. *See Tokenizer*, OPENAI, https://platform.openai.com/tokenizer.

[33] *See* Alec Radford et al., IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING 1 (2018) ("IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING"), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. The predominant architecture used for generative language modelling is the "transformer." One of the key features of the transformer architecture is the ability to generate predictions by processing input tokens in parallel. *See* Ashish Vaswani et al., *Attention is All You Need* at 6, ARXIV (Aug. 2, 2023), https://arxiv.org/abs/1706.03762. For a visual introduction to generative pre-trained transformers, *see* 3Blue1Brown, *Transformers (how LLMs work) explained visually*, YOUTUBE (Apr. 1, 2024), https://youtu.be/wjZofJX0v4M.

input and expected output, with performance measured by how well the model predicts each next token (output) based on preceding tokens (input).[34]

Consider a training example beginning: "*There are few people in England, I suppose, who have more true enjoyment of music than myself, or a better natural taste. If I had ever learnt, I should have been a great proficient . . .*"[35] Generative pre-training would generate predictions for each token in the example (except the first, which has no prior context), evaluate those predictions compared to the correct tokens (i.e., the ones that appeared in the training example), and then make small adjustments to the model's weights to increase the likelihood of the correct tokens. In other words, pre-training would adjust the model weights to increase the likelihood of the word "people" following the phrase "there are few," and so on for each token throughout the length of the training example.[36] This process is then repeated across many examples or batches of examples—some with similar introductions, *e.g.*, "*There are few sights sadder than a ruined book . . .*"[37]—with the goal of learning a general model of language that can then be adapted for specific tasks.[38]

Several years ago, researchers realized that by scaling this process—in other words, pre-training language models with more parameters, on more data, and with more computing power—it was possible to develop general purpose models that could perform well on a variety of diverse language-based tasks *without* the need for additional task-specific training.[39] Simply providing these models with natural language directions and then using them to iteratively predict each next token led to surprisingly good results. For example, early pre-trained models

---

[34] For a more precise mathematical definition, *see* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 3.

[35] JANE AUSTEN, PRIDE AND PREJUDICE ch. 31 (1813).

[36] The length of training examples is limited by a "context window." *See* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 3–4 (explaining that generative pre-training maximizes the next-token likelihood based on a fixed-sized "context window," which in the case of GPT-1 was 512 tokens). Since the introduction of generative pre-training, context windows have scaled to millions of tokens. *See, e.g.*, *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*, META, https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (introducing models with 1M and 10M token context windows).

[37] LEMONY SNICKET, THE WIDE WINDOW 109 (2000).

[38] The paper introducing generative pre-training provided several examples of such tasks, including sentiment classification, *i.e.*, classifying a snippet of text as "positive" or "negative." *See* IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING at 6; *Datasets: stanfordnlp, sst2*, HUGGING FACE, https://huggingface.co/datasets/stanfordnlp/sst2.

[39] Tom B. Brown et al., *Language Models are Few-Shot Learners* at 1, ARXIV (July 22, 2022) ("*Language Models are Few-Shot Learners*"), https://arxiv.org/abs/2005.14165; Alec Radford et al., *Language Models are Unsupervised Multitask Learners* at 1 (2019) ("*Language Models are Unsupervised Multitask Learners*"), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

could answer SAT analogy questions and translate English sentences to French with prompting alone (*e.g.*, "Q: what is the French translation of {sentence} A:").[40]

Although we have been discussing language models, the same general principles apply to generative models for other types of content such as images, video, and audio.[41] For example, image models can be trained using a combination of text and image tokens and a similar next-token prediction objective.[42] The text tokens come from descriptive captions for the images (whether human-authored or computer-generated) and provide context for iteratively predicting the image tokens.[43] Like language models, generative models for other types of content demonstrate sophisticated abilities when their training is scaled to large numbers of examples.[44]

C. Training Data

Below we discuss the characteristics that developers look for in training data, how they acquire it, and how they curate it for use in training.

1. Data Characteristics

The developers of generative AI models may consider many factors when compiling data for training. These include the quantity of data, its quality, and the ultimate purpose(s) of the model.

---

[40] *See Language Models are Few-Shot Learners* at 7, 14–15, 24–25, 60.

[41] *See, e.g.*, Elman Mansimov et al., *Generating Images from Captions with Attention*, ARXIV (Feb. 29, 2016), https://arxiv.org/abs/1511.02793; Aditya Ramesh et al., *Zero-Shot Text-to-Image Generation* at 1–2, ARXIV (Feb. 26, 2021) ("*Zero-Shot Text-to-Image Generation*"), https://arxiv.org/abs/2102.12092.

[42] *See, e.g.*, *Zero-Shot Text-to-Image Generation* at 2 ("We concatenate . . . text tokens with . . . image tokens, and train an autoregressive transformer to model the joint distribution over the text and image tokens.").

[43] *See id.* at 2. Diffusion models are another popular approach to image generation. These models are trained on images with random noise added to them and corresponding text captions. The training objective is to accurately predict the added noise and thus remove it from the image. Through this training, diffusion models develop the ability to generate novel images from pure noise and text captions alone. *See generally*, Jay Alammar, *The Illustrated Stable Diffusion* (rev. Nov. 2022), https://jalammar.github.io/illustrated-stable-diffusion/; Robin Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models* at 3–5, ARXIV (Apr. 13, 2022), https://arxiv.org/abs/2112.10752.

[44] *See, e.g.*, *Zero-Shot Text-to-Image Generation* at 9 ("We find that scale can lead to improved generalization, both in terms of zero-shot performance relative to previous domain-specific approaches, and in terms of the range of capabilities that emerge from a single generative model."); Jiahui Yu et al., *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation* at 2, ARXIV (June 22, 2022), https://arxiv.org/abs/2206.10789; Andrew Agostinelli et al., *MusicLM: Generating Music from Text* at 2, ARXIV (Jan. 26 2023), https://arxiv.org/abs/2301.11325 ("When trained on a large dataset of unlabeled music, MusicLM learns to generate long and coherent music at 24 kHz, for text descriptions of significant complexity, such as "*enchanting jazz song with a memorable saxophone solo and a solo singer*" or "*Berlin 90s techno with a low bass and strong kick*.").

Quantity. Generative AI models "are well-known for requiring . . . millions or billions of works for training purposes."[45] When not bottlenecked by other factors, such as computing power, increasing the quantity of training data typically increases a model's "performance," that is, its ability to make accurate predictions on test data not seen during training.[46] That performance has, so far, been associated with the ability of generative AI models to perform well on downstream tasks.[47] The scaling phenomenon has created a strong demand for data.[48] Some researchers have even suggested that, if current trends continue, language model training will soon exhaust the stock of publicly available text.[49]

It is an open question, however, how much data an AI developer needs, and the marginal effect of more data on a model's capabilities. Not everyone agrees that further increases in data and test performance will necessarily lead to continued real world improvements in utility.[50] Developers have also begun exploring techniques for training

---

[45] DMLA Initial Comments at 10–11. IBM stated that foundation models, *i.e.*, large models trained on broad data for a variety of downstream use cases, "require massive amounts of data (currently on the scale of terabytes) to enhance the model's quality, accuracy and flexibility." IBM Initial Comments at 3. Stable Diffusion reported that their image-generation model "was pre-trained on a filtered subset of two billion image and caption pairs." Stable Diffusion Initial Comments at 10. Rightsify stated that "[a] rudimentary model could be trained on a small music dataset of as little as 1,000 songs. However, high-quality music models that can resemble professionally produced music would require datasets of at least 1 million songs to be commercially viable." Rightsify Initial Comments at 5.

[46] *See, e.g.*, Jared Kaplan et al., *Scaling Laws for Neural Language Models* at 3, ARXIV (Jan. 23, 2020) ("*Scaling Laws for Neural Language Models*"), https://arxiv.org/abs/2001.08361; Tom Henighan et al., *Scaling Laws for Autoregressive Generative Modeling* at 3, ARXIV (Nov. 6, 2020), https://arxiv.org/abs/2010.14701; Hao Li et al., *On the Scalability of Diffusion-based Text-to-Image Generation* at 7, ARXIV (Apr. 3, 2024), https://arxiv.org/abs/2404.02883. The same is true for scaling compute (the computing power expended on training) and model size (the number of trainable parameters in the model), and all three are often scaled together to avoid "overfitting," a situation in which the data becomes a bottleneck for performance. *Scaling Laws for Neural Language Models* at 3.

[47] *See, e.g.*, *Language Models are Few-Shot Learners* at 4 (Suggesting "log loss"—a metric of how well a model's predictions align with expected outputs— "correlates well with many downstream tasks" and "follows a smooth trend of improvement with scale."); *Scaling Laws for Neural Language Models* at 10 ("In the domain of natural language, it will be important to investigate whether continued improvement on the loss translates into improvement on relevant language tasks.").

[48] For example, in 2022, Google researchers reported that "current large language models are significantly undertrained." Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* at 1, ARXIV (Mar. 29, 2022), https://arxiv.org/abs/2203.15556. They claimed that by training a smaller language model on substantially more data, they were able to outperform GPT-3 on various benchmarks. *Id.* To achieve this result, they trained a model using over 2 billion documents, including 4 million books (comprising approximately 20% of the dataset by size). *Id.* at 22.

[49] Pablo Villabos et al., *Will we run out of data? Limits of LLM scaling based on human-generated data* at 1, ARXIV (June 4, 2024), https://arxiv.org/abs/2211.04325 ("Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained.").

[50] *See, e.g.*, Arvind Narayanan & Sayash Kapoor, *AI scaling myths*, AI SNAKE OIL (June 27, 2024), https://www.aisnakeoil.com/p/ai-scaling-myths.

competitive models with less data. For example, researchers from Cornell trained a generative image model, Common Canvas, on approximately 70 million Creative-Commons-licensed images.[51] They claim the model has "comparable performance" to Stability AI's Stable Diffusion 2, even though it was trained on a substantially smaller dataset.[52]

Quality. The performance of models also depends heavily on the quality of the data used to train them. As reflected in the saying "garbage in, garbage out," poor quality training data can lead to poor quality outputs.[53] Recent research from major developers suggests that quality may even be a more important consideration than quantity.[54]

Some assessments of quality are more objective than others. Text scraped from the internet often contains error messages or other content with limited or negative training value.[55] Images may have inaccurate or misleading labels, such as a picture of an angry dog labeled as a "wolf,"[56] or they may be highly compressed with significant information loss and distortion.[57]

---

[51] Aaron Gokaslan et. al., *CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images* at 1–2, 6, ARXIV (Oct. 25, 2023), https://arxiv.org/abs/2310.16825.

[52] *Id.* at 1 (using a dataset that was less than 3% the size of the one used to train Stable Diffusion 2). The creation and use of synthetic data is another approach to reduce the dependency on large collections of human-authored data. *See* BigBear.ai Initial Comments at 8 ("synthetic data is created using algorithms or simulations and can help address limitations in the availability of real-world data."). However, use of synthetic data may lead to a phenomenon called model collapse where "the outputs quickly start denigrating into nonsense." *See* Authors Guild Initial Comments at 14; Illia Shumailov, AI models collapse when trained on recursively generated data, Ian Shumaylov et al., *AI models collapse when trained on recursively generated data*, NATURE (Jul. 24, 2024), https://www.nature.com/articles/s41586-024-07566-y; Xiaodan Xing et al., *On the Caveats of AI Authophagy* at 4–7, ARXIV (Nov. 8, 2024), https://arxiv.org/abs/2405.09597.

[53] *See, e.g.*, Youdi Gong et al., *A survey on dataset quality in machine learning*, 162 INFO. SOFTWARE TECH. 107268 (Oct. 2023), https://doi.org/10.1016/j.infsof.2023.107268.

[54] *See. e.g.*, Apple, *Apple Intelligence Foundation Language Models* at 4, ARXIV (July 29, 2024) ("*Apple Intelligence Foundation Language Models*"), https://arxiv.org/abs/2407.21075 ("We find that data quality, much more so than quantity, is the key determining factor of downstream model performance."); Marah Abdin et al., *Phi-4 Technical Report* at 1, ARXIV (Dec. 12, 2024) ("*Phi-4 Technical Report*"), https://arxiv.org/abs/2412.08905 ("[S]ignificant improvements in data quality can rival, and sometimes surpass, the performance gains traditionally achieved by scaling compute with model and dataset size.").

[55] *See* Colin Raffel et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* at 6, ARXIV (Sept. 19, 2023) https://arxiv.org/abs/1910.10683.

[56] Sedir Mohammed et al., *The Effects of Data Quality on Machine Learning Performance* at 9, ARXIV (Dec. 12, 2024), https://arxiv.org/abs/2207.14529.

[57] Qinhong Yang et al., *HQ-50K: A Large-scale, High-quality Dataset for Image Restoration* at 3, ARXIV (June 8, 2023), https://arxiv.org/abs/2306.05390.

Otherwise high-quality content may be watermarked, which has been described as "a big problem" for scraped image data.[58]

Other assessments are more subjective. Books, encyclopedias, academic papers, and legal opinions are generally considered high-quality sources of text because they are edited, factually rich, and cover diverse topics.[59] Works in the public domain may be older, leading to worse performance on modern language tasks,[60] while other readily available sources may reflect biases or contain "toxic" content.[61]

Purpose. The purpose for which a model is developed also governs the selection of data for training. Developers often seek to align the content of their training data with the expected use of the model.[62] For example, a language model for legal work would benefit from extensive training on legal documents,[63] and a language model for medical diagnostics would benefit from training on medical papers.[64] Likewise, an image model trained primarily on outdoor,

---

[58] Romain Beaumont, *Laion-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/. To address this problem, the LAION developers used automated tools to compute the probability of an image containing a watermark and deemed images with a probability exceeding 80% as "unsafe." *Id.* Nevertheless, one developer was sued by Getty Images based on allegations that it trained on LAION data and its model output images with distorted versions of Getty watermarks. Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del., Feb. 3, 2023).

[59] *See*, *e.g.*, Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3-4, ARXIV (Dec. 31, 2020) ("*The Pile: An 800GB Dataset of Diverse Text for Language Modeling*"), https://arxiv.org/abs/2101.00027; *Apple Intelligence Foundation Language Models* at 4.

[60] *See* Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity* at 9–11, ARXIV (Nov. 13, 2023), https://arxiv.org/abs/2305.13169.

[61] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 611 (2018); Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity* at 8, ARXIV (Nov. 13, 2023), https://arxiv.org/abs/2305.13169.

[62] In addition to its substantive content, sometimes developers seek data with a specific structure. For example, multilingual models often rely on a parallel corpus, that is, a set of aligned translations between two or more languages. However, obtaining high-quality parallel corpora is resource intensive, and some developers have turned to subtitles because they are readily available, cover many different languages, and are easy to align in parallel based on timestamps. *See, e.g.*, Reid Pryzant et al., *JESC: Japanese-English Subtitle Corpus*, ARXIV (Feb. 21, 2018), https://arxiv.org/abs/1710.10639; *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 26.

[63] Pierre Colombo et al., *SaulLM-7B: A pioneering Large Language Model for Law*, ARXIV (Mar. 7, 2024), https://arxiv.org/abs/2403.03883.

[64] *See* Zeming Chen et. al., *MediTron-70B: Scaling Medical Pretraining for Large Language Models* at 2–5, ARXIV (Nov. 27, 2023), https://arxiv.org/abs/2311.16079 (presenting "a pair of generative LLMs for medical reasoning, adapted from Llama-2 through continued pretraining on carefully curated high-quality medical sources," including clinical practice guidelines, open-access research papers, abstracts from non-open-access papers, and "diverse medical guidelines from the internet.").

natural images of land and seascapes is unlikely to perform as well on indoor or abstract images, such as quilt designs, posters, or cartoons.[65]

When training foundation models (*i.e.*, large models trained for a wide variety of use cases), developers use diverse training materials.[66] According to Meta, for a model to "realistically emulate all facets of human language," it is necessary to use data "reflecting a broad range of speech—from casual banter, to literary prose, to scientific jargon."[67] Public reporting on major technology companies has highlighted efforts to collect materials covering very specific content. In one instance, the developer of a generative video model apparently sought videos for "doing boxing," "hitting a pinata," "cracking neck," and "jaywalking."[68] If a model is intended to be general-purpose, able to generate videos of domains as varied as cross-country skiing, tropical fish, and modern dance, it will likely perform best if it has been trained on least some examples from each of those domains.[69]

2. Acquisition and Curation

Training data can be acquired in various ways from a variety of sources. One common practice is downloading "publicly available" data from the internet.[70] This can mean using automated tools to systematically "scrape" data from online sources, such as deploying stream-

---

[65] *See generally* Qinhong Yang et al., *HQ-50K: A Large-scale, High-quality Dataset for Image Restoration* at 1, ARXIV (June 8, 2023), https://arxiv.org/abs/2306.05390.

[66] *See, e.g.*, *Language Models are Unsupervised Multitask Learners* at 10 ("When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets.").

[67] Meta Initial Comments at 2.

[68] Samantha Cole, *AI Video Generator Runway Training on Thousands of YouTube Videos Without Permission*, 404 MEDIA (July 25, 2024), https://www.404media.co/runway-ai-image-generator-training-data-youtube/ (referring to one YouTube channel as "THE HOLY GRAIL OF CAR CINEMATICS SO FAR."). Another report on Nvidia's data collection suggested efforts to gather videos in somewhat broader categories, including a directive to focus on finding more "cinematic, drone footage, egocentric, and some travel and nature" videos. Samantha Cole, *Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI*, 404 MEDIA (Aug. 5, 2024), https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project/.

[69] In some cases, developers use general-purpose foundation models as a starting point to build new models with narrower purposes, by training on more domain-specific data to improve performance on a narrower set of tasks. *See infra* Section II.D.1.

[70] *See, e.g.*, Anthropic Initial Comments at 5; OpenAI Initial Comments at 5. "Publicly available" is not synonymous with "authorized." It may simply be used to mean "available on the internet." GenLaw Participants Initial Comments at 44–45; IBM RESEARCH, GRANITE FOUNDATION MODELS 2 (2024), https://www.ibm.com/downloads/cas/X9W4O6BM. For example, the Pile dataset included data from Wikipedia and Books3. Although Wikipedia text is often available under a Creative Commons Attribution-ShareAlike license, Books3 contains 196,640 books sourced from an unauthorized BitTorrent tracker. *See The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3 (citing Shawn Presser (@theshawwn), TWITTER (Oct. 25, 2020, 4:32 AM), https://twitter.com/theshawwn/status/1320282149329784833).

ripping software to download millions of video or subtitle files from YouTube.[71] Or it can mean downloading pre-existing databases, such as an entire copy of Wikipedia using one of the regularly provided backups offered by the site.[72] One particularly common source of training data is text scraped by web crawlers,[73] often obtained from Common Crawl.[74] Some developers have also turned to well-known pirate sources, such as shadow libraries with large collections of full, published books.[75]

Developers may also incorporate training data from licensed or non-public sources. Some own or have access to data acquired through interactions with customers or users.[76] They may also license data from third parties,[77] such as traditional publishers, intermediaries, and specialized data providers. Developers may find such material particularly desirable because it may not be available to competitors, is reliably high-quality, or promotes particular characteristics during training.[78]

---

[71] *See, e.g.*, Samantha Cole, *Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI*, 404 MEDIA (Aug. 5, 2024), https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project/; THE PILE: AN 800GB DATASET OF DIVERSE TEXT FOR LANGUAGE MODELING at 26.

[72] *See* WIKIMEDIA, Wikimedia Downloads, https://dumps.wikimedia.org/. *See also Language Models are Few-Shot Learners* at 8; *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 25.

[73] GPT-3, for example, was trained with several datasets comprised of scraped material, such as Common Crawl and WebText, comprising 82% of the weight of the training mix. *See Language Models are Few-Shot Learners* at 9.

[74] Common Crawl "maintains a free, open repository of web crawl data," which contains "over 250 billion pages spanning 17 years." COMMON CRAWL, https://commoncrawl.org/. "The corpus contains raw web page data, metadata extracts, and text extracts." *Id.* In their comment, Common Crawl stated that it is "the Primary Training Dataset for every LLM [and contributed to] 82% of raw tokens used to train GPT-3." Common Crawl Initial Comments at 1.

[75] *See, e.g.*, Thomas Heldrup, *Report on Pirated Content Used in the Training of Generative AI*, RIGHTS ALLIANCE 5–6 (2025), https://rettighedsalliancen.dk/wp-content/uploads/2025/03/Report-on-pirated-content-used-in-training-of-AI.pdf; Ashley Belanger, *Meta claims torrenting pirated books isn't illegal without proof of seeding*, ARS TECHNICA (Feb. 20, 2025), https://arstechnica.com/tech-policy/2025/02/meta-defends-its-vast-book-torrenting-were-just-a-leech-no-proof-of-seeding/. There are also allegations that some developers bypass paywalls to obtain data, whether for training or retrieval-augmented generation, a topic discussed in section III.C, *infra*. *See, e.g.*, Compl. ¶¶ 96, 98, Advance Local Media et al. v. Cohere, Inc., No. 1:25-cv-01305 (S.D.N.Y. Feb, 13, 2025).

[76] *See, e.g.*, Steven Vaughan-Nichols, *Meta uses your Facebook data to train its AI. Here's how to opt out (sort of)*, ZDNET (Aug. 30, 2023), https://www.zdnet.com/article/meta-uses-your-facebook-data-to-train-its-ai-heres-how-to-opt-out-sort-of/; *How your data is used to improve model performance*, OPENAI, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance ("ChatGPT, for instance, improves by further training on the conversations people have with it, unless you opt out."); *Terms of Service*, X.COM (effective Nov. 15, 2024).

[77] *See infra* Section IV.D.3 (characterizing licensing market activity).

[78] *See* Katie Paul & Anna Tong, *Inside Big Tech's Underground Race to Buy AI Training Data*, REUTERS (Apr. 5, 2024), https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

Regardless of its source, raw data typically undergoes a curation process to prepare it for training.[79] Because processing data on a massive scale is resource intensive, some developers rely, in whole or in part, on datasets that were initially collected and curated by third parties.[80] Examples of curation include filtering, cleaning, and compiling data.

Filtering. Filtering is a common practice, especially for data scraped from the internet, which often includes content that is undesirable for training.[81] Developers may use automated techniques to remove explicit, watermarked, mislabeled, or low-quality content, or to identify "aesthetic"[82] or high-quality subsets.[83] Other reasons for filtering include deduplication,[84] which may have the effect of reducing memorization, discussed below in Section II.D.2, and compliance with legal regimes.[85] For example, Getty Images states that when it licenses works for use in a commercial text-to-image model, it "curates a dataset that includes content that has been released for commercial use in respect of rights of publicity, privacy, trademark, and other intellectual property rights."[86]

---

[79] *See, e.g.*, BSA Initial Comments at 6 ("'Raw data' is frequently 'messy,' requiring significant work to transform the data into a usable form.").

[80] DATASET PROVIDERS ALLIANCE, SHAPING THE FUTURE OF AI DATA (2024) ("While some data types are abundant, there's a scarcity of high-quality, labeled data in specialized fields. . . a direct licensing model encourages creation and curation of high-quality datasets, driving innovation in both AI development and content creation"), https://www.thedpa.ai/ai-data-lilcensing-position-paper.

[81] Stefan Baack & Mozilla Insights, *Training Data for the Price of a Sandwich*, MOZILLA FOUND. (Feb. 6, 2024), https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/ ("Due to Common Crawl's deliberate lack of curation, AI builders do not use it directly as training data for their models. Instead, builders choose from a variety of filtered Common Crawl versions to train their LLMs.").

[82] *See, e.g.*, LAION-Aesthetics V1, LAION-AI, https://projects.laion.ai/laion-datasets/laion-aesthetic.html ("Laion aesthetic is a subset of laion5B that has been estimated by a model trained on top of clip embeddings to be aesthetic. The intended usage of this dataset is image generation.").

[83] *See, e.g.*, *Language Models are Few-Shot Learners* at 43 ("In order to improve the quality of Common Crawl, we developed an automatic filtering method to remove low quality documents."); Christoph Schuhmann et al., *LAION-5B: An open large-scale dataset for training next generation image-text models* at 5–6, ARXIV (Oct. 16, 2022), https://arxiv.org/abs/2210.08402 (filtering 90% of a 50 billion text-image dataset by removing images with short labels; small, malicious, or unusually large or redundant images; and images with low computed similarity to their text caption).

[84] *See, e.g.*, Lee et al., *Deduplicating Training Data Makes Language Models Better*, ARXIV (Mar. 24, 2022) https://arxiv.org/abs/2107.06499; *Language Models are Few-Shot Learners* at 43; *but see The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3, 27 (describing weighted sampling process that results in duplicates of documents from certain subsets).

[85] Christoph Schuhmann et al., *LAION-5B: An open large-scale dataset for training next generation image-text models* at 6, ARXIV (Oct. 16, 2022), https://arxiv.org/abs/2210.08402 ("In the case of illegal content, we computed CLIP embeddings to filter out such samples.").

[86] Getty Images Initial Comments at 9.

Cleaning. Documents that are not filtered may nevertheless benefit from some form of automated processing or "cleaning." Text scraped from the internet may contain excerpts with limited or negative training value, such as those related to navigation ("next" buttons), calls to action ("Read more…"), or social media counters ("likes").[87] Rather than excluding the entire document, these undesirable portions can be removed.[88] In some instances, this may include copyright-related information such as the author or owner of the work.[89]

Compiling. During curation, it is common to compile multiple datasets into a larger dataset with desirable properties and diverse coverage.[90] For example, the developers of the Pile—a dataset that has been used to train a number of generative language models—created the final dataset by sampling from 22 subsets. These included PubMed Central, an archive of nearly five million biomedical journal articles, to "benefit potential downstream applications in the medical domain," and Books3, a dataset of full-length books, for "long-range context modeling research and coherent storytelling."[91] During this process, developers sometimes "up-sample" or weight certain desired subsets, like Wikipedia, meaning they configure the sampling process to select examples from those subsets more often than others, resulting in greater representation and duplicates in the final dataset.[92]

---

[87] *See, e.g.*, Guilherme Penedo et al., *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* at 5, 27, ARXIV (June 1, 2023), https://arxiv.org/abs/2306.01116.

[88] *Id.* at 5.

[89] *See, e.g.*, *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 25 (describing the removal of "copyright information" from PubMed abstracts); Def.'s Mot. Partial Summ. J. at 37, Kadrey v. Meta Platforms, No. 23-cv-3417 (N.D. Cal., Mar. 24, 2025) ("unrebutted testimony from Meta employees and the parties' experts conclusively establish that Meta removed CMI from training data alongside other repetitive text as a part of industry standard procedures to improve performance"); Am. Compl. ¶¶ 45–54, The Intercept Media. v. OpenAI, No. 24-cv-01515, (S.D.N.Y. Jun 21, 2024).

[90] Van Lindberg Initial Comments at 7 ("Many new datasets are being created by compiling, converting, and annotating previously available training material into new, larger datasets."); Llama Team, AI @ Meta, *The Llama 3 Herd of Models* at 14, ARXIV (Nov. 23, 2024) ("*The Llama 3 Herd of Models*"), https://arxiv.org/abs/2407.21783 ("We made a several adjustments to the pre-training data mix during training to improve model performance on particular downstream tasks. In particular, we increased the percentage of non-English data during pre-training to improve . . . multilingual performance.").

[91] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3–4; *see also supra* note 70 (discussing unauthorized sourcing of Books3).

[92] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3, 27; Guilherme Penedo et al., *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* at 3, 9, 21 ARXIV (June 1, 2023), https://arxiv.org/abs/2306.01116 (describing and exploring tradeoffs with up-sampling); *The Llama 3 Herd of Models* at 14 (describing up-sampling on mathematical data and down sampling on "lower quality" data).

D. Training

Training is the procedure that uses data (*e.g.*, text or images) to develop generative AI models. As previously discussed, this requires identifying a formal measure or "objective" for how well the model performs, and then repeatedly adjusting the model's parameters based on that objective as the model is exposed to training data.[93] Two aspects of the training process are particularly relevant to the copyright analysis: training phases and memorization.

1. Training Phases

The training of generative AI models is rarely a single event, but an iterative process that may be stopped at any point and continued with different data, a different objective, or even a different actor guiding the process. For example, when training its Intelligence Foundation Language Models, Apple started with lower-quality, bulk web-crawl data before shifting to a mixture with higher-quality, longer, and licensed data over several stages.[94] After Meta trained the Llama 3 models, it publicly released their weights, which were then further trained by third parties to create new models, such as Perplexity's Sonar and Nvidia's Nemotron.[95] Thus, broad references to a model's "training" may obscure which data was used, for what purpose, and by whom.

Some commenters drew a distinction between two phases called "pre-training" and "post-training" or "fine-tuning."[96] OpenAI described the pre-training for language models as the step "in which a massive amount of computing power and data is spent to teach the model the broad foundations of language, grammar, and reasoning,"[97] and post-training as the step "where the pre-trained model is further trained on a (relative to pre-training) smaller amount of carefully curated data of specific tasks, like summarization or text classification."[98]

While commonly used, this terminology can be misleading. The term "pre-training" often distinguishes a *type* of training focused on accurately predicting examples from a large

---

[93] For resources introducing the technical details of model training *see, e.g.*, 3Blue1Brown, Neural Networks, Eps. 1–4, YOUTUBE, https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi (introducing neural networks and the training process with visuals); DEEP LEARNING WITH PYTHON at 48–62 (providing a technical introduction to gradient-based optimization).

[94] *Apple Intelligence Foundation Language Models* at 5.

[95] Perplexity Team, *Meet new Sonar: A Blazing Fast Model Optimized for Perplexity Search*, PERPLEXITY (Feb. 11, 2025), https://www.perplexity.ai/hub/blog/meet-new-sonar; Kari Briski, *NVIDIA Announces Nemotron Model Families to Advance Agentic AI*, NVIDIA (Jan. 6, 2025), https://blogs.nvidia.com/blog/nemotron-model-families/.

[96] OpenAI Reply Comments at 6; Hugging Face Initial Comments at 8.

[97] OpenAI Reply Comments at 6.

[98] *Id*.

dataset.[99] Thus, a third-party may engage in "continued pre-training" on a model that has been trained already,[100] and there may be multiple pre-training phases with different data.[101] The term may also imply that it is merely a preliminary stage with minor importance. Yet pre-training often requires orders of magnitude more data and computing power than other training; and it is the stage responsible for many of the sophisticated capabilities of generative AI models.[102] OpenAI's research papers introducing GPT-2 and GPT-3 made the point that by pre-training on a massive quantity of data, a model could perform well on a variety of tasks *without* additional training.[103]

"Post-training" or "fine-tuning" may refer to a variety of activities conducted for different purposes.[104] Some techniques focus on adapting a general-purpose model to perform narrowly defined tasks or generate specific content.[105] Others maintain the general-purpose

---

[99] *See, e.g.*, *Improving Language Understanding by Generative Pre-Training* at 3 (Describing pre-training as "learning a high-capacity language model on a large corpus of text); *The Llama 3 Herd of Models* at 14 (describing pre-training as a "stage in which the model is trained at massive scale using straightforward tasks such as next-word prediction or captioning").

[100] For example, the developers of SaulLM-7B, a large language model tailored for the legal domain, used a previously trained model as a starting point and then conducted "continued pretraining" on legal text to improve performance in that domain. *See* Pierre Colombo et. al., *SaulLM-7B: A pioneering Large Language Model for Law* at 2, 7, ARXIV (Mar. 7, 2024), https://arxiv.org/abs/2403.03883.

[101] *See Apple Intelligence Foundation Language Models* at 5.

[102] *See* Chunting Zhou et al., *LIMA: Less Is More for Alignment* at 1, ARXIV (May 18, 2023), https://arxiv.org/abs/2305.11206 (suggesting that "almost all knowledge in large language models is learned during pretraining, and only limited instruction tuning data is necessary to teach models to produce high quality output."); *but see* Mohit Raghavendra et al., *Revisiting the Superficial Alignment Hypothesis* at 1, ARXIV (Sept. 27, 2024), https://arxiv.org/abs/2410.03717 (suggesting the hypothesis "that almost all of a language model's abilities and knowledge are learned during pre-training, while post-training is about giving a model the right style and format" is, "at best, an over-simplification.").

[103] *See Language Models are Unsupervised Multitask Learners* at 1–2, 9; *Language Models are Few-Shot Learners* at 1, 5, 40–41.

[104] Meta defines "post-training" to refer to "any model training that happens outside of pre-training," which for modern foundation models can include "tun[ing] to follow instructions, align with human preferences, and improve specific capabilities (for example, coding and reasoning)." *The Llama 3 Herd of Models* at 1, 3.

[105] *See, e.g.*, *Improving Language Understanding by Generative Pre-Training* at 3, 6 (describing fine-tuning to adapt a language model to tasks such as sentiment classification); Nataniel Ruiz et al., *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation* at 1, ARXIV (Mar. 15, 2024), https://arxiv.org/abs/2208.12242 (describing the fine-tuning of an image model on a subject, *e.g.*, one's pet, to generate new images of that subject in different contexts).

nature of the model but focus on improving its ability to follow instructions or generate outputs that "align" with human preferences or intent.[106]

The upshot is that broad labels like "pre-training," "post-training," and "fine-tuning" do not fully convey the purpose, necessity, or impact of any particular training.[107] What an AI developer does with specific training data, and why, is necessarily case-specific.

2. Memorization

The extent to which models retain or "memorize" training data, which would then travel with the model in subsequent distributions, was disputed by commenters. Some AI companies asserted that "[t]here is no copy of the training data — whether text, images, or other formats — present in the model itself."[108] OpenAI characterized contrary arguments as based on "a common and unfortunate misperception of the technology," and argued that model weights are just "large strings of numbers" that reflect "statistical relationship[s]" among the training tokens.[109]

But others pointed to "numerous examples" of models generating "verbatim, near identical, or substantially similar outputs," arguing that they can "embody the expressive works they were trained on."[110] News/Media Alliance stated that "regardless of the exact technical processes employed," such behavior "has the same effect as memorization and retention."[111]

---

[106] *See, e.g.*, Long Ouyang et al., *Training language models to follow instructions with human feedback* at 1–2, ARXIV (Mar. 4, 2022), https://arxiv.org/abs/2203.02155; Yuntao Bai, *Constitutional AI: Harmlessness from AI Feedback* at 1–2, ARXIV (Dec. 15, 2022), https://arxiv.org/abs/2212.08073.

[107] New terminology also continues to proliferate. In a blog post introducing its latest model, Meta started using the term "mid-training" to describe "new training recipes including long context extension using specialized datasets." *See, e.g.*, *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*, META, https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (introducing models with 1M and 10M token context windows).

[108] Google Initial Comments at 3–4. *See also* Public Knowledge Initial Comments at 10.

[109] OpenAI Initial Comments at 6. *See also* Google Initial Comments at (emphasizing that models are simply the "encapsulation" of "statistical facts"). However, the research community has long touted the ability of language models to "implicitly store and retrieve knowledge." Adam Roberts, et al., *How Much Knowledge Can You Pack Into the Parameters of a Language Model?* at 1, ARXIV (Oct. 5, 2020), https://arxiv.org/abs/2002.08910. Since knowledge is implicitly stored in the model's parameters—or weights—it is sometimes referred to as "parametric" memory. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 1, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401.

[110] N/MA Reply Comments at 10–11; UMG Reply Comments at 6 ("In its recently filed lawsuit against Anthropic, UMG and other major music publishers identified 500 illustrative instances where simply asking for the lyrics to popular, copyrighted songs generated nearly identical infringing copies."). The Office plans to discuss liability for outputs that infringe copyright in Part 4 of this Report.

[111] N/MA Reply Comments at 12.

Seeking to reconcile these positions, A. Feder Cooper and James Grimmelmann explain that "the problem is that the [statistical] 'patterns" learned by a model can be highly abstract, highly specific, or anywhere in between."[112] Where the learned pattern is highly specific, "the pattern *is* the memorized training data."[113] Put another way, training involves comparing model outputs with examples and making small adjustments to the model's weights so that it is more likely to generate outputs closer to those examples.[114] While the *goal* may be to learn abstract patterns across training examples, the process does not appear to be inherently restricted to a particular level of abstraction.[115] In some cases, memorization may even be useful, with models exhibiting a "Goldilocks phenomenon; [they] are most useful when they memorize just the right amount, neither too little nor too much."[116]

OpenAI and other commenters acknowledged the potential for *some* memorization, but described it as rare, unintended, difficult to detect, and inconsistent with the purpose of training—"a bug, not a feature."[117] For example, Meta cited a study finding that one language

---

[112] A. Feder Cooper & James Grimmelmann, *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24, ARXIV (forthcoming 2025) ("*The Files are in the Computer: Copyright, Memorization and Generative AI*"), https://arxiv.org/abs/2404.12590. Consider the Noam Chomsky quote: "Colorless green ideas sleep furiously," which is an example of language that is "grammatically well-formed but semantically nonsensical." Giorgio Franceschelli et al., *Training Foundational Models as Data Compression: On Information, Model Weights and Copyright Law* at 2, ARXIV (Mar. 12, 2025), https://arxiv.org/abs/2407.13493. Researchers tested this quote on Meta's Llama 3 and found that the probability of "green" when given "Colorless" was 20%, while the probability of each of the subsequent words was always greater than 90%. *Id.*

[113] *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24.

[114] *See supra* Sections II.A, II.B.

[115] *See The Files are in the Computer: Copyright, Memorization and Generative AI* at 52–53, 55–56.

[116] *Id.* at 56; *see also* Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, ARXIV (Jan. 30, 2023) ("*Extracting Training Data from Diffusion Models*"), https://arxiv.org/abs/2301.13188 ("Our results also suggest that [the] theory that memorization is *necessary* for generalization in classifiers may extend to generative models, raising the question of whether the improved performance of diffusion models compared to prior approaches is precisely *because* diffusion models memorize more." (emphasis added)).

[117] *See, e.g.*, OpenAI Reply Comments at 9 n.23 (explaining that pre-trained language models can, "on rare occasions, 'memorize' training data such that it may output a verbatim excerpt of that data when prompted with a different portion of that data. This is considered a bug, not a feature, and . . . developers take steps both to prevent memorization from occurring and to prevent the output of verbatim copies of training data when it does"); Meta Initial Comments at 15–16.

model had a memorization rate of approximately one percent.[118] Given the scale of the training datasets, however, even one percent may not be trivial.[119]

Considerable research has been done on the extent to which and reasons why models memorize data.[120] A variety of factors appear to influence the extent of memorization, including the number of model parameters, the presence of duplicates in training data, training repeatedly on the same example, whether an example is unusual or an "outlier," at what point an example is seen during training, and how broadly memorization is defined.[121]

E. Deployment

In practice, users do not interact directly with the statistical models powering generative AI. Instead, these models are deployed in larger AI systems, which process and control the information flowing into and out of the models, connect them with other software tools, and provide a more convenient user interface.[122] The choices made during this deployment can have substantial impacts on what models can do and what material they use.

---

[118] Meta Initial Comments at 16 n.68 (citing Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* at 3–4, 9, ARXIV (Mar. 6, 2023), https://arxiv.org/abs/2202.07646).

[119] For example, the paper cited by Meta found that GPT-J 6B memorized at least 1% of its training dataset, the Pile, which is dataset that includes over 200 million documents. *See* Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* at 3–4, 9, ARXIV (Mar. 6, 2023), https://arxiv.org/abs/2202.07646; Stella Biderman et al., *Datasheet for the Pile* at 8, ARXIV (Jan. 20, 2022), https://arxiv.org/abs/2201.07311.

[120] *See, e.g.*, Milad Nasr. et al., *Scalable Extraction of Training Data from (Production) Language Models*, ARXIV (Nov. 28, 2023), https://arxiv.org/abs/2311.17035; *Extracting Training Data from Diffusion Models*; Mireshghallah et al., *An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models*, PROCEEDINGS OF THE 2022 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (December 7–11, 2022), https://aclanthology.org/2022.emnlp-main.119/; Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, ARXIV (Dec. 12, 2022), https://arxiv.org/abs/2212.03860.

[121] *See, e.g.*, *Extracting Training Data from Diffusion Models* at 6, 7; Milad Nasr. et al., *Scalable Extraction of Training Data from (Production) Language Models* at 6, ARXIV (Nov. 28, 2023), https://arxiv.org/abs/2311.17035. If "memorization" extends beyond extractable verbatim copies or excerpts of works, but also material that is close enough to be recognizable (*e.g.*, visual characteristics of an animated character separate from any specific image of that character), it may be more common. *See* Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295, 327 (2023) ("The Snoopy problem is that the more abstractly a copyrighted work is protected, the more likely it is that a generative AI model will 'copy' it."); *The Files are in the Computer: Copyright, Memorization and Generative AI* at 56–59.

[122] *See, e.g.*, *Introducing Canvas*, OPENAI, https://openai.com/index/introducing-canvas/ (describing "a new interface for working with ChatGPT on writing and coding projects that go beyond simple chat."); Benj Edwards, *Certain names make ChatGPT grind to a half, and we know why*, ARS TECHNICA (Dec. 2, 2024), https://arstechnica.com/information-technology/2024/12/certain-names-make-chatgpt-grind-to-a-halt-and-we-know-why/ ("OpenAI's ChatGPT is more than just an AI language model with a fancy interface. It's a system consisting of a stack of AI models and content filters that make sure its outputs don't embarrass OpenAI or get the company into legal trouble."); *Tool use with Claude*, ANTHROPIC, https://docs.anthropic.com/en/docs/build-with-claude/tool-use/overview (describing how developers can create systems that combine Claude with other software tools to "perform a wider variety of tasks").

The same AI model can be deployed in systems that perform very different tasks. OpenAI and Anthropic advertise their models' use for everything from keyword extraction[123] and classifying customer support tickets at scale,[124] to document summarization[125] and translation,[126] to fully generative tasks like writing a class lesson plan[127] or rap lyrics.[128] Although language models are particularly flexible, there are diverse use cases for other types of models as well.[129]

The nature of the model's deployment can also affect what materials it uses when generating outputs. Techniques have been developed to enable models to retrieve content from outside their training data when the system is responding to a specific request.[130] Researchers affiliated with Facebook coined the term "retrieval-augmented generation," or "RAG," to describe this process.[131] Many models use search engines for RAG, meaning they can generate queries that will be executed by the system, with the top results returned to the model in the form of an expanded prompt.[132] For example, given the question "What show won the Outstanding Drama award at the 2024 Emmys?", the generative AI assistant Claude can generate several queries such as "2024 emmy awards outstanding drama winner," send those queries to a third-party search engine (Brave Search), pull the full-text of the top results—

---

[123] *Default Keywords*, OPENAI, https://platform.openai.com/docs/examples/default-keywords.

[124] Anthropic's user guide describes a template prompt with natural language directions on how to classify customer support tickets and a placeholder for the text of specific tickets. The template prompt can then be integrated into a system that applies it to incoming support tickets and parses Claude's response to extract a classification label. *See Ticket Routing*, ANTHROPIC, https://docs.anthropic.com/en/docs/about-claude/use-case-guides/ticket-routing.

[125] *Ticket Routing*, OPENAI, https://platform.openai.com/docs/examples/default-summarize.

[126] *Translation*, OPENAI, https://platform.openai.com/docs/examples/default-translation.

[127] *Lesson Plan Writer*, OPENAI, https://platform.openai.com/docs/examples/default-lesson-plan-writer.

[128] *Rap Battle Writer*, OPENAI, https://platform.openai.com/docs/examples/default-rap-battle.

[129] For example, image generation models can erase unwanted objects "such as blemishes on portraits or items on desks," *API Reference*, *Erase*, STABILITY AI, https://platform.stability.ai/docs/api-reference#tag/Edit/paths/~1v2beta~1stable-image~1edit~1erase/post.

[130] *See, e.g.*, Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 1–2, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401.

[131] *Id*.

[132] *See, e.g.*, *Models*, *Llama 3.1*, LLAMA, https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/ (describing Llama's built-in tool calling for Brave search).

articles from CBS, Billboard, and others—and answer the user's question using the retrieved text as additional context.[133]

Beyond training and content retrieval, there are techniques developers can use to enhance models' capabilities during deployment.[134] Recently, advanced systems have begun to employ processes that allow language models to "think" and "act" before responding to user prompts. They "think" by generating text that verbally reasons through a problem before answering,[135] and they "act" by generating text that directs the system to take actions.[136] OpenAI's Deep Research can independently run from five to thirty minutes on a question, iteratively searching, copying, and analyzing various sources.[137]

In addition to augmenting models' outputs, systems can also constrain them. The developers of generative AI models and systems may employ a variety of "guardrails" to prevent them from generating objectionable content.[138] External filters can intercept prompts before they reach the generative model, or intercept model outputs before they reach the user.[139] "Safety prompting" uses hidden system prompts to reduce the likelihood of generating

---

[133] For an example of this process, see *Anthropic, Anthropic Cookbook, Web search using Brave Search Engine*, GITHUB, https://github.com/anthropics/anthropic-cookbook/blob/main/third_party/Brave/web_search_using_brave.ipynb.

[134] For example, an earlier Part of this Report described prompt optimization, where another generative AI model rewrites user prompts to make them more likely to generate more appealing outputs. U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE – PART 2: COPYRIGHTABILITY at 5–6 & n. 23 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf.

[135] *See* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* at 2–3, ARXIV (Jan. 10, 2023), https://arxiv.org/abs/2201.11903.

[136] *See* Shunyu Yao et al., *ReAct: Synergizing Reasoning and Acting in Language Models* at 9, ARXIV (Mar. 10, 2023), https://arxiv.org/abs/2210.03629.

[137] *Introducing Deep Research*, OPENAI, https://openai.com/index/introducing-deep-research/. One professor reported that it automatically explored "alternative ways of getting access to paywalled articles." Ethan Mollick, *The End of Search, The Beginning of Research*, ONE USEFUL THING (Feb. 3, 2025), https://www.oneusefulthing.org/p/the-end-of-search-the-beginning-of.

[138] *See* Peter Henderson et al., *Foundation Models and Fair Use* at 20–25, ARXIV, https://arxiv.org/abs/2303.15715 (suggesting a range of technical mitigation strategies employed by AI systems).

[139] *See, e.g.*, Traian Rebedea et al., *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails* at 12, ARXIV (Oct. 16, 2023), https://arxiv.org/abs/2310.10501; Gelei Deng et al., *Masterkey: Automated Jailbreaking of Large Language Model Chatbots*, NETWORK AND DISTRIBUTED SYSTEM SECURITY SYMPOSIUM 7 (Feb. 26, 2024), https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf.

undesirable outputs.[140] Alignment training is a type of continued model training designed to bring its behavior in line with human preferences or values.[141]

None of these approaches is infallible, however. The line between desired and undesired behavior is often subjective,[142] and users can intentionally, or sometimes unintentionally, bypass or degrade guardrails.[143] The implementation and efficacy of guardrails against copyright-infringing outputs has already been the subject of litigation.[144]

An additional point about deployment is that developers exert varying degrees of control over trained models, and the decisions shaping a model's use can be made by different actors. Some companies, like OpenAI and Anthropic, retain control over their models by deploying them on cloud services, providing access through consumer-facing products or an application programing interface ("API"), which lets third parties develop products without accessing or controlling the model directly.[145] Others, like Apple, have designed models for "on-device" use, which involves distributing weights to end users via embedded software or

---

[140] For example, the system prompt for ChatGPT appears to have included directions on how to generate Dall-E prompts. They instruct the system to "not create images in the style of artists, creative professionals or studios whose latest work was created after 1912 (*e.g.*, Picasso, Kahlo)." They further state: "Do not name or directly / indirectly mention or describe copyrighted characters. Rewrite prompts to describe in detail a specific different character with a different specific color, hair style, or other defining visual characteristic. Do not discuss copyright policies in responses." Pascal Hetscholdt, *Asking ChatGPT-4 about its 'system prompts', to prevent copyright infringement. GPT-4: Not all users may appreciate or understand the technicalities or reasoning behind system prompts*, PASCAL'S SUBSTACK (Feb. 9, 2024), https://p4sc4l.substack.com/p/asking-chatgpt-4-about-its-system (reporting ChatGPT's response to a user-prompt designed to elicit the full system prompt in response).

[141] *See, e.g.*, Long Ouyanget al., *Training language models to follow instructions with human feedback* at 1, ARXIV, https://arxiv.org/abs/2203.02155; Yuntao Bai, *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* at 1, 4–5 ARXIV, https://arxiv.org/abs/2204.05862; *see also* Peter Henderson et al., *Foundation Models and Fair Use* at 25, ARXIV, https://arxiv.org/abs/2303.15715 (discussing the potential use of alignment training to reduce copyright risks).

[142] *See, e.g.*, Traian Rebedea et al., *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails* at 14, ARXIV (Oct. 16, 2023), https://arxiv.org/abs/2310.10501 ("It should also be noted that evaluation of the output moderation rail is subjective and each person/organization would have different subjective opinions on what should be allowed to pass through or not.").

[143] *See, e.g.*, Daniel Kang et al., *Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks* at 1–2, (Feb. 2023), https://arxiv.org/abs/2302.05733; Xiangyu Qi et al., *Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend It To!* at 3, ARXIV (Oct. 5, 2023), https://arxiv.org/abs/2310.03693.

[144] In one ongoing lawsuit, Concord Music Group and Anthropic agreed to an order requiring Anthropic to maintain guardrails designed to prevent infringing outputs and to address future disputes regarding their efficacy. Stipulation and Order Regarding Prelim. Inj. Mot., Concord Music Grp. v. Anthropic PBC, No. 5:24-cv-3811 (N.D. Cal. Jan. 2, 2025), ECF No. 291.

[145] *See, e.g.*, *ChatGPT*, OPENAI, https://openai.com/chatgpt/overview/; *Build with Claude*, ANTHROPIC, https://www.anthropic.com/api/.

software updates.[146] And some major companies, including Meta, Microsoft, and Google, have released "open" models to the public, meaning that their downloadable weights can be shared, used, retrained, or deployed by anyone.[147] According to Hugging Face, the weights for one version of Meta's Llama 3 have been downloaded over 6 million times in the last month.[148]

[146] *See. e.g.*, *Apple Intelligence Foundation Language Models* at 6–7, 31 (describing the creation of a smaller "on-device" model, designed to run efficiently on an iPhone, iPad, or Mac.).

[147] This is subject to the enforceability of any licensing terms or terms of use, which can be quite permissive. *See, e.g.*, *Microsoft/Phi-4*, HUGGING FACE, https://huggingface.co/microsoft/phi-4 (MIT license); *but see Llama 4 Community License Agreement*, LLAMA, https://www.llama.com/llama4/license/ ("If . . . the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users . . . you are not authorized to exercise any of the rights under this Agreement.").

[148] *Meta/Llama-3.1-8B-Instruct*, HUGGING FACE, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

III. PRIMA FACIE INFRINGEMENT

The Copyright Act grants copyright owners a set of exclusive rights: to reproduce, distribute, publicly perform, and publicly display their works, as well as the right to prepare derivative works.[149] Establishing a prima facie case of infringement requires two elements: "(1) ownership of a valid copyright, and (2) copying of constituent elements of the work that are original."[150] Creating and deploying a generative AI system using copyright-protected material involves multiple acts that, absent a license or other defense, may infringe one or more rights.

A. Data Collection and Curation

The steps required to produce a training dataset containing copyrighted works clearly implicate the right of reproduction.[151] Developers make multiple copies of works by downloading them; transferring them across storage mediums; converting them to different formats; and creating modified versions or including them in filtered subsets.[152] In many cases, the first step is downloading data from publicly available locations,[153] but whatever the source, copies are made—often repeatedly.[154]

---

[149] 17 U.S.C. § 106.

[150] *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 361 (1991). *See also* 17 U.S.C. § 501(a). "The word 'copying' is shorthand for the infringing of any of the copyright owner's five exclusive rights, described at 17 U.S.C. § 106." *S.O.S., Inc. v. Payday, Inc.*, 886 F.2d 1081, 1085 n.3 (9th Cir. 1989).

[151] The right of reproduction extends to the protected elements of the copyrighted work, in whole or in part, and to non-literal copies with significant variations, as long they are substantially similar to the original. 17 U.S.C. § 106(a). *See Feist*, 499 U.S. at 361.

[152] *See supra* Section II.C.2. Some of these steps may also implicate the right to prepare derivative works, which includes translations, abridgments, condensations, or any other form in which a work may be recast, transformed, or adapted. 17 U.S.C. §§ 106(2), 101 (definition of "derivative work"). Developers may, as part of the curation process, abridge, rewrite, reorganize, or augment existing works. *See, e.g.*, Aaron Gokaslan et al., *CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images* at 2, ARXIV (Oct. 25, 2023), https://arxiv.org/abs/2310.16825 (using an image-to-text model to generate synthetic captions for images prior to training); *Phi-4 Technical Report* at 1–2, 5–6 (using human-generated works as "seeds" for generating synthetic data, e.g., by directing a pre-existing language model to rewrite or augment the content). To the extent this process removes text or metadata concerning the author, title, or other identifying information, it may also implicate section 1202's prohibition on the removal of copyright management information. *See* 17 U.S.C. § 1202(b). Potential violations of section 1202 in generative AI development will be addressed in a later part of this Report.

[153] *See supra* note 70 (discussing the distinction between publicly available and authorized).

[154] For instance, in addition to the various intermediate stages, the final Pile dataset contains multiple copies of documents from many of the sources, including Books3 and Wikipedia. *See The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3.

Most commenters agreed with or did not dispute that copying during the acquisition and curation process implicates the reproduction right.[155] As Professors Pamela Samuelson, Christopher Jon Sprigman, and Matthew Sag explained: "the process of training Generative AI models is generally preceded by massive amounts of web scraping that results in the creation of locally stored copies of millions or billions of copyrighted works."[156] Although some commenters noted that data may be discarded after the training process, that does not affect the infringement analysis.[157] Moreover, public reporting indicates that major developers often maintain training datasets for use in future projects.

B. Training

The training process also implicates the right of reproduction. First, the speed and scale of training requires developers to download the dataset and copy it to high-performance storage prior to training.[158] Second, during training, works or substantial portions of works are temporarily reproduced as they are "shown" to the model in batches.[159] Those copies may

[155] One professor asserted that the developers of the LAION-5B image dataset never downloaded image files from the internet, only textual information about those images. *See* Michael Murray Initial Comments at 4. But the LAION developers state that after downloading images they "subsequently discarded" them and further acknowledge that "any researcher using the datasets must . . . download[] the subset they are interested in." *FAQ*, LAION, https://laion.ai/faq/; *see also* Romain Beaumont, *Laoin-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ ("We download the raw images from the parsed URLs . . . .").

[156] Pamela Samuelson et al. Initial Comments at 7.

[157] *See* 17 U.S.C. § 101 (definition of "fixed"); *see also* DMLA Initial Comments at 7 ("Retention practices vary among AI developers. Some will delete the training sets used in their AI models, while some will store them. However, retention policies do not actually have much bearing on copyright infringement. The right of reproduction may be violated regardless of whether a work is retained or stored.").

[158] *Cf. FAQ*, LAION, https://laion.ai/faq/ ("Any researcher using the datasets must reconstruct the images data by downloading the subset they are interested in. For this purpose, we suggest the img2dataset tool."); *The Llama 3 Herd of Models* at 9 (explaining how "Tectonic, Meta's general-purpose distributed file system, [was] used to build a storage fabric for Llama 3 pre-training." (internal cites omitted)); Daniel Gervais Initial Comments at 3 ("[I]t is almost always the case that a copy is kept . . . because training from a local copy is more efficient.).

[159] *The Llama 3 Herd of Models* at 19 (describing training on batches of examples, with examples ranging from 4K to 128K tokens in length); AAP Initial Comments at 11 ("The training process may involve a number of optimization processes such as mini-batching, shuffling, or caching, each of which may involve a temporary reproduction.") *see also supra* note 36.

persist long enough to infringe the right of reproduction,[160] depending on the model at issue and the specific hardware and software implementations used by developers.[161]

Third, the training process—providing training examples, measuring the model's performance against expected outputs, and iteratively updating weights to improve performance—may result in model weights that contain copies of works in the training data. If so, then subsequent copying of the model weights, even by parties not involved in the training process, could also constitute prima facie infringement.

As discussed in the Technological Background, the extent to which models memorize training examples is disputed.[162] When, however, a specific model can generate verbatim or substantially similar copies of a training example, without that expression being provided externally in the form of a prompt or other input, it must exist in some form in the model's weights.[163] When a model takes the prompt "Ann Graham Lotz" and outputs an image that is nearly identical to a portrait found in the training data, the expression in that image clearly comes from the model.[164] As A. Feder Cooper and James Grimmelmann put it, "a model is not a magical portal that pulls fresh information from some parallel universe into our own."[165]

In such instances, there is a strong argument that copying the model's weights implicates the right of reproduction for the memorized examples. Like other digital files that encode or compress content using mathematical representations, the content need not be directly perceivable to constitute a copy.[166] The relevant question is whether the work is "fixed"

---

[160] 17 U.S.C. § 101 ("A work is 'fixed' in a tangible medium of expression when its embodiment in a copy or phonorecord . . . is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration"); *Compare MAI Sys. Corp. v. Peak Comput., Inc.*, 991 F.2d 511, 518–19 (9th Cir. 1993) (fixation where software copy loaded into random-access memory for a sufficient duration to enable technician to view or diagnose computer errors), *with Cartoon Network LP, v. CSC Holdings, Inc.*, 536 F.3d 121, 129–30 (2d Cir. 2008) (no fixation where a programming stream was copied "one small piece at a time" in a buffer, and not maintained for a duration "more than a fleeting 1.2 seconds"). Such temporary copying may, however, be of little significance given the other acts of reproduction in AI training. *See* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 63 (2017).

[161] *See, e.g.*, *Data Input Pipelines: Optimize Pipeline Performance – Better Performance with the tf.data API*, TENSORFLOW (Aug. 15, 2024), https://www.tensorflow.org/guide/data_performance (discussing techniques, including caching, to achieve peak performance by building "an efficient input pipeline that delivers data for the next step before the current step has finished").

[162] *See supra* Section II.D.2.

[163] N/MA Initial Comments at 25–26, 29.

[164] *See Extracting Training Data from Diffusion Models* at 1.

[165] *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24.

[166] For example, the JPEG format represents images as a series of cosine waves oscillating at different frequencies, stored as corresponding numerical coefficients. *See, e.g.*, International Telecommunication Union, Recommendation

and "can be perceived, reproduced, or otherwise communicated . . . with the aid of a machine or device."[167] Since model weights are lists of numbers that do not change (barring further training), they are fixed, and because memorized works can be generated and displayed using software, those works can be perceived or reproduced with the aid of a machine.

Model weights that have memorized protectable expression from training data may also infringe the derivative work right. Some commenters asserted that model weights are necessarily abstractions or transformations of *all* the original training data.[168] NMPA, for example, stated that "[u]ltimately, the model becomes an abstract agglomeration of its training material capable of generating (i.e., communicating) verbatim copies of works within the training set, many of which are copyrighted. Such qualities fall squarely within the Copyright Act's definition of a derivative work."[169] Others argued that models cannot be derivative works because they do not contain training examples—they only learn from them through an abstraction process.[170] Citing *Authors Guild* for this point, TechNet asserted that models "do not represent any protected aspects of the original works to users."[171]

Courts that have addressed infringement claims regarding model weights have reached varying conclusions. In *Kadrey v. Meta Platforms*, the court described allegations that the Llama models themselves were infringing derivative works as "nonsensical."[172] In that case, however, the complaint did not allege that the models could "spit[] out actual copies of their protected

---

T.81 at 3–4, 14–15 (Sept. 1992), https://www.w3.org/Graphics/JPEG/itu-t81.pdf ("JPEG 1"); Computerphile, *JPEG DCT, Discrete Cosine Transform (JPEG Pt2)- Computerphile*, YOUTUBE (May 22, 2015), https://www.youtube.com/watch?v=Q2aEzeMDHMA.

[167] 17 U.S.C. § 101.

[168] *See* NMPA Initial Comments at 10–11 ("AI researchers have themselves explained that, in the training process, training material is '*transformed* and modeled in a very different representation of weights and biases . . . . *[I]t is derivative work*[.]'" (quoting Sharon Goldman, *The Data that Trains AI is Under the Spotlight — And Even I'm Weirded Out | The AI Beat*, VENTUREBEAT (Apr. 24, 2023), https://venturebeat.com/ai/the-data-that-trains-ai-is-under-the-spotlight-and-even-im-weirded-out-the-ai-beat/)); A2IM-RIAA Joint Initial Comments at 28.

[169] *See* NMPA Initial Comments at 10–11.

[170] *See* TechNet Initial Comments at 3 n.6; University Library of the University of California, Berkeley Initial Comments at n.24; *see also* Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 302 (2023) ("[T]he link between copyrighted works in the training data and generative AI outputs is highly attenuated by a process of decomposition, abstraction, and remix."); Public Knowledge Initial Comments at 18 ("One can download the fully trained Stable Diffusion model weights at the size of 4 GB, while the LAION-2B dataset it is trained on — the smallest version — contains around 80,000 GB of images; no amount of compression would allow for the model to contain all that information.").

[171] *See* TechNet Initial Comments at 3 n.6 (citing *Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202, 225–26 (2d Cir. 2015)) (internal marks omitted).

[172] *Kadrey v. Meta Platforms, Inc.*, No. 23-cv-3417, 2023 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023).

works" or outputs that are "similar enough … to be infringing derivative works."[173] In *Andersen v. Stability AI*, by contrast, the court denied a motion to dismiss filed by a third party that was not involved in the training process but had downloaded and used an already-trained model.[174] It found sufficient allegations that copies or protected elements remained, in some format, within the model.[175] The court distinguished *Kadrey* on the ground that the "necessary allegations regarding the products' training and operations, [were] materially different."[176]

The Office agrees with this distinction. Whether a model's weights implicate the reproduction or derivative work rights turns on whether the model has retained or memorized substantial protectable expression from the work(s) at issue.[177] As discussed above, the use of those works in preparing a training dataset and training a model implicates the reproduction right, but copying the resulting weights will only infringe where there is substantial similarity.[178]

C. RAG

RAG also involves the reproduction of copyrighted works.[179] Typically, RAG works in one of two ways. In one, the AI developer copies material into a retrieval database, and the generative AI system can later access that database to retrieve relevant material and supply it to the model along with the user's prompt.[180] In the other, the system retrieves material from an

---

[173] *Id.*

[174] *Andersen v. Stability AI Ltd.*, 744 F. Supp. 3d 956, 982–84 (N.D. Cal. 2024).

[175] *Id. See also id.* at 974 ("That these works may be contained in Stable Diffusion as algorithmic or mathematical representations – and are therefore fixed in a different medium than they may have originally been produced in – is not an impediment to the claim at this juncture.").

[176] *Andersen*, 744 F. Supp. 3d at 975 n.16.

[177] Use of the works in preparing the model is not enough—there must be evidence of substantial similarity. *See, e.g.*, *Litchfield v. Spielberg*, 736 F.2d 1352, 1357 (9th Cir. 1984) (rejecting the argument that the derivative work right does not require substantial similarity: "[t]o prove infringement, one must show substantial similarity."); *Castle Rock Ent., Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 143 n.9 (2d Cir. 1998); *Alcatel USA, Inc. v. DGI Techs., Inc.*, 166 F.3d 772, 787 n.55 (5th Cir. 1999); *Kohus v. Mariol*, 328 F.3d 848, 858 (6th Cir. 2003); *Bucklew v. Hawkins, Ash, Baptie & Co., LLP.*, 329 F.3d 923, 930 (7th Cir. 2003).

[178] Other facts related to memorization may be relevant to the fair use analysis, such as frequency of memorization and the ability to access memorized content. *See infra* Sections IV.B.1.c; IV.D.3.

[179] *See supra* notes 130–33.

[180] For example, in the paper introducing RAG, the researchers downloaded a copy of Wikipedia and split it into 100-word chunks. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* at 4, ARXIV (Apr. 12, 2021), https://arxiv.org/abs/2005.11401; *see also* Compl. at 4–5, *Dow Jones & Co., v. Perplexity AI, Inc.*, No. 24-cv-7984 (S.D.N.Y. Oct 21, 2024), ECF No. 1 (alleging "copying without authorization massive amounts of [news publisher's] copyrighted works for inclusion into Perplexity's RAG index.").

external source (for example, a search engine or a specific website).[181] Both methods involve making reproductions, including when the system copies retrieved content at generation time to augment its response.[182] We note that RAG is an important feature of many AI products, and that RAG-related uses are of particular concern for news media stakeholders.[183]

D. Outputs

Generative AI models sometimes output material that replicates or closely resembles copyrighted works. Users have demonstrated that generative AI can produce near exact replicas of still images from movies,[184] copyrightable characters,[185] or text from news stories.[186] Such outputs likely infringe the reproduction right and, to the extent they adapt the originals, the right to prepare derivative works. Some commenters noted that, depending on the content type and the audience, they may implicate the public display and public performance rights as well.[187] These infringement issues, including enforcement challenges and the allocation of potential liability, will be addressed in a later Part of this Report.

---

[181] *See supra* notes 132–33.

[182] In the case of materials retrieved from an external database, the reproductions may be short-lived. *See supra* note 160. However, many chat systems maintain a "state" or history of the conversation, which may preserve information, including the results of retrieval, across repeated interactions with the user. *See, e.g.*, *Conversation State*, OPENAI PLATFORM, https://platform.openai.com/docs/guides/conversation-state?api-mode=chat. In one pending case, the plaintiffs alleged that the full text of retrieved articles was visible to users in an "Under the Hood" view available on the company's developer interface. *See* Compl. ¶¶ 79–81, Advance Local Media et al. v. Cohere, Inc., No. 25-cv-1305 (S.D.N.Y. Feb, 13, 2025).

[183] *See, e.g.*, Letter from N/MA, Summary of *Ex Parte* Meeting on Apr. 29, 2024 Regarding the Office's AI Study, to U.S. Copyright Office (May 3, 2024), https://www.copyright.gov/policy/artificial-intelligence/ex-parte-communications/letters/NewsMedia-Alliance-May-3-2024.pdf; Raptive Initial Comments at 2 (describing the use of AI in connection with online search series as "particularly concerning" since summaries "would substitute for and drive traffic from the materials that were wholesale ingested to create the service.").

[184] Gary Marcus & Reid Southern, *Generative AI Has a Visual Plagiarism Problem*, IEEE SPECTRUM (Jan. 6, 2024), https://spectrum.ieee.org/midjourney-copyright.

[185] Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295, 327–37 (2023) (describing the "Snoopy problem"); Gary Marcus & Reid Southern, *Generative AI Has a Visual Plagiarism Problem*, IEEE SPECTRUM (Jan. 6, 2024), https://spectrum.ieee.org/midjourney-copyright.

[186] First Am. Compl. at 29–47, New York Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. Aug. 12, 2024), ECF No. 170; *see also* Compl. at 25–27, 30, Dow Jones & Co., Inc. v. Perplexity AI, Inc., No. 24-cv-7984 (S.D.N.Y. Oct. 21, 2024), ECF No. 1 (displaying examples of verbatim and detailed summary text outputs from *The Wall Street Journal* and *New York Post*).

[187] *See* Katherine Lee et al. Initial Comment at 65; ASCAP Initial Comments at 11.

IV. FAIR USE

To the extent that acts involved in developing and deploying a generative AI model constitute prima facie infringement, the primary defense available is fair use.

Fair use is a judge-made doctrine now codified in Section 107 of the 1976 Copyright Act. It provides that "the fair use of a copyrighted work . . . is not an infringement of copyright" and lists four non-exclusive factors that must be considered in determining whether a particular use is fair:

> *(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;*
>
> *(2) the nature of the copyrighted work;*
>
> *(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and*
>
> (4) *the effect of the use upon the potential market for or value of the copyrighted work.*[188]

These statutory factors are not to be applied mechanically.[189] Rather, they "set forth general principles, the application of which requires judicial balancing, depending upon relevant circumstances."[190] Fair use is, fundamentally, an "equitable rule of reason."[191] It is an affirmative defense, with the defendant bearing the burden of proof.[192] In approaching fair use claims involving new technologies, courts have sought to further copyright's "basic purpose" of promoting progress by striking a balance between protecting authors' exclusive rights in their works and enabling others to build upon those works.[193]

The comments the Office received in response to the NOI were sharply divided on the applicability of fair use. On one side, commenters painted a dire picture of what unlicensed

---

[188] 17 U.S.C. § 107.

[189] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 577 (1994) (fair use "is not to be simplified with bright-line rules, . . . [it] calls for case-by-case analysis" (citing *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 560 (1985))).

[190] *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 19 (2021).

[191] *Id.* at 18.

[192] *See Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 459 (9th Cir. 2020) ("Not much about the fair use doctrine lends itself to absolute statements, but the Supreme Court and our circuit have unequivocally placed the burden of proof on the proponent of the affirmative defense of fair use."). *See also Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 547 n.21 (2023) ("*Warhol*"); *Campbell*, 510 U.S. at 590; *Harper & Row*, 471 U.S. at 561.

[193] See *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 19 ("When technological change has rendered its literal terms ambiguous, the Copyright Act must be construed in light of its basic purpose" (citing *Twentieth Century Music Corp. v. Aiken*, 422 U.S. 151,156 (1975))); *Kirtsaeng v. John Wiley & Sons, Inc.*, 579 U.S. 197, 198 (2016) ("[T]he well-settled

training would mean for artists' livelihoods. The Copyright Alliance warned that "the widespread unauthorized ingestion of copyrighted works would certainly appear to cause immeasurable harm to creators and copyright owners—both by destroying existing, nascent, and to-be-developed licensing markets and by flooding the market with low-quality substitutional material."[194] One creator wrote that "[t]he unregulated use of AI tools by companies and individuals is actively threatening my ability to get jobs in my field. It makes me feel sick that all the art I posted online to build my career, can be stolen at any time and used without my permission."[195]

Many of these commenters argued that the use of copyrighted works to create new expressive works that compete with or serve as substitutes for the originals cannot be considered fair.[196] AAP compared it to compelling authors to subsidize their competition.[197]

---

objectives of the Copyright Act . . . are to enrich the general public through access to creative works by striking a balance between encouraging and rewarding authors' creations and enabling others to build on that work." (citations and quotations omitted)).

[194] Copyright Alliance Initial Comments at 51; *see also* EGAIR Initial Comments at 367 ("[T]he pursuit of art could be relegated to the independently wealthy . . . . This will disproportionately harm the development of artists from marginalized communities, like disabled artists, and artists with dependents."); AAP Initial Comments at 1–2 ("The un-permissioned and uncompensated use of copyrighted works to create the datasets for training generative AI . . . presents a direct assault on the livelihoods and professions of authors, publishers, and all those who are integral to the publishing endeavor."); Center for Art Law Initial Comments at 2 ("[T]he nature of the output produced by AI, as well as the unprecedented scale, threatens the livelihood and rights of human copyright holders."); SGA-SCL-MCNA Joint Reply Comments at 8 ("[S]hould the fair use doctrine be perverted . . . to sanction such unauthorized generative AI uses of copyrighted works . . ., within a generation the . . . professional creator class would all but cease. Music creation would finally and fully be rendered a hobby, not a career. And without new human creation, the entire music sector would eventually suffer the same fate, to the grave detriment of American consumers, culture, the US trade balance, and the overall US economy." (footnote omitted)).

[195] Bonnie Smith Initial Comments at 1; *see also* Molly Gur Initial Comments at 1 ("I have worked as a creative professional for 15 years, and things have now become so troubling that something must change.").

[196] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 17 ("[T]he unauthorized reproduction of copyrighted works by AI developers to develop models that produce AI-generated works that actually or potentially compete with the inputted works comes as close as a use can come to being presumptively not fair use."); Artist Rights Alliance-FMC Joint Reply Comments at 1 ("Using copies of some or all of our songs and recordings to create new ones that compete with ours in the market is clearly not allowable under even the most expansive, wishful reading of 'fair use.'"); Marsha Blackburn Initial Comments at 2 ("[I]t is my belief that unlicensed AI ingestion of copyrighted works should not constitute fair use when the AI output supplants or competes commercially with the human-created works it was trained on."); Graphic Artists Guild Initial Comments at 9 ("It is difficult to square the scraping of images into datasets to train AI models with fair use" because AI platforms "generate visual content that mimics the images in the training dataset" and "AI output competes with, if not replaces, the original images."); Ed Newton-Rex Reply Comments at 1 (Dec. 6, 2023) ("Today's generative AI models can clearly be used to create works that compete with the copyrighted works they are trained on. So I don't see how using copyrighted works to train generative AI models of this nature can be considered fair use.").

[197] AAP Initial Comments at 13 ("[T]here are very limited circumstances in which the use of copyrighted works to train AI models would constitute fair use, given the lack of a transformative purpose combined with the harm to the

Rightsify stated, "[t]here is no legal precedent for the massive scraping of data for the purposes of creating data sets that can be commercially exploited to potentially create competing works."[198] Others contended that the unauthorized copying of expressive works in AI training adds nothing new while usurping an emerging market for works as training materials.[199]

On the other side, many warned that requiring AI companies to license works in training data would stifle development of a critical technology, entrench the power of those companies that are capable of acquiring or already own sufficient data, and impair national competitiveness. As summarized by the venture capital firm a16z, "imposing the cost of actual or potential copyright liability on the creators of AI models will either kill or significantly hamper their development . . . . The result will be far less competition, far less innovation, and very likely the loss of the United States' position as the leader in global AI development."[200] Stability AI called it "doubtful" that generative AI would be possible without the fair use defense and maintained that "[t]he U.S. has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation."[201]

---

market for and value of copyrighted works. . . . [A] rule that unauthorized use of copyrighted works to train AI models constitutes fair use essentially compels authors and publishers to subsidize the development of AI models.").

[198] Rightsify Initial Comments at 4. *See also* IAC-DDM Joint Initial Comments at 7 ("IAC-DDM Joint Initial Comments") ("[T]he massive and systematic copying of copyrighted content for an avowedly commercial and substitutive purpose does not present a hard or close case."); MPA Reply Comments at 20 ("None of the cases cited by AI developers involve the wholesale copying of expressive, non-functional works and the creation of a model that is then used to generate expressive works."); N/MA Reply Comments at 15 ("Case law has generally not permitted copying for purposes that do not comment on or at least point to the original works, outside of defined, limited exceptions, such as to access functional computer code for interoperability purposes."); UMG Initial Comments at 39–40 ("We can think of no precedent for finding this kind of wholesale, commercial taking that competes directly with the copyrighted works appropriated to be fair use.").

[199] *See* AP Initial Comments at 2–3 ("AI training adds nothing new to the original works; it merely uses them to create new works that supersede or supplant them. Because one purpose of news publishers' content is to license it to AI developers for model training, the use by AI developers 'share[s] the objectives' of news publisher's content."); New York Times Initial Comments at 4 ("New York Times Initial Comments") ("But taking expressive, protected content and using it to power tools that output close, sometimes verbatim, summaries of that very content is not transformative for purposes of copyright law. . . . GAI products use our content for purposes that are clearly commercial and harm The Times by creating output that is substitutive of our content.").

[200] a16z Initial Comments at 8. *See also* EFF Initial Comments at 4 ("The effect of requiring authorization would be to limit competition to companies that have their own trove of images or strike a deal with such a company, resulting in all the usual harms of limited competition (higher costs, worse service, security risks)."); Qualcomm Reply Comments at 3 ("[O]ther countries have embraced . . . the use of copyrighted material in AI training. To the extent the law in the United States differs or remains uncertain, there will be unavoidable incentives to move the development of AI tools . . . to other countries. These incentives jeopardize the United States' longstanding position as a global leader in the creative economy, particularly in the field of high technology and software development."); CCIA Initial Comments at 16.

[201] Stability AI Initial Comments at 13.

These commenters saw the use of copyrighted works to train AI models as consistent with fair use precedent. Some asserted that fair use generally favors technological advancements,[202] particularly where "intermediate copying" facilitates the development of new technologies.[203] Authors Alliance stated that "[i]n the vast majority of cases, the use of copyrighted works to train AI models constitutes fair use" because it is done as an intermediate step in producing non-infringing content and serves the public benefit by reducing bias in datasets and improving performance of AI models.[204] Meta asserted that AI training does not harm rightsholder interests because "the purpose and effect of training is not to extract or reproduce the protectable expression in training data, but rather to identify language patterns across a broad body of content."[205]

A. Factor One

The first fair use factor "focuses on whether an allegedly infringing use has a further purpose or different character, which is a matter of degree, and the degree of difference must be weighed against other considerations."[206] Courts typically stress two main elements: transformativeness and commerciality. Some courts have also evaluated whether the defendant had lawful access to the work.[207]

---

[202] *See* a16z Initial Comments at 7 ("Where copies of copyrighted works are created for use in the development of a productive technology with non-infringing outputs, our copyright law has long endorsed and enabled those productive uses through the fair use doctrine. Without the safeguard of fair use, we could not have now-ubiquitous technologies like internet search engines, online book search tools, and video game emulators." (internal citations omitted)); Adobe Initial Comments at 3 ("Fair use precedent dealing with 'significant changes in technology' make clear that use of copyrighted works for purposes like training AI models is transformative."); Meta Initial Comments at 13 ("Many courts . . . have recognized that the creation of copies of copyrighted works (especially copies that are not perceivable to the public) in the course of technological development of non-infringing, competing products is protected by fair use.").

[203] *See* TechNet Initial Comments at 4–5 ("But the creation of intermediate copies in furtherance of the creation of a new and useful technological tool is not the kind of copying that violates copyright law."); Adobe Initial Comments at 3–4 (describing *Sega v. Accolade* and *Sony Computer Entertainment, Inc. v. Connectix Corp.* applying fair use to intermediate copying necessary to reverse engineer access to unprotected functional elements within a program and analogizing to AI model training: "Inputs are temporarily accessed for the unprotected ideas, concepts, and styles contained in the dataset—say, the number of fingers a human hand has, or what cars look like—to help the AI model learn facts about the world" (footnotes omitted)).

[204] Authors Alliance Initial Comments at 9–10.

[205] Meta Initial Comments at 11.

[206] *Warhol*, 598 U.S. at 525.

[207] *See infra* Section IV.A.4.

1. Identifying the Use

"The fair use provision, and the first factor in particular, requires an analysis of the specific 'use' of a copyrighted work that is alleged to be an 'infringement . . . [as t]he same copying may be fair when used for one purpose but not another."[208] In *Andy Warhol Foundation v. Goldsmith* ("*Warhol*"), the photographer Lynn Goldsmith challenged the Foundation's unauthorized licensing of a screenprint of the musician Prince that Andy Warhol had created based on her copyrighted photograph. The Court's fair use analysis was based on the licensing of the screenprint rather than its initial creation decades before. On the first factor, it found that the licensing use had the purpose of display in a magazine, which was "substantially the same purpose" as Goldsmith's original photo.[209]

As described above, copyrighted works are used in different ways during the development and deployment of generative AI models. The use of a work in initial pre-training, for instance, may be distinct from its use in subsequent training or RAG. A number of commenters opined that the fair use analysis requires treating these different uses separately.[210] One observed that "[e]ven if a base model is deemed [noninfringing], downstream fine-tuned or aligned models may have a substantively different fair-use analysis."[211]

The Office agrees that different uses during AI development and deployment require separate consideration.[212] But while it is important to identify the specific act of copying during

---

[208] *Warhol*, 598 U.S. at 533. In *Warhol*, the Supreme Court identified the use of the copyrighted work in its analysis of factor one. For this reason, we likewise discuss this issue in connection with the first factor, but note that identifying the use is also a prerequisite to consideration of other factors.

[209] *Id.* at 550–51. Justice Gorsuch, concurring in the judgment, emphasized that "[i]f, for example, the Foundation had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art," the analysis would be different, and "the purpose and character of that use might well point to fair use." *Id.* at 557–58 (Gorsuch, J., concurring).

[210] Copyright Alliance Initial Comments at 60 ("[U]nder *Warhol*, different uses of a particular work should be considered separately, and it is possible that one use is considered to be transformative while the other is not."); CCC Initial Comments at 9 ("As fair use is fact dependent, different stages of training may have different analyses."); MPA Initial Comments at 21 ("[T]he relevant use will vary, both with the stage of training, scope of material used, and ultimate use of the outputs."); Public Knowledge Initial Comments at 8–12 (separately analyzing fair use as to the creation of AI training datasets and the use of those datasets in AI training). *But see* Rightsify Initial Comments at 5 ("As the training sets are created for the ultimate purpose of developing commercial models, the end purpose should be the only issue that matters. The intermediate steps are all part of the process and should not be analyzed separately under the first factor.").

[211] Katherine Lee et al. Initial Comments at 101.

[212] *See Warhol*, 598 U.S. at 533; *see also Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202 (2d Cir. 2015) (distinguishing between digitization of copyrighted works, creation of search functionality, display of snippets, and distribution of digital copies as separate uses); *Fioranelli v. CBS*, 551 F. Supp. 3d 199 (S.D.N.Y. 2021) (distinguishing between uses of video footage in certain documentary films, works focusing on conspiracy theories, political documentaries, and a feature film); *Chapman v. Maraj*, No. 18-cv-9088, 2020 WL 6260021 (C.D. Cal. Sept. 16, 2020)

development, compiling a dataset or training alone is rarely the ultimate purpose. Fair use must also be evaluated in the context of the overall use.[213]

2. Transformativeness

a) Legal Framework

In assessing transformativeness, the question is "whether the new work merely 'supersedes the objects' of the original creation, or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message. . . ."[214] Such a use is less likely to substitute for the original in the marketplace and more likely to advance the purposes of copyright.[215]

In *Warhol*, the Supreme Court clarified the concept of transformativeness. The Court explained that while adding new expression can be relevant to evaluating whether a use has a different purpose and character, it does not necessarily make the use transformative.[216] Even significant alterations will not be enough if the use ultimately serves a purpose similar to that of

(distinguishing between using a musical work to experiment in creating a new musical work with distributing a sound recording embodying that new work); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 176 (2d Cir. 2018) ("It is useful to analyze separately distinct functions of the secondary use (i.e., the use by TVEyes of Fox's copyrighted material), considering whether each independent function is a fair use.").

[213] *See Google Books*, 804 F.3d at 216–25 (looking beyond the digitization of the books to consider the ultimate uses of those copies to enable the system's search and snippet functions); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014*); Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606 (9th Cir. 2000).

[214] *Campbell*, 510 U.S. at 579. The concept of transformativeness was described as follows by Judge Pierre Leval in his influential article, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("I believe the answer to the question of justification turns primarily on whether, and to what extent, the challenged use is *transformative.* The use must be productive and must employ the quoted matter in a different manner or for a different purpose from the original. A quotation of copyrighted material that merely repackages or republishes the original is unlikely to pass the test; in Justice Story's words, it would merely 'supersede the objects' of the original. If, on the other hand, the secondary use adds value to the original -- if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings -- this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society.").

[215] *Campbell*, 510 U.S. at 579. Most of the paradigmatic examples listed in the preamble of section 107 ("criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research") reflect the types of purposes that courts have found to be transformative. *Warhol*, 598 U.S. at 528 (quoting 17 U.S.C. § 107; *Campbell*, 510 U.S. at 577–78).

[216] *Warhol*, 598 U.S. at 525 ("Although new expression may be relevant to whether a copying use has a sufficiently distinct purpose or character, it is not, without more, dispositive of the first factor."); *id.* at 544–45 ("[T]he meaning of a secondary work, as reasonably can be perceived, should be considered to the extent necessary to determine whether the purpose of the use is distinct from the original, for instance, because the use comments on, criticizes, or provides otherwise unavailable information about the original. (citation omitted)).

the original,[217] and may instead produce a derivative work and demonstrate the "need for licensing."[218]

The Court explained that "a use that has a distinct purpose is justified because it furthers the goal of copyright, namely, to promote the progress of science and the arts, without diminishing the incentive to create."[219] Such justification may be found when the copying "is reasonably necessary to achieve the user's new purpose."[220] For example, where a work is targeted for parody, criticism, or commentary, there is a need to use that particular work to effectively accomplish that purpose.[221] Using a work to communicate a new meaning or message unrelated to commenting on the work itself, however, does not provide such a justification.[222]

---

[217] *See Warhol*, 598 U.S. at 522, 525–526, 533–534 (Warhol's "Orange Prince crops, flattens, traces, and colors the photo but does not otherwise alter it. . . . [It] adds new expression to Goldsmith's photograph . . . [and] the first fair use factor still favors Goldsmith."); *id.* at 529 ("[T]he owner has a right to derivative transformations of her work. Such transformations may be substantial, like the adaptation of a book into a movie. To be sure, this right is '[s]ubject to' fair use. The two are not mutually exclusive. But an overbroad concept of transformative use, one that includes any further purpose, or any different character, would narrow the copyright owner's exclusive right to create derivative works. To preserve that right, the degree of transformation required to make 'transformative' use of an original must go beyond that required to qualify as a derivative." (citations omitted)).

[218] *See id.* at 529 n.5, 541.

[219] *Id.* at 532. *See also* Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("Factor One's direction that we 'consider[] . . . the purpose and character of the use' raises the question of justification. Does the use fulfill the objective of copyright law to stimulate creativity for public illumination?").

[220] *Warhol*, 598 U.S. at 532.

[221] *Id.* at 539–40 (2023); *Campbell*, 510 U.S. at 588 ("When parody takes aim at a particular original work, the parody must be able to 'conjure up' at least enough of that original to make the object of its critical wit recognizable."). Even where the use and the original share the same or highly similar purposes, the first factor may favor fair use where the use is justified. *See Warhol*, 598 U.S. at 532 ("An independent justification . . . is particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it."); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606–07 (9th Cir. 2000) (holding a use to be transformative "despite the similarities in function and screen output" between the use and the original work because the user's "product created a new platform," which was an "innovation [that] affords opportunities for game play in new environments").

[222] *Warhol*, 598 U.S. at 547 ("[B]ecause AWF's commercial use of Goldsmith's photograph to illustrate a magazine about Prince is so similar to the photograph's typical use, a particularly compelling justification is needed. Yet AWF offers no independent justification, let alone a compelling one, for copying the photograph, other than to convey a new meaning or message. As explained, that alone is not enough for the first factor to favor fair use."). *See also Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 452–55 (9th Cir. 2020) (concluding that defendant's use of creative elements of plaintiff's works was not transformative as it did not critique or comment on them, but rather mimicked them and paralleled their purpose).

The *Warhol* Court further emphasized that both transformativeness and justification are matters of degree.[223] "[T]he first factor (which is just one factor in a larger analysis) asks 'whether *and to what extent*' the use at issue has a purpose or character different from the original."[224] As the Court previously stated in *Campbell v. Acuff-Rose*, "the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use."[225] The degree to which a use is transformative can inform the analysis of market harm under the fourth factor, because less transformative uses are more likely to serve as market substitutes.[226] Further, although transformativeness often leads to a finding of fair use,[227] not every transformative use is a fair one.[228]

[223] *See Warhol*, 598 U.S. at 525, 528–29, 532.

[224] *Warhol*, 598 U.S. at 528 (quoting *Campbell*, 510 U.S. at 579). *See id.* at 532 ("Once again, the question of justification is one of degree."); Leval, *supra* note 214, at 1111 ("[I]t is not sufficient simply to conclude whether or not justification exists. The question remains how powerful, or persuasive, is the justification, because the court must weigh the strength of the secondary user's justification against factors favoring the copyright owner.").

[225] *Campbell*, 510 U.S. at 579; *see Warhol*, 598 U.S. at 529 ("The larger the difference [in purpose and character between the use and the original], the more likely the first factor weighs in favor of fair use. The smaller the difference, the less likely.").

[226] *Campbell*, 510 U.S. at 591. Similarly, the degree of transformativeness is relevant to the third factor analysis. *See id.* at 586–87 ("[T]he extent of permissible copying varies with the purpose and character of the use."); *infra* Section IV.C.2.

[227] Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 UNIV. PA. L. REV. 549, 605 (2008), https://scholarship.law.upenn.edu/penn_law_review/vol156/iss3/1 (finding that of fair use cases decided between 1978 and 2005, "each of the 13 circuit court opinions and 27 of the 29 district court opinions that found the defendant's use to be transformative also found it to be a fair use-and one of the two district court outliers was reversed on appeal").

[228] *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 177 (2d Cir. 2018) (finding that TVEyes service was "at least somewhat transformative" but that the balance of factors opposed fair use because "[a]t bottom, TVEyes is unlawfully profiting off the work of others by commercially re-distributing all of that work that a viewer wishes to use, without payment or license").

Beyond these general principles, case law provides additional guideposts. Uses that merely change the medium,[229] or spare the user inconvenience,[230] are not transformative. By contrast, copying to make available information about the content of the works copied can be transformative where it does not provide substitutes for those works.[231] For example, in *Google Books*, the Second Circuit found that scanning books to create a full-text searchable database to provide information about the books' contents served a "highly transformative purpose."[232]

Copying a work in order to remove functional impediments to competition may also be transformative even where the use enables the creation of competing works.[233] In *Google LLC v. Oracle America, Inc.*, the Supreme Court concluded that "reimplementation" of copied code was transformative because it "furthered the development of computer programs" by enabling programmers to use their existing skills in a new mobile platform.[234] Similarly, the Second Circuit held in *Sega v. Accolade* that copying computer code to learn the functional requirements

---

[229] *See Capitol Recs., LLC v. ReDigi Inc.*, 910 F.3d 649, 661 (2d Cir. 2018) ("ReDigi makes no change in the copyrighted work. It provides neither criticism, commentary, nor information about it. Nor does it deliver the content in more convenient and usable form to one who has acquired an entitlement to receive the content."); *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1262–63 (11th Cir. 2014) (finding no transformative use where excerpts of plaintiffs' works were digitized and used for "the same intrinsic purpose—or at least one of the purposes—served by Plaintiffs' works: reading material for students in university courses"); *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 108–09 (2d Cir. 1998) (holding that retransmission of radio broadcasts over the telephone merely repackaged or republished the original such that there was a "total absence of transformativeness in [defendant's] act of retransmission"); *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001) (affirming that it was not transformative to convert copyrighted songs from CDs to MP3 files for download because the "original work[s] [were] merely retransmitted in a different medium").

[230] *See Wall Data Inc. v. L.A. Cnty. Sheriff's Dep't*, 447 F.3d 769, 779–80 (9th Cir. 2006) (determining that the use was not transformative where, to speed up installation, defendant made exact copies of licensed software and used it for the same purpose as the original); *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 920, 922–24 (2d Cir. 1995) (concluding that photocopying for convenience was not a transformative use, but rather "part of a systematic process of encouraging employee researchers to copy articles so as to multiply available copies while avoiding payment").

[231] *See Google Books*, 804 F.3d at 216–17; *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014).

[232] *Google Books*, 804 F.3d at 216–18; *see also HathiTrust*, 755 F.3d at 97 (concluding that creation of a full-text searchable database was "quintessentially transformative" because "the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn"); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165 (9th Cir. 2007) ("Although an image may have been created originally to serve an entertainment, aesthetic, or informative function, a search engine transforms the image into a pointer directing a user to a source of information."); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 819 (9th Cir. 2003) (reasoning that the search engine "functions as a tool to help index and improve access to images on the internet and their related web sites"). *Cf. VHT, Inc. v. Zillow Grp., Inc.*, 918 F.3d 723, 740–43 (9th Cir. 2019) (finding that creation of a searchable database of full-size versions of real estate photographs was nontransformative because the database images ultimately served the same purpose as the originals, that is, "to artfully depict rooms and properties").

[233] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 30; *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 606 (9th Cir. 2000) (holding a use to be transformative "despite the similarities in function and screen output" between the use and the original work).

[234] *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 30.

for hardware-compatible games served a legitimate purpose that increased the "number of independently designed video game programs offered for use with the [hardware]."[235]

b) Commenters' Views

Commenters disagreed as to whether or to what extent the use of copyrighted works in the development of AI systems is transformative. Many viewed the process of generative AI training as highly transformative.[236] They saw the statistical analysis of works in machine learning as far removed in purpose and character from that of the original works.[237] The University Library of the University of California, Berkeley stated that "training [a] model to predict or classify aspects of copyright-protected inputs is a distinct purpose, and one that is highly transformative from the original 'consumptive' purpose."[238] Professors Samuelson, Sag, and Sprigman asserted that "[d]eriving uncopyrightable abstractions and associations from the

[235] *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522–23 (9th Cir. 1992); *see also Connectix Corp.*, 203 F.3d at 606 (holding that a PlayStation emulator was "modestly transformative" because it "create[d] a new platform, the personal computer, on which consumers can play games designed for the Sony PlayStation" and "[t]his innovation affords opportunities for game play in new environments"). Another court determined that copying an entire operating system was "moderately transformative" because the use added features to the program for the purpose of enabling security research. *Apple Inc. v. Corellium, Inc.*, No. 21-cv-12835, 2023 U.S. App. LEXIS 11225 at *20 (11th Cir. 2023) (per curiam). *But see Thomson Reuters v. Ross*, No. 20-cv-613-SB, slip op. at 17–18 (D. Del. Feb. 11, 2025) (finding legal research tool had same purpose as headnotes used to develop it and distinguishing reverse engineering cases on the basis that copying there was required to access functional elements of code).

[236] *See, e.g.*, CCIA Initial Comments at 8, 10; Van Lindberg Initial Comments at 20–21; Microsoft-Github Joint Initial Comments at 8; Meta Initial Comments at 12–13; Program on Information Justice and Intellectual Property ("PIJIP") Initial Comments at 5; EFF Initial Comments at 2; Creative Commons Initial Comments at 3–4. A few commenters discussed the creation of datasets separately from the training process as a whole. *See* Public Knowledge Initial Comments at 8 ("[Data sets] are transformative works, with minimal contribution of each constituent work to the overall value of the complete work, and the nature of their use is preliminary to non-infringing creative activity."); CCIA Initial Comments at 7.

[237] *See* Meta Initial Comments at 14 ("[M]odels use training data not to copy their content or challenge authors' ability to sell copies of their works, but rather to develop an entirely new and innovative service that, in turn, produces valuable new content—thereby vastly expanding the capacity for human creative productivity and the progress of science and the useful arts." (internal citations omitted)); Duolingo Initial Comments at 2–3; Tim Boucher Initial Comments at 11; BSA Initial Comments at 8; University of Illinois, Urbana-Champaign iSchool Initial Comments at 11 ("[A]bstracting high-level patterns from a corpus is in itself a transformative activity."); Engine Initial Comments at 6 ("Engine Initial Comments"); EFF Initial Comments at 3 ("[A]s *Google v. Oracle* teaches, a use that facilitates the creation of new works is more likely to be fair. As in *Google*, a model can be used for a range of expression informed by user prompts, conveying messages devised by users."); Katherine Lee et al. Initial Comments at 100 ("Many models *qua* models are arguably highly transformative. They represent works internally in new and very different ways.").

[238] University Library of the University of California, Berkeley Initial Comments at 5. *See also Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 640 (4th Cir. 2009); *Perfect 10 v. Amazon.com, Inc.*, 508 F.3d at 1165; *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003).

training data and then using that knowledge to confect new digital artifacts is not just transformative, it is highly transformative."[239]

Several commenters described the use of copyrighted works to train AI models as fundamentally different from the purposes of those works because it is "non-expressive." For example, Anthropic asserted that "[t]o the extent copyrighted works are used in training data, it is for analysis (of statistical relationships between words and concepts) unrelated to any expressive purpose of the work."[240] Google stated that because training is a process for "deconstructing existing works for the purposes of modeling mathematically how language works," it serves a different purpose than the "communicative, expressive purpose for which these works were created."[241] Another commenter opined that the difficulty in determining whether a model has been trained on a work is evidence that it is not intended to replicate the expressive material in its training data.[242] Some compared AI training to human learning, as evidence that it was productive and transformative.[243]

---

[239] Pamela Samuelson et al. Reply Comments at 14–15.

[240] Anthropic Initial Comments at 7. *See also* IBM Initial Comments at 4 ("The material is not being used for its expression, and the foundation model is not being trained to reproduce or compete with the original content."); Pamela Samuelson et al. Initial Comments at 14–15 ("Moreover, the non-expressive use of copyrighted works by generative AI use does not usurp the copyright owner's interest in communicating her original expression to the public because that expression is not communicated."); BSA Initial Comments at 7; C4C Initial Comments at 3 ("[Fair use] allows for copying for non-expressive (or non-consumptive) uses. . . . Copyrighted works used as input in AI models become part of a data set. They become tokens that integrate in a data collection and copyright protection does not apply to facts or ideas.").

[241] Google Initial Comments at 11.

[242] TechNet Initial Comments at 11 (Oct. 30, 2023) ("[T]he fact that the works on which a model has been trained cannot be readily be determined by users of that model is proof that training works . . . is not meant to replicate the content of the works but to simply extract the unprotectable elements of those works."); *cf.* Meta Initial Comments at 15–16 ("The purpose of the models is to extract enough *statistical* information about language and abstract concepts to enable the creation of *new* content—not to capture and reproduce expressive material from the training data itself."); Univ. of Ill. Urbana Initial Comments at 10 ("[M]odels are not created or marketed to reproduce specific works, and do not excel at the task.").

[243] Chamber of Progress Initial Comments at 6 ("Generative AI aligns more with human learning, where exposure to existing works shapes and influences fresh creations, rather than simply piecing together existing content."); Committee for Justice ("CFJ") Initial Comments at 6 ("[P]ointing to temporary copying by the developers of AI models is a weak peg to hang one's hat on. For one thing, it is not fundamentally different from what humans do when they learn from examples."); Meta Initial Comments at 13 ("[T]he process behind Generative AI is similar to human learning."). Some of these commenters draw on Mark A. Lemley and Bryan Casey's article, *Fair Learning*, which argues that "the use of copyrighted works by ML systems should be fair," because "people, like machines, often copy expression when they are only interested in learning the ideas conveyed by that expression," and frequently such uses are fair. 99 TEX. L. REV. 743, 749, 775–76 (2021). As discussed below, the Office sees the analogy between AI learning to human learning as faulty. *See infra* at Section IV.A.2.c.

A few commenters, citing *Warhol* for the proposition that a justification for a use may support its transformativeness,[244] argued that the mass use of works is justified as important or necessary to the development of AI technology.[245] IBM for example observed that "[t]he countless scientific, societal, and economic benefits that foundation models can provide more than justify the reproductions of copyrighted material in their training datasets."[246]

On the other side, many disagreed with the proposition that using copyrighted works in AI training is transformative.[247] Some described such use as similar to non-transformative processes like compression,[248] where the expressive elements of the works are simply represented in a different way. Others compared an AI model to a device loaded with copyright-infringing content: "Unlike a camera or VCR, generative AI is 'pre-loaded' by the developer with copyrighted content, and unlike a camera or VCR, AI uses that copyrighted content to generate its own (uncopyrightable) synthetic content."[249] They asserted that copying for AI training is unjustified because no individual work is necessary to train AI, and other means of acquisition, such as licensing, are available.[250]

---

[244] *Warhol*, 598 U.S. at 531–33.

[245] *See* Meta Initial Comments at 14 ("[T]here is an overwhelming justification for the copying; the ability to accurately distill the desired facts (whether about language or images or sounds) requires the ingestion of massive amounts of content that cannot reasonably be individually licensed."); Data Provenance Initiative Initial Comments at 10 (as a result of the increased size of datasets and new machine learning techniques, machine learning capabilities have greatly improved; "[t]he compelling justification for using underlying raw data is that machine learning models are designed to work better when trained on a broad range of content").

[246] IBM Initial Comments at 4 ("A contrary finding would severely limit the data available for foundation model training and significantly encumber AI development, thereby impeding the useful arts and sciences.").

[247] *See* NMPA Initial Comments at 16 ("[T]he use of musical works to train AI models is done for the purpose of creating new musical works that serve purposes that are substantially the same as those of the originals. If such a purpose could be considered 'transformative' it would make the copying of any musical work to create a new musical work a 'transformative' fair use, a notion the Supreme Court rejected."); Evangelical Christian Publishers Association ("ECPA") Initial Comments at 4; Center for Art Law Initial Comments at 3 ("Center for Art Law Initial Comments"); *see also* Graphic Artists Guild Initial Comments at 10; New York Times Initial Comments at 4; Writers Guild of America ("WGA") Initial Comments at 2; AP Initial Comments at 2–3.

[248] *See* A2IM-RIAA Joint Initial Comments at 15 ("We believe that AI models store representations of all or part of our recordings within their models, even if this is in compressed form."); John Patterson Initial Comments at 1 ("[T]he ability of generative AI to store works as 'training' is just a more sophisticated form of compression.").

[249] UMG Initial Comments at 12–13. *See* Getty Images Reply Comments at 7–8.

[250] *See, e.g.*, AAP Initial Comments at 15–17 ("No one specific work will be necessary for [machine learning], and developers will generally have a range of substitutes or alternatives that achieve the same purpose. . . . Thus, the justification for using copyrighted works in training AI models is negligible."); Copyright Alliance Initial Comments at 58–60 ("[S]imply because licensing is not a financially desirable avenue for AI developers does not mean unauthorized use is justified.").

A number of commenters opined that when analyzing the purpose and character of an AI developer's use of copyrighted material, courts should not view the training process in isolation but consider the ultimate use of the model.[251] In addition, one commenter observed that "[d]ifferent stages like pre-training and fine-tuning could . . . raise distinct considerations under the first fair use factor" as "[f]ine-tuning . . . usually narrows down the model's capabilities and might be more aligned with the original purpose of the copyrighted material."[252]

Several commenters disputed the characterization of training on copyrighted works as "non-expressive."[253] As an initial matter, the MPA observed that courts have never said there is a "non-expressive use" doctrine: "The relevant inquiry is not whether the 'use' is 'expressive' or 'non-expressive'; rather, it asks whether the 'use' is transformative, as one consideration in the four-factor analysis."[254] Others rejected the claim that AI training uses only the ideas or facts embodied in a work.[255] In the words of the Authors Guild, "AI companies seek out published books for [training] precisely because of their expressive content, as high-quality, professionally authored works are vital to enabling an LLM to produce outputs that mimic human language, story structure, character development, and themes."[256] AAP asserted that "Gen AI training . . . does not extract the ideas, facts, or concepts being conveyed by an author, it solely extracts the exact expressive choices made to convey those ideas—i.e., the words an author used, and the order in which they were placed."[257] These commenters further argued that the cases cited in

---

[251] *See, e.g.*, Copyright Alliance Reply Comments at 4 ("[T]he purpose of generative AI cannot be considered in a vacuum of 'training.'"); Brooklyn Law Incubator & Policy Clinic ("BLIP") Initial Comments at 10; Lee Hollaar Initial Comments at 1; Rightsify Initial Comments at 5.

[252] Seth Polansky Initial Comments at 11.

[253] *See, e.g.*, Katherine Lee et al. Initial Comments at 102 ("[T]he non-expressive use argument fails once the dataset is an input into generative models that can produce outputs that reproduce copyrighted expression."); Center for Art Law Initial Comments at 4; David Opderbeck Initial Comments at 16–26.

[254] MPA Reply Comments at 18; *see also* Copyright Alliance Reply Comments at 12–13 ("[N]either the courts nor Congress have ever espoused a broad category of fair use called 'non-expressive use.'").

[255] N/MA Reply Comments at 10; HarperCollins Publishers Reply Comments at 2 ("[W]hat all these [companies] are extracting or analyzing is the way in which authors have expressed themselves – precisely so [they] can emulate that expression convincingly for AI users."); NMPA Initial Comments at 14 ("They train on expressive works to generate other expressive works. They copy expression for expression's sake."); International Confederation of Societies of Authors and Composers ("CISAC") Reply Comments at 3 n. 16 ("[I]n the AI developers' selection and filtering of training data, expressive works are sought out specifically for their expressive value, meaning that an original work is capable of being replicated by AI in an identical way to the original expression") (citing Michael Frank Initial Comments at 4–6 (discussing aesthetic scoring)); AAP Reply Comments at 5–6.

[256] The Authors Guild Initial Comments at 17.

[257] AAP Reply Comments at 5–6.

support of the concept of non-expressive use relate to computer programs and are distinguishable from the use of expressive works in generative AI training.[258]

c) Analysis

As discussed above, *Warhol* requires examining not just the immediate act of copying but its ultimate goal.[259] Accordingly, whether copying a work to compile a training dataset is transformative depends on whether the dataset will be used for a transformative purpose.

In the Office's view, training a generative AI foundation model on a large and diverse dataset will often be transformative. The process converts a massive collection of training examples into a statistical model that can generate a wide range of outputs across a diverse array of new situations. It is hard to compare individual works in the training data—for example, copies of *The Big Sleep* in various languages—with a resulting language model capable of translating emails, correcting grammar, or answering natural language questions about 20th-century literature, without perceiving a transformation. The purpose of creating works of authorship is to disseminate them for human enjoyment and education. Many AI models, however, are meant to perform a variety of functions, some of which may be distinct from the purpose of the copyrighted works they are trained on.[260] For example, a language model can be used to help learn a foreign language by chatting with users on diverse topics and offering corrective feedback.[261]

---

[258] *See* Anonymous AI Technical Writer Reply Comments at 9 ("Google v. Oracle America . . . is not relevant . . . because, as the ruling notes, the code in question was an API. . . . It would be like trying to copyright the words of a formal wedding invitation."); The Authors Guild Reply Comments at 5; DMLA Initial Comments at 9.

[259] *See Warhol*, 598 U.S. at 550–51; *see also Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97–105 (2d Cir. 2014) (looking beyond copying books for "digitization" and "storage" to consider the ultimate purposes of those uses: full-text search functionality, accessibility for the print-disabled, and preservation); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 601, 606–07 (9th Cir. 2000) (evaluating purpose of copying to reverse engineer software by considering the final product that the reverse engineering enabled); *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 108–09 (2d Cir. 1998); *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001); *Flo & Eddie, Inc. v. Sirius XM Radio, Inc.*, 821 F.3d 265, 270 n.4 (2d Cir. 2016) ("The fair-use analysis applicable to [creating an internal database of recordings] … is bound up with whether the ultimate use of the internal copies [to make public performances] is permissible."). *See supra* Section IV.A.1.

[260] *Cf. Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202, 216–18 (2d Cir. 2015); *HathiTrust*, 755 F.3d at 97; *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1165; *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 819 (9th Cir. 2003).

[261] *Cf.* Stavros Athanassopoulos, *The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom*, Advances in Mobil Learning Educational Research (2023), https://www.syncsci.com/journal/AMLER/article/view/AMLER.2023.02.009; *see also OpenAI and the California State University System Bring AI to 500,000 Students and Faculty*, OPENAI (Feb. 4, 2025), https://openai.com/index/openai-and-the-csu-system/; Linda Kane, *Space Force Generative AI Challenge Empowers Guardians Through Education, Collaboration*, U.S. SPACE FORCE (Dec. 10, 2024), https://www.spaceforce.mil/News/Article-Display/Article/3992437/space-force-generative-ai-challenge-empowers-guardians-through-education-collab/; Severin Rodler et al., *Generative Artificial Intelligence in Surgery*, 175 SURGERY 1496 (2024), https://www.sciencedirect.com/science/article/abs/pii/S0039606024001193?via%3Dihub.

But transformativeness is a matter of degree, and *how* transformative or justified a use is will depend on the functionality of the model and how it is deployed. On one end of the spectrum, training a model is most transformative when the purpose is to deploy it for research,[262] or in a closed system that constrains it to a non-substitutive task. For example, training a language model on a large collection of data, including social media posts, articles, and books, for deployment in systems used for content moderation does not have the same educational purpose as those papers and books.

On the other end of the spectrum is training a model to generate outputs that are substantially similar to copyrighted works in the dataset. For example, a foundation image model might be further trained on images from a popular animated series and deployed to generate images of characters from that series. Unlike cases where copying computer programs to access their functional elements was necessary to create new, interoperable works, using images or sound recordings to train a model that generates similar expressive outputs does not merely remove a technical barrier to productive competition. In such cases, unless the original work itself is being targeted for comment or parody, it is hard to see the use as transformative.[263]

Many uses fall somewhere in between. The use of a model may share the purpose and character of the underlying copyrighted works without producing substantially similar content. Where a model is trained on specific types of works in order to produce content that shares the purpose of appealing to a particular audience, that use is, at best, modestly transformative. Training an audio model on sound recordings for deployment in a system to generate new sound recordings aims to occupy the same space in the market for music and satisfy the same consumer desire for entertainment and enjoyment. In contrast, such a model could be deployed for the more transformative purpose of removing unwanted distortion from sound recordings.

Because generative AI models may simultaneously serve transformative and non-transformative purposes,[264] restrictions on their outputs can shape the assessment of the purpose and character of the use. As described above, developers can apply training techniques or deployment guardrails so that the model rejects requests for excerpts of

---

[262] *Cf. Google Books*, 804 F.3d at 217 ("[T]he purpose of Google's copying of the original copyrighted books is to make available significant information about those books, permitting a searcher to identify those that contain a word or term of interest, as well as those that do not include reference to it. In addition, through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell as strongly favoring satisfaction of the first factor.").

[263] The decision to train on expressive works when there are available alternatives may itself reflect a lack of transformative purpose. For example, an image model could be trained on mass image data collected through automated means (street-view cars, body cameras, security cameras), yet developers often choose aesthetic images such as stock photography. This suggests the purpose is not simply to generate images of the physical world, but to generate images that have expressive qualities like the originals.

[264] As described above, the strength of generative pre-trained language models is their ability to perform well on a variety of tasks when given natural language directions. *See supra* notes 39–41, 102–103.

copyrighted works or even refuses to generate expressive works. Where such restrictions are effective, the system will be less capable of fulfilling the purpose of the original works, and their use in training may be more transformative.

The use of copyrighted works by RAG[265] requires separate consideration. Unlike pre-training where a large, diverse dataset is used to train a model for a wide variety of tasks, RAG retrieves individual works because they are relevant to a user's prompt, for the purpose of enhancing the response. The use of RAG is less likely to be transformative where the purpose is to generate outputs that summarize or provide abridged versions of retrieved copyrighted works, such as news articles, as opposed to hyperlinks.[266]

In providing this analysis, the Office rejects two common arguments about the transformative nature of AI training. As noted above, some argue that the use of copyrighted works to train AI models is inherently transformative because it is not for expressive purposes.[267] We view this argument as mistaken. Language models are trained on examples that are hundreds of thousands of tokens in length, absorbing not just the meaning and parts of speech of words, but how they are selected and arranged at the sentence, paragraph, and document level—the essence of linguistic expression.[268] Image models are trained on curated datasets of aesthetic images because those images lead to aesthetic outputs.[269] Where the

---

[265] *See supra* text accompanying notes 130–133.

[266] *Cf. L.A. News Serv. v. Reuters Television Int'l, Ltd.*, 149 F.3d 987, 990, 993–94 (9th Cir. 1998); *Monge v. Maya Mags., Inc.*, 688 F.3d 1164, 1174 (9th Cir. 2012) (summarizing Ninth Circuit cases); *Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.*, 166 F.3d 65, 72 (2d Cir. 1999) (concluding defendants' "abstracts" of news articles were "not in the least 'transformative'" because they were for the most part direct translations with little or no new expression added); *Twin Peaks Prods., Inc. v. Publ'ns Int'l, Ltd.*, 996 F.2d 1366, 1375–76 (2d Cir. 1993) (concluding that "detailed report of the plots goes far beyond merely identifying their basic outline for the transformative purposes of comment or criticism" and that defendant's plot synopses were essentially "abridgments"); *Penguin Random House LLC v. Colting*, 270 F. Supp. 3d 736, 750–51 (S.D.N.Y. 2017) (holding defendants' abridgements to remove adult themes were not fair use); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 551–57 (S.D.N.Y. 2013) (finding defendant's republication of text from news articles without adding commentary or insight was nontransformative); UNITED STATES COPYRIGHT OFFICE, COPYRIGHT PROTECTIONS FOR PRESS PUBLISHERS 37–44 (June 2022), available at https://www.copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf. *Cf. Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1168; *Kelly v. Arriba Soft*, 336 F.3d 811, 818–20 (9th Cir. 2003); *Google Books*, 804 F.3d 202, 215–18 (2d Cir. 2015).

[267] *See supra* text accompanying notes 240–242.

[268] *Cf. The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3–4 (describing books as "invaluable for . . . coherent storytelling."); Jack Lindsey et al., On the Biology of a Large Language Model, TRANSFORMER CIRCUITS (Mar. 27, 2025), https://transformer-circuits.pub/2025/attribution-graphs/biology.html (identifying internal model mechanisms believed to be associated with writing poetry, including planning mechanisms related to rhyming); Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RSCH. 1137, 1138 (2003), https://jmlr.csail.mit.edu/papers/volume3/bengio03a/bengio03a.pdf (describing a statistical model of language focused on learning "the distribution of word sequences, rather than learning the role of words in a sentence").

[269] *See supra* note 82.

resulting model is used to generate expressive content, or potentially reproduce copyrighted expression, the training use cannot be fairly characterized as "non-expressive."[270]

Nor do we agree that AI training is inherently transformative because it is like human learning.[271] To begin with, the analogy rests on a faulty premise, as fair use does not excuse all human acts done for the purpose of learning.[272] A student could not rely on fair use to copy all the books at the library to facilitate personal education; rather, they would have to purchase or borrow a copy that was lawfully acquired, typically through a sale or license.[273] Copyright law should not afford greater latitude for copying simply because it is done by a computer. Moreover, AI learning is different from human learning in ways that are material to the copyright analysis. Humans retain only imperfect impressions of the works they have experienced, filtered through their own unique personalities, histories, memories, and worldviews. Generative AI training involves the creation of perfect copies with the ability to analyze works nearly instantaneously. The result is a model that can create at superhuman speed and scale. In the words of Professor Robert Brauneis, "Generative model training transcends the human limitations that underlie the structure of the exclusive rights."[274]

3. Commerciality

The commerciality inquiry relates to the potential unfairness of using copyrighted works to obtain a financial benefit while forgoing payment.[275] Because even paradigmatic fair uses, such as news reporting or criticism, are often done for profit, "the crux of the profit/nonprofit distinction is not whether the sole motive of the use is monetary gain but whether the user stands to profit from exploitation of the copyrighted material without paying the customary price."[276]

---

[270] *Cf.* Katherine Lee et al. Initial Comments at 102 ("Even if [a training dataset] is also used to train non-generative-AI models, the non-expressive use argument fails once the dataset is an input into generative models that can produce outputs that reproduce copyrighted expression.").

[271] *See supra* note 243.

[272] *See* Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 48 Colum. J.L. & Arts 1, 12–13 (2025); Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45 (2017) ("No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on.").

[273] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *see also* Sobel, *supra* note 275.

[274] Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 48 Colum. J.L. & Arts 1, 38–39, 58–59 (2025).

[275] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *Soc'y of Holy Transfiguration Monastery, Inc.*, 689 F.3d 29, 61 (1st Cir. 2012) ("'Profit,' in this context, is thus not limited simply to dollars and coins; instead, it encompasses other non-monetary calculable benefits or advantages." (citation omitted)).

[276] *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562 (1985).

The Office's NOI asked how to assess commerciality in the context of generative AI, particularly in circumstances when curating datasets or training on those datasets may be done for noncommercial or research purposes, but the dataset or model is later adapted to commercial use.[277] Several commenters warned against considering such practices as noncommercial and described them as "data laundering."[278] News/Media Alliance stated that "in light of concerning practices of . . . initially nonprofit models that transition into commercial entities or assist them in building competitive, commercial products, the Office should be careful in drawing any kind of a bright line between commercial and noncommercial uses."[279]

The Office also asked whether it made a difference if the funding for non-commercial research uses came from for-profit companies.[280] While one commenter stated that funding from a commercial source "may be evidence of a commercial purpose," particularly if done as part of a "data laundering" arrangement,[281] others believed that it made no difference.[282]

---

[277] *See* NOI at 59946.

[278] *See, e.g.*, Copyright Alliance Initial Comments at 31 ("Data laundering entails private, commercial AI companies funding research or nonprofit institutions to develop training datasets and sometimes even the AI tools themselves, which often use copyright-protected works, under the guise of supporting noncommercial research activities. Once these training sets or models are developed, the funding AI company then uses them to develop proprietary commercial AI platforms."); A2IM-RIAA Joint Initial Comments at 12 ("When it comes to collection and curation, we have witnessed a disturbing practice of willfully disaggregating the creation of datasets for AI training, often by entities that claim to be non-profit or research-focused, and the actual training of AI models, often by for-profit, commercial entities. . . . [S]uch disaggregation can easily result in so-called 'data laundering,' whereby the developer of a commercial AI model seeks to avoid copyright infringement liability by claiming that the dataset from which it ingested copyrighted works was built for research purposes."); AAP Initial Comments at 15 & n.51 ("Once trained, an AI model can be deployed by multiple third parties, including commercial entities, so the noncommercial nature of the entity developing the model does not mitigate the likely harm to copyright owners that will result."); EGAIR Initial Comments Attachment at 370; N/MA Initial Comments at 37–38; NMPA Initial Comments at 18; UMG Initial Comments at 55.

[279] N/MA Initial Comments at 37–38; *see also* MPA Initial Comment at 23–24 ("Even a noncommercial purpose in the creation of a work (like Andy Warhol's art) can become a commercial purpose when that work is put to a different use (*i.e.*, licensing to a magazine). Thus, 'for-profit' and 'nonprofit' labels are not dispositive."). Another commenter suggested that this practice meant that commerciality analysis would often turn on the activities of downstream actors rather than dataset curators. Katherine Lee et. al. Initial Comments at 102 ("[M]any training datasets are made publicly available noncommercially. Some observers have argued that this amounts to a kind of ethical and legal laundering by the commercial companies that then train on those datasets — especially when there is a funding relationship between the two. The factor-one commerciality analysis of the dataset may therefore turn on the activities of parties besides the dataset curator.").

[280] NOI at 59946.

[281] Copyright Alliance Initial Comments at 65.

[282] AIPLA Initial Comments at 8 ("The source of funding currently does not play a meaningful role in the law of fair use, nor do we see any unique characteristics of AI, as currently constituted, which would compel a different result with respect to this technology."); DMLA Initial Comments at 10 ("It should not make a significant difference if

Because there are distinct acts and often multiple actors involved in the creation of AI systems, identifying the use with particularity is critical here too. The creation and distribution of a training dataset, the copying of that dataset for training, and the copying and distribution of model weights for use in a system may be conducted by different entities, each of whose activities may or may not be considered "commercial." Accordingly, in assessing whether the transfer of training datasets, synthetic data, or model weights is obscuring a commercial benefit and constitutes "data laundering," the financial relationships between the actors are relevant.

Moreover, commerciality does not turn solely on whether an organization is designated as "profit" or "nonprofit," but whether the use itself furthers commercial purposes.[283] A for-profit company with a substantial research arm could train a model with a novel architecture or training technique and release a research paper without commercializing the model. It could also "open source" the resulting model weights (*i.e.*, provide them to the public for free), leaving it to others to experiment or build products with them.[284] Although these activities could indirectly further the financial interests of the company, the connection between the copying and any commercial gain may be too attenuated to render the use commercial.[285]

Similarly, the nonprofit status of an organization should not in itself preclude a finding of commerciality.[286] Nonprofits may engage in commercial activity by directly monetizing

---

funding for noncommercial or research uses is provided by for-profit developers of AI systems, as much research is funded by for-profit tech organizations.").

[283] *See Warhol*, 598 U.S. at 537; *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001).

[284] Companies that release open weight models may nevertheless engage in a commercial use if they deploy them in their own monetized products and services or require licensing from large-scale commercial users.

[285] *See Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 185 (2d Cir. 2024) ("Any link between the funds [from an affiliate partnership] and its use of the Works is too attenuated for us to characterize the use as commercial on that basis."); *Bouchat v. Balt. Ravens Ltd. P'ship*, 737 F.3d 932, 948 (4th Cir. 2013), as amended (2014) (finding the "commercial character" of defendant's use of a logo in a display in a football stadium's higher-priced club level to be "attenuated," as "[n]o one is putting down hundreds of dollars" to see it); *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 922; *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 83 (2d Cir. 2014); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003) (finding "use of [copyrighted work] was more incidental and less exploitative in nature than more traditional types of commercial use" in which commercial defendant did not use works directly to promote itself or sell them); *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992) ("[T]he use at issue was an intermediate one only and thus any commercial "exploitation" was indirect or derivative.").

[286] *Compare Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1118 (9th Cir. 2000) (concluding that use "unquestionably profits [defendant]" because defendant "gained an "advantage" or "benefit" from its distribution and use" in the form of new members that contributed to defendant's growth "at no cost"), *and Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory*, 689 F.3d 29, 61 (1st Cir. 2012) (finding that a monastery archbishop "profited" from the use, and even if it did not generate "actual financial income . . . he benefitted by being able to provide, free of cost, the core text of the Works to members of the Orthodox faith, and by standing to gain at least some recognition within the Orthodox religious community"), *with Hachette Book Grp. v. Internet Archive*, 115 F.4th at 186 (use noncommercial where defendant "obtain[ed] only those nonmonetary benefits that attend most other legitimate, secondary uses, including advancing its mission and bolstering its reputation" and noting that

datasets or models through licensing or subscription-based products. Such direct monetization would be commercial notwithstanding an organization's corporate structure or charitable goals.[287]

In short, the analysis should not turn on the status of any individual entity but on the reality of whether the specific use in question serves commercial or nonprofit purposes.

4. Unlawful Access

A number of commenters contended that the first factor analysis should also take into account whether the AI developer had lawful access to the works used in training.[288] They reported that it is common for training datasets to include pirated works or works accessed by circumventing paywalls.[289] Some concluded that the "[i]f generative AI developers know or should have known that their systems are ingesting works that have been made available illegally, these acts would reflect bad faith or unclean hands."[290] Professors Samuelson,

---

"[c]haracterizing these general benefits as commercial profits would render commercial the activities of virtually any nonprofit organization that bolsters its reputation through its own nonprofit activities") *and Am. Soc'y for Testing & Materials, et al. v. Public.Resource.Org, Inc.*, 896 F.3d 437, 449 (D.C. Cir. 2018) (dismissing argument that free distribution of copyrighted industry standards was commercial because it enhanced a nonprofit organization's fundraising appeal as "hardly ris[ing] to the level of making [it] a 'commercial' use).

[287] *Cf. Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 186 (concluding non-profit organization's use was noncommercial in nature because defendant "d[id] not profit directly from its exploitation of the Works").

[288] *See, e.g.*, Copyright Alliance Initial Comments at 52–53; The Authors Guild Reply Comments at 6 (Dec. 6, 2023) ("[T]he knowing use of pirated copies in the service of a commercial enterprise should weigh against fair use."). Other commenters suggested that lawfulness of the training materials would be relevant to the fourth factor, insofar as using a pirated copy deprives the rightsholder of a sale or licensing fee. *See, e.g.*, DMLA Initial Comments at 8–9. Still others agreed that the use of pirated works should weigh against fair use generally but did not specify under which factor. *See* Pearson Initial Comments at 6; STM Initial Comments at 8. The FTC commented that "under certain circumstances, the use of pirated or misuse of copyrighted materials could be an unfair practice or unfair method of competition under Section 5 of the FTC Act." U.S. Federal Trade Commission Initial Comments at 5. *See infra* Section IV.F.

[289] *See supra* note 75; Copyright Alliance Initial Comments at 26–28 ("*The Washington Post* discovered that Google's C4 dataset . . . contains copyrighted works that are located behind a firewall on subscription-based websites . . . . Moreover, this dataset also included pirated books scraped from . . . a notorious pirate website . . . . Many other training datasets have the same issues." (footnote omitted)); StakeOut.AI Reply Comments at 2 ("AI researchers have found that generative AI training sets consist of files downloaded from pirate book repositories such as Library Genesis and Z-Library."); European Writers' Council ("EWC") Initial Comments at 8 ("AI companies have been pulling copyrighted book works from bit torrent piracy sites since 2013[]. The corpus Book3 and The Pile was proven to contain 194,000 identified titles." (emphasis omitted) (footnotes omitted)); CCC Initial Comments at 5; Graphic Artists Guild Initial Comments at 8 ("The LAION Database which was used to develop the diffusion model powering many of the AI image generators indiscriminately scraped over 5 billion images from online sources, including . . . . piracy websites"); AAP Initial Comments at 8–9; IBM Initial Comments at 3; DMLA Initial Comments at 6.

[290] N/MA Initial Comments at 44; Copyright Alliance Initial Comments at 52–53; The Authors Guild Reply Comments at 6 (Dec. 6, 2023).

Sprigman, and Sag, however, cautioned that "context matters[,]" and "[i]t would be unwise to elevate lawful access to a *per se* rule."[291]

In the Office's view, the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use without being determinative.[292] Courts have expressed some uncertainty about whether good or bad faith generally is relevant to the fair use analysis.[293] The cases in which they have done so, however, involved defendants who used copyrighted works despite the owners' denial of permission. Training on pirated or illegally accessed material goes a step further.[294] Copyright owners have a right to control access to their works, even if someone seeks to obtain them in order to make a fair use.[295] Gaining unlawful access therefore bears on the character of the use.

---

[291] Pamela Samuelson et al. Initial Comments at 24 ("Moreover, prohibiting academic research on illegal text corpuses will generally not benefit copyright owners or further the interests copyright is designed to promote."). *See, e.g.*, *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 482 (2d Cir. 2004) ("Even a finding of bad faith by defendants would not automatically preclude finding that their use was fair use").

[292] *See Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562–63 (1985) (finding a user's knowing exploitation of a "purloined manuscript" to be relevant to the character of the use); *Chi. Bd. of Educ. v. Substance, Inc.*, 354 F.3d 624, 628 (7th Cir. 2003) (explaining that the fact that the access was unauthorized "does not exclude the possibility of a fair use defense."); *L.A. News Serv. v. KCAL-TV Channel 9*, 108 F.3d 1119, 1122 (9th Cir. 1997) (finding relevant to the propriety of the user's conduct that, after being refused a license, defendant then "obtained a copy of the tape. . . directly copied the original, superimposed its logo . . . and used it for the same purpose for which it would have been used had it been paid for"); *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."); *cf. Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d at 1164 n.8 ("[W]e conclude that Google's inclusion of thumbnail images derived from infringing websites in its Internet-wide search engine activities does not preclude Google from raising a fair use defense.").

[293] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 32–33 ("As for bad faith, our decision in *Campbell* expressed some skepticism about whether bad faith has any role in a fair use analysis. We find this skepticism justifiable, as '[c]opyright is not a privilege reserved for the well-behaved.'" (citations omitted)); *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 585 n.18 (1994) ("[W]e reject [plaintiff's] argument that [defendant's] request for permission to use the original should be weighed against a finding of fair use."); *see also* Pierre N. Leval, *Campbell As Fair Use Blueprint?*, 90 WASH. L. REV. 597, 612–14 (2015); Simon Frankell & Matt Kellogg, *Bad Faith and Fair Use*, 60 J. COPYRIGHT SOC'Y USA 1 (2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2165468; Pierre N. Leval, *Toward A Fair Use Standard*, 103 HARV. L. REV. 1105, 1126–28 (1990).

[294] In addition, some commenters suggested that, to the extent copyright owners "opt out" of having their material used to train AI, whether through terms of use, the robots.txt instructions, or other means, a defendant's decision to ignore such opt-outs might inform the fair use analysis. *See, e.g.*, N/MA Initial Comments at 44; Pamela Samuelson et al. Initial Comments at 24 ("A defendant's . . . disregard of robots.txt exclusions and similar mechanisms could each be framed in terms of an argument against fair use under the fourth factor.").

[295] *Cf.* 17 U.S.C. § 1201 (prohibiting circumvention of technological protection measures used by copyright owners to control access to their works); Samuelson, et al. Initial Comments at 24 ("One might argue that although copyright owners do not have a right to charge for fair uses as such, they do have a right to charge for access to their works. As such, it may be deemed harmful or unfair for commercial users to bypass the market for access to train their LLMs without a compelling reason.").

B. Factor Two

The second factor, the nature of the copyrighted work, "calls for recognition that some works are closer to the core of intended copyright protection than others."[296] The use of more creative or expressive works (such as novels, movies, art, or music) is less likely to be fair use than use of factual or functional works (such as computer code).[297] The unpublished nature of a work can also weigh against a fair use determination.[298]

Those commenters who discussed the second factor asserted that it will often weigh against fair use because training datasets usually include expressive works, even if they contain less creative or unprotectable material as well.[299] While some noted that datasets may include unpublished works,[300] most works will have been published, which "modestly supports a fair

---

[296] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994); *see also* 17 U.S.C § 107(2); *Google LLC v. Oracle Am., Inc.*, 593 U.S. at 29.

[297] *See Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 563 (1985) ("The law generally recognizes a greater need to disseminate factual works than works of fiction or fantasy."); *Stewart v. Abend*, 495 U.S. 207, 237 (1990) (instructing that "fair use is more likely to be found in factual works than in fictional works" whereas "a use is less likely to be deemed fair when the copyrighted work is a creative product"). *See also Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 455 n.40 ("Copying a news broadcast may have a stronger claim to fair use than copying a motion picture."). Although Congress amended section 107 to clarify that the unpublished nature of a work is not dispositive, see Pub. L. No. 102–492, 106 Stat. 3145 (1992) (adding the text "[t]he fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors"), courts continue to consider publication status under factor two. *See Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d at 87 ("Whether or not a work was published thus enters into our analysis of this factor as a judicial gloss on 'the nature of the copyrighted work.' That gloss, of course, is firmly grounded in fair use's common law origins and the legislative history of the 1976 Copyright Act.").

[298] *See Harper & Row*, 471 U.S. at 549 (finding significant that the defendant, in copying unpublished excerpts for its magazine, had "arrogated to itself the right of first publication, an important marketable subsidiary right"); *Swatch Grp. Mgmt. Servs.*, 756 F.3d at 87 ("Whether or not a work was published thus enters into our analysis of this factor as a judicial gloss on 'the nature of the copyrighted work.'").

[299] *See, e.g.*, The Authors Guild Initial Comments at 20 ("The second factor would weigh against fair use where the works are highly creative and closer to the heart of copyright. As such, training on books for instance would weigh against fair use, whereas perhaps the use of functional and standard code would weigh in favor of fair use."); DMLA Initial Comments at 8–9 ("[W]orks ingested will frequently be expressive or creative, as in the case of visual artworks."); NMPA Initial Comments at 18 ("Where AI models are trained on expressive works, such as musical compositions or sound recordings, this factor will almost always weigh against a finding of fair use."); Graphic Artist Guild Initial Comments at 10. *But see* Authors Alliance Initial Comments at 11.

[300] *See, e.g.*, Copyright Alliance Initial Comments at 49 (stating that not all the expressive material posted to the internet and scraped for AI training will meet the legal definition of "published" (internal citations omitted)); Graphic Artist Guild Initial Comments at 10.

use argument."[301] Several observed, however, that the second factor rarely plays a substantial role in the overall fair use balancing.[302]

As generative AI models are regularly trained on a variety of works—both expressive and functional, published as well as unpublished—the facts will vary depending on the model and works at issue. Language models are often trained on highly creative works like novels, alongside those with more factual or functional content, like computer code or scholarly articles.[303] Where the works involved are more expressive, or previously unpublished, the second factor will disfavor fair use.[304]

C. Factor Three

On the third factor, the question is whether "'the amount and substantiality of the portion used in relation to the copyrighted work as a whole,' . . . are reasonable in relation to the purpose of the copying."[305] This factor "harken[s] back to the first [factor]" because "[t]he extent of permissible copying varies with the purpose and character of the use."[306] It also bears on the fourth factor insofar as more extensive copying can increase the risk that the use will serve as a market substitute for the original.[307] Relevant considerations may include how much of each work is used; the reasonableness of the amount in light of the purpose of the use; and the amount made accessible to the public.

---

[301] New Media Rights Initial Comments at 16 ("Here, it appears that ChatGPT is only using published works (which modestly supports a fair use argument)"). *See also* Data Provenance Initiative Initial Comments at 10–11 ("Most of the raw data will typically be published within the meaning of copyright law, which may also tilt this factor in favor of fair use. For use of raw data that is unpublished within the meaning of copyright law, this factor would likely disfavor fair use."); Katherine Lee et al. Initial Comments at 102.

[302] *See, e.g.*, Engine Initial Comments at 7 ("The nature of the content used will vary for each AI system. However, this factor rarely plays a significant role—standing alone—in determining fair use."). One commenter viewed the second factor as particularly unhelpful in evaluating generative AI because "[t]o the AI application, the exact type of content used for training is irrelevant." Van Lindberg Initial Comments at 26.

[303] The use of more expressive works is generally not incidental—when training a general-purpose model to perform well on diverse tasks, such as generating poetry or animated content, developers rely on correspondingly diverse training materials. *See supra* text accompanying notes 66–67.

[304] The nature of the copyrighted work may also be relevant to whether its use for training serves a transformative purpose. *See supra* note 265 (distinguishing the use of aesthetic stock photography and automated imagery for training generative image models).

[305] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994) (quoting 17 U.S.C. § 107). It is the amount and substantiality with respect to the copied work and not the infringing work that matters: "a taking may not be excused merely because it is insubstantial with respect to the infringing work." *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 565 (1985).

[306] *Campbell*, 510 U.S. at 586–87.

[307] *Id.* at 587; *Google Books*, 804 F.3d 202, 221 (2d Cir. 2015).

1. The Amount Used

The Supreme Court has said that courts assessing the amount and substantiality must consider both the quantity of material used and its quality and importance.[308] Copying even a small portion of a work may weigh against fair use where it is the "heart" of the work.[309] In general, "[t]he larger the amount, or the more important the part, of the original that is copied, the greater the likelihood that the secondary work might serve as an effectively competing substitute for the original, and might therefore diminish the original rights holder's sales and profits."[310]

Downloading works, curating them into a training dataset, and training on that dataset generally involve using all or substantially all of those works.[311] Such wholesale taking ordinarily weighs against fair use.[312]

2. Reasonableness in Light of Purpose

Copying an entire work may weigh less heavily against a finding of fair use, however, where it is reasonable in relation to a transformative purpose.[313] In several cases, courts have

---

[308] *Campbell*, 510 U.S. at 586–87.

[309] *See Harper & Row Publishers*, 471 U.S. at 565. Conversely, copying a large amount may not weigh against fair use where it "captures little of the material's creative expression." *Google LLC v. Oracle, Inc.*, 593 U.S. at 33 (citing *Campbell*, 510 U.S. at 588; *New Era Publications Int'l, ApS v. Carol Publishing Group*, 904 F.2d 152, 158 (2d Cir. 1990)).

[310] *Campbell*, 510 U.S. at 587.

[311] *See* NMPA Initial Comments at 10 ("When a pre-existing work is used to train an AI model, it is analyzed in its entirety. For some models, developers will compress each training example into a compact representation and then cause the developing model to predictively reconstruct it."); Karla Ortiz Initial Comments ("I found that almost the entire body of my work, the work of almost every artist I know, and the work of hundreds of thousands of other artists, was taken without our consent, credit or compensation to train these for-profit technologies."); Katherine Lee et al. Initial Comments at 100. Some have argued that generative AI training in fact uses little of the training data. *See* Meta Initial Comments at 15 (stating the process "meant to extract, relatively speaking, a miniscule amount of information from each piece of training data."); Oren Bracha, *The Work of Copyright in the Age of Machine Production*, UNIV. OF TEXAS LAW 1, 25 (last updated Feb. 16, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4581738 (arguing that what developers take in the training process is not protectible expression at all but a form of "meta information" about the works). As discussed above, however, the Office finds that the training process regularly uses expressive elements of the underlying works. *See supra* text accompanying notes 268–71.

[312] 4 NIMMER ON COPYRIGHT § 13F.07 (As a rule, "the more of a copyrighted work that is taken, the less likely the use is to be fair); *Capitol Records v. ReDigi*, 910 F.3d 649, 662 (2d Cir. 2018) ("[U]se of the entirety of a digital file . . . tends to disfavor a finding of fair use." (citations omitted)); *Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1118 (9th Cir. 2000) ("[C]opying an entire work militates against a finding of fair use. (citations omitted)).

[313] *See Google v. Oracle*, 593 U.S. at 35 ("The 'substantiality' factor will generally weigh in favor of fair use where . . . the amount of copying was tethered to a valid, transformative purpose."); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 613 (2nd Cir. 2006) (copying of the entire image copied did not weigh against a finding of fair use); *Kelly v. Arriba Soft*, 336 F.3d 811, 821 (9th Cir. 2003).

found mass copying of entire works to be justified when it enabled transformative uses, such as to develop search engines[314] or plagiarism detection software.[315] In *Google Books*, Google's scanning of millions of books was excused in part because "not only is the copying of the totality of the original [books] reasonably appropriate to Google's transformative purpose [of creating a search engine of books], it is literally necessary to achieve that purpose."[316] The Ninth Circuit similarly found that copying entire images was reasonable in relation to creating a visual search engine: "If Arriba only copied part of the image it would be more difficult to identify it, thereby reducing the usefulness of the visual search engine."[317]

Commenters disagreed about the need to use entire copyrighted works in AI training. Some believed that because the most powerful generative AI models need massive amounts of data, it is "reasonable for developers to try to maximize the amount of data these models ingest in order to increase the public benefit of these tools."[318] Others disputed either the amount of data needed or the justification for taking it.[319] NMPA argued that AI models' insensitivity to any particular copyrighted work make the third factor analysis different from search engine cases like *Google Books*: "Even if copying more portions of more works results in an AI model that is incrementally more commercially competitive, that is very different from the binary necessity for making complete copies in [Google Books]."[320] More fundamentally, several

---

[314] *See Kelly*, 336 F.3d at 821; *Google Books*, 804 F.3d 202, 221 (2d Cir. 2015); *Authors Guild v. HathiTrust*, 755 F.3d 87, 98 (2nd Cir. 2014).

[315] *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638–40 (4th Cir. 2009).

[316] *Google Books*, 804 F.3d at 221. *See also HathiTrust*, 755 F.3d at 98 (Libraries' digitization of full books to enable full-text search of their digital repository was not excessive "[b]ecause it was reasonably necessary . . . to make use of the entirety of the works in order to enable the full-text search function.").

[317] *Kelly*, 336 F.3d at 821. The copied images were displayed as "smaller, lower-resolution [thumbnail] images." *Id.*; *cf. VHT, Inc. v. Zillow Grp., Inc.*, 918 F.3d 723, 744 (9th Cir. 2019) (unlike other search engine cases involving transformative uses, third factor disfavored favor use because "nothing justifies [defendant's] full copy display of [plaintiff's] photos").

[318] Authors Alliance Initial Comments at 11. *See also* OpenAI Initial Comments at 12–13 ("In order to research, analyze and reflect the full breadth of human reasoning and understanding, AI models need to learn from as broad an array of examples as possible.").

[319] A2IM and RIAA, for example, claim that "[v]ery powerful tools can be built with less data." A2IM-RIAA Joint Initial Comments at 20. Getty Images states that for visual machine learning, "the quantity of visual work used for training may be less important than its quality." Getty Images Initial Comments at 16. *Cf.* Pls.' Reply in Supp. of Mot. for Prelim. Inj. at 6, Concord Music Grp., Inc. v. Anthropic PBC, No. 23-cv-1092 (M.D. Tenn. Feb. 14, 2024) ("First, Anthropic does not need to copy Publishers' artistic expression in its entirety to achieve its claimed purpose. Anthropic protests that Publishers' lyrics are a tiny fraction of its training data; it could easily exclude those lyrics and retain the remaining 'trillions of tokens of pre-existing text' it allegedly requires." (citation omitted)).

[320] NMPA Initial Comments at 18.

commenters argued that scale should not affect the fair use analysis.[321] In the words of one rightsholder association, "Fair use should not provide a 'volume discount.'"[322]

The Office agrees that the use of entire copyrighted works is less clearly justified in the context of AI training than it was for Google Books or a thumbnail image search. Those services made information available about the content of the works copied, making the extent of the copying definitionally necessary for full-text search to work. Generative AI, by contrast, is not limited to providing information about the works in the training dataset.[323] Moreover, there may be cases where a more targeted round of training has more limited data requirements; in such circumstances, the developer may be able to reduce the amount taken from individual works without compromising the training goal.[324]

Nevertheless, the use of entire works appears to be practically necessary for some forms of training for many generative AI models. While for large, general-purpose models, there is no need to copy any amount of any specific work,[325] research supports commenters' assertions that internet-scale pre-training data, including large amounts of entire works, may be necessary to achieve the performance of current-generation models.[326] To the extent there is a transformative purpose, the use of entire works on that scale could be reasonable.

3. The Amount Made Available to the Public

In several cases where the defendant made non-public, intermediate copies, courts have concluded that the question is "not so much 'the amount and substantiality of the portion used' in making a copy, but rather the amount and substantiality of what is thereby made accessible

---

[321] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 20 ("To argue, as many in the AI community do, that massive infringement is somehow permissible fair use while more limited infringement is not would turn established principles of copyright law on their head."); Copyright Alliance Initial Comments at 67; Getty Images Initial Comments at 16.

[322] AAP Initial Comments at 19.

[323] In addition, where the Second Circuit deemed Google Books' use of copyrighted works "highly transformative," *Google Books*, 804 F.3d 202, 222, as we noted above, uses of copyrighted works to train generative AI will vary in their degree of transformativeness. *See supra* Section IV.A.2.c.

[324] *See Dr. Seuss Enters. v. ComicMix*, 983 F.3d 443, 458 (9th Cir. 2020) (explaining that the third factor inquiry concerns the amount taken from the work(s) at issue, not the number of works used).

[325] AI companies themselves describe the impact of individual training examples as negligible. *See* Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J. at 8, Kadrey v. Meta Platforms Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489 ("Because any given work is a tiny fraction of total training data . . . No individual text materially contributes to the performance of the model.").

[326] *See* Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain* at 27, ARXIV (last updated Mar. 1, 2024), https://arxiv.org/abs/2309.08133 (contrasting pre-2017 machine learning datasets, which contained in the tens of thousands or tens of millions of images, with generative AI datasets, which include billions); *see also supra* Section II.C.1 (discussing quantity).

to a public for which it may serve as a competing substitute."[327] In *Sony v. Connectix* and *Sega v. Accolade*, the Ninth Circuit held that although defendants made, respectively, complete copies of a game console's basic input/output system and a video game in order to access their functional requirements, this carried "very little weight" when the ultimate material accessible to the public (a console emulator and an original video game) did not include the works' protectible expression.[328]

A few courts have extended this focus on outputs beyond the context of functional computer code.[329] In *Google Books*, described by the Second Circuit as "test[ing] the boundaries of fair use," although Google made complete copies of books, the third factor nevertheless did not weigh against Google because only carefully-limited "snippets" incapable of substituting for the original works were made available to the public.[330] And in a recent decision about copying legal summaries to train a (non-generative) AI search tool, the court found that factor three favored the defendant because its use did not make copyrighted material available to the public.[331] In contrast, where a defendant copied television broadcasts and allowed users to

---

[327] *Google Books*, 804 F.3d at 222.

[328] *See Sony Comput. Entm't v. Connectix*, 203 F.3d 596, 606 (9th Cir. 2000) ("But as we concluded in Sega, in a case of intermediate infringement when the final product does not itself contain infringing material, this factor is of 'very little weight.'"); *Sega v. Accolade*, 977 F.2d 1510, 1526–27 (9th Cir. 1992) ("Where the ultimate (as opposed to direct) use is as limited as it was here, the factor is of very little weight.").

[329] *See, e.g.*, *Google Books*, 804 F.3d at 221–22; *Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, No. 20-cv-613, 2025 WL 458520 at *9 (D. Del. Feb. 11, 2025) ("Because Ross did not make West headnotes available to the public, Ross benefits from factor three."). *But see Disney Enters., Inc. v. VidAngel, Inc.*, 869 F.3d 848, 862 n.12 (9th Cir. 2017) ("VidAngel also argues that creating an 'intermediate copy' for filtering is a 'classic fair use.' The cases it cites are inapposite, because VidAngel does not copy the Studios' works to access unprotected functional elements it cannot otherwise access.").

[330] *Google Books*, 804 F.3d at 221–22. In response to a user's search, the system would display a maximum of three snippets of a book containing the search term—the same three snippets regardless of how many times the user entered the search; each snippet was around an eighth of a page in length; and Google permanently blocked the system from displaying part of every page and all of one page out of ten. *Id.* at 210. Moreover, Google disabled snippet view entirely for certain categories, such as cookbooks and dictionaries, where a snippet might be all the searcher needs. Beginning in 2005, it did the same for any book upon request of the rightsholder. *Id.* "The result," according to the Second Circuit, was that "a searcher cannot succeed, even after long extended effort to multiply what can be revealed, in revealing through a snippet search what could usefully serve as a competing substitute for the original." *Id.* at 222–23. The court cautioned that "[i]f snippet view could be used to reveal a coherent block amounting to 16% of a book, that would raise a very different question."

[331] *Thomson Reuters*, 2025 WL 458520 at *9. The court nevertheless rejected the fair use defense, distinguishing the reverse engineering cases as being "about copying computer code," where copying expression was necessary to reach unprotectible ideas. *Id.* at *8.

view ten-minute clips, with no restrictions on the number they could view,[332] the Second Circuit found that the third factor clearly weighed against fair use.[333]

Professors Samuelson, Sag, and Sprigman described this line of cases as showing that "making complete literal copies [for generative AI training] . . . is reasonable as an intermediate technical step in an analytical process that does not lead to the communication of the underlying original expression to a new audience."[334] The Copyright Alliance disagreed, contending that the reverse engineering cases were specific to the use of functional code,[335] and that Google Books served a more clearly transformative purpose than generative AI training, in that it provided information about the works used rather than generating new outputs to compete with those works.[336] Moreover, Google Books had "significant safeguards" to reduce the risk that the copies could serve as substitutes.[337]

In the Office's view, while there are meaningful distinctions from the intermediate copying cases,[338] their logic suggests that the third factor may weigh less heavily against generative AI training where there are effective limits on the trained model's ability to output protected material from works in the training data. As in the intermediate copying cases, generative AI typically do not make all of what was copied available to the public. Most outputs from generative AI systems do not contain any protected expression from their training data, and models can be deployed in ways that entirely obscure outputs from users or result in non-expressive outputs.[339]

Where a model can output expression, however, the question is whether, like Google Books, the AI developer has adopted adequate safeguards to limit the exposure of copyrighted material. At least for some "memorized" works, generative AI users can potentially obtain far

---

[332] *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 175 (2nd Cir. 2018).

[333] *Id.* at 178.

[334] Pamela Samuelson et al. Initial Comments at 16.

[335] Copyright Alliance Initial Comments at 55.

[336] *Id.* at 56–57.

[337] *Id.*

[338] *Sony* and *Sega* concerned intermediate copying that was necessary to access functional material. *See Sony Comput. Entm't v. Connectix*, 203 F.3d 596, 606 (9th Cir. 2000); *Sega v. Accolade*, 977 F.2d 1510, 1526–27 (9th Cir. 1992).

[339] For example, as described above, Anthropic advertises its model's use for classifying customer support tickets. *See supra* note 124.

more protectible expression than the snippets made available in *Google Books*.[340] Commenters disagree about how much effort this requires.[341] They do not dispute that it happens.[342]

But many generative AI companies with chatbot and other public-facing services employ guardrails and other methods to prevent potentially infringing outputs.[343] These include input filters that block user prompts likely to result in generations that reproduce copyrighted content; training techniques designed to make infringing outputs less likely; internal system prompts that instruct it not to generate names of copyrighted characters or create images in the style of living artists; and output filters that block copyrighted content from being displayed.[344] Although there are factual disputes over the efficacy of these guardrails,[345] where they do prevent the generation of infringing content, the third factor will weigh less heavily against fair use.

In sum, AI developers ordinarily copy entire works and make use of their expressive content for training, weighing against fair use. But in cases where there is a transformative purpose, and where there is a need to train on a large volume of works to effectively generalize, the copying of entire works may be reasonable. This is especially true where little or none of the copied material will be made accessible to the public, whether due to training techniques or choices made in deployment.[346] In those circumstances, the third factor may not weigh against fair use.

---

[340] *See supra* note 330.

[341] *See supra* Section II.D.2; UMG Initial Comments at 6; Johan Brandstedt Initial Comments at 30; Rich Campanella Reply Comments at 11; John-Edgar Martin Lopez Reply Comments at 11.

[342] *See* OpenAI Initial Comments at 7; Meta Initial Comments at 16 n. 68.

[343] *See supra* Section II.E. *See also* Winston Cho, *Music Publishers Reach Deal With AI Giant Anthropic Over Copyrighted Song* Lyrics, HOLLYWOOD REP. (Jan. 2, 2025) ("Under the agreement, Anthropic will apply already-implemented guardrails in the training of new AI systems. The deal also provides an avenue for music publishers to intervene if the guardrails aren't working as intended."); Matthew Finnegan, *Microsoft Pledges to defend Copilot customers against copyright lawsuits*, COMPUTERWORLD (Sept. 8, 2023) ("Microsoft said it already has content filters in place to reduce the likelihood of Copilot generating copyright-infringing material in its responses.").

[344] *See supra* Section II.E.

[345] *Compare* Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J. at 23–24, Kadrey v. Meta Platforms Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489 (arguing that Meta's Llama model cannot be used to produce more than 1% of any work), *with* Pls.' Opp. to Def.'s Mtn. to Dismiss at 3–4, 9, 15, Concord Music Grp., Inc. v. Anthropic PBC, No. 24-cv-3811 (N.D. Cal. Sept. 5, 2024), ECF No. 222 (arguing that Anthropic's models generated outputs in response to user prompts that include "identical or nearly identical copies" of works and that Anthropic's guardrails are ineffective, inconsistent, and easily evaded).

[346] *See Google Books*, 804 F.3d 202, 222–23 (2d Cir. 2015).

D. Factor Four

The fourth and final statutory factor is "the effect of the use upon the potential market for or value of the copyrighted work."[347] "The enquiry must take account not only of harm to the original but also of harm to the market for derivative works."[348] The Supreme Court has twice described this factor as "undoubtedly the single most important element of fair use,"[349] although its importance "will vary, not only with the amount of harm, but also with the relative strength of the showing on the other factors."[350] Although the copyright owners might "bear some initial burden of *identifying* relevant markets," they "need not present empirical data of their own in connection with [the] asserted affirmative defense."[351]

This section evaluates different ways in which the use of copyrighted works for generative AI can affect the market for or value of those works, including through lost sales, market dilution, and lost licensing opportunities.[352] It also addresses broader claims that the public benefits of unlicensed training might shift the fair use balance.

---

[347] 17 U.S.C. § 107(4).

[348] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994) (internal quotation marks and citation omitted). The harm must, however, arise from an interest protected by copyright. *See Google Books*, 804 F.3d at 224 (acknowledging that the "snippet" feature might "cause some loss of sales" if, for instance, a consumer needed only to locate a specific fact that the snippet displayed, but noting that such a loss would be related to an interest not protected by copyright, which does not extend to the facts described in the book).

[349] *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 566 (1985). *See also Stewart v. Abend*, 495 U.S. 207, 238 (1990). Some courts have attributed this factor's importance to the relationship between market effect and copyright's underlying goal of rewarding authors for their creations. *See, e.g.*, *Google Books*, 804 F.3d at 214; *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 642 (4th Cir. 2009).

[350] *Campbell*, 510 U.S. at 590 n.21.

[351] *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 194.

[352] Because section 107(4) directs courts to consider "the effect of the use upon the potential market for *or value of* the copyrighted work," 17 U.S.C. § 107(4) (emphasis added), courts have sometimes looked beyond "strictly monetary terms" and considered other negative effects on the value of the copyrighted work. *See Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory*, 689 F.3d 29, 64 (1st Cir. 2012); *Chicago Bd. of Educ. v. Substance, Inc.*, 354 F.3d 624, 630 (7th Cir. 2003); *cf.* Jane C. Ginsburg, Essay, *Fair Use Factor Four Revisited: Valuing the "Value of the Copyrighted Work,"* 67 J. COPYRIGHT SOC'Y USA 19, 21 (2020). In one case, the Ninth Circuit held that although plaintiff's religious text might not have traditional monetary value, defendant's use could injure its evangelizing value. *Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1119 (9th Cir. 2000). Non-monetary value has also been found to be harmed by a loss of advertising potential and by damage to reputation. *See Video Pipeline, Inc. v. Buena Vista Home Entm't, Inc.*, 342 F.3d 191, 203 (3d Cir. 2003), *abrogated on other grounds by TD Bank N.A. v. Hill*, 928 F.3d 259 (3d Cir. 2019); *Penguin Grp. (USA) Inc. v. Am. Buddha*, No. 13-cv-2075, 2015 WL 11170727 at *6 (D. Ariz. May 11, 2015).

1. Lost Sales

The first harm to consider is "actual or potential market substitution"[353]—that is, whether a market for the original work is supplanted "so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."[354] Courts consider not only the harm from a particular use but also whether there would be a "substantially adverse impact" on the market if that use were to become "unrestricted" and "widespread."[355]

Commenters offered competing perspectives on whether or how the outputs of generative AI can substitute for the originals. Several asserted that use of copyrighted works for training was clearly substitutional insofar as the model generates copies of the work.[356] The National Association of Broadcasters provided an example of "nearly word for word" copies of a local station's news stories being reproduced by a generative AI system without permission from the station or its owner,[357] "illustrat[ing] how AI-generated 'news' has the potential to substitute for and supplant the market for copyrighted broadcast content on which the AI systems have been trained."[358]

Other commenters argued that the substitution that may occur is broader than the harm cognizable under the fourth factor.[359] As Meta put it, "while it is possible (at least in theory) for Generative AI to create works 'of the same type' that compete in the overall market with the originals, this is not the kind of substitution that implicates the fourth fair use factor."[360] The

---

[353] *Warhol*, 598 U.S. 508, 536 n. 12 (2023).

[354] *Google Books*, 804 F.3d 202 at 223.

[355] *See Campbell*, 510 U.S. at 590; *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1276 (11th Cir. 2014).

[356] *See* A2IM-Recording Academy-RIAA Joint Reply Comments at 13 ("When copyrighted works are used in the development of a generative AI model that outputs the same type of works, the potential for those outputs to supplant the market for the input works is clear."); Center for Art Law Initial Comments at 5; N/MA Initial Comments at 46–47; Yelp Reply Comments at 10.

[357] NAB Initial Comments at 4.

[358] *Id.* at 5. *See also* IAC-DDM Joint Initial Comments at 2–3 (noting, in the digital publishing context, that in search engines' uses of generative AI, "[t]he long-familiar links to trusted, authoritative websites where users may access that original content are replaced by unsourced, unattributed, synthetic 'answers' based on that content").

[359] *See, e.g.*, Anthropic Initial Comments at 8 ("Courts have held that generating new works in the same 'class of works' can still be fair use under the fourth factor. The key question is whether the use substitutes for the original in the market, not simply whether the use creates a more competitive marketplace."); Hugging Face Initial Comments at 10 ("The fourth factor analysis centers on the communicable expression of a given work (or works). It has not to our knowledge previously been interpreted as preventing competition generally among users and developers of new tools.").

[360] Meta Initial Comments at 18.

Authors Alliance likewise contended that the effect on the market "is unlikely to be significant based on the lack of a substitutional effect between the individual works themselves and the generative AI systems based on AI models that use them as training materials."[361]

There are instances, however, where the use of works in generative AI training can lead to a loss in sales. The use of pirated collections of copyrighted works to build a training library, or the distribution of such a library to the public, would harm the market for access to those works. And where training enables a model to output verbatim or substantially similar copies of the works trained on, and those copies are readily accessible by end users, they can substitute for sales of those works.[362]

A potential loss of sales is particularly clear in the case of works specifically developed for AI training. There is a thriving industry focused on developing training datasets that improve the ability of language models to follow instructions, format and structure outputs, use tools, act consistently with human values, or improve domain performance.[363] Where the content of those datasets is copyrightable, or the datasets themselves evince human selection and arrangement of data, and the datasets are primarily or solely targeted at AI training, widespread unlicensed use would likely cause market harm.[364]

Uses involving the retrieval of copyrighted works by RAG can also result in market substitution. As described above, RAG augments AI model responses by retrieving relevant content during the generation process, resulting in outputs that may be more likely to contain protectable expression, including derivative summaries and abridgments.[365] A user for whom

---

[361] Authors Alliance Initial Comments at 11.

[362] *Cf. Google Books*, 804 F.3d 202, 224 (2d Cir. 2015) ("Especially in view of the fact that the normal purchase price of a book is relatively low in relation to the cost of manpower needed to secure an arbitrary assortment of randomly scattered snippets, we conclude that the snippet function does not give searchers access to effectively competing substitutes.").

[363] *See, e.g.*, *Nvidia/Llama-Nemotron-Post-Training-Dataset-v1*, HUGGING FACE, https://huggingface.co/datasets/nvidia/Llama-Nemotron-Post-Training-Dataset-v1 (dataset described as "support[ing] improvements of math, code, general reasoning, and instruction following capabilities").

[364] *See, e.g.*, Data Provenance Initiative Initial Comments at 2–3 ("[D]atasets containing data created for the sole purpose of training machine learning models (mainly for finetuning and alignment) . . . may likely contain copyrightable contributions from the dataset creators in the form of annotations. . . . [T]he unauthorized use of [such] datasets for training machine learning is identical to its original purpose."); Regulosity-Pangea Joint Initial Comments at 8; EWC Initial Comments at 8.

[365] *See supra* Sections II.E, III.C, text accompanying notes 265–266; *see also* New York Times Initial Comments at 3 ("[S]ome GAI products go so far as to retrieve and copy our most recent and relevant content in order to 'ground' generative AI output, through a process known as 'retrieval augmented search.' . . . GAI products are designed to keep readers on the companies' own tools and websites by providing expressive, satisfying summaries in response to queries that obviate the need for users to travel to publishers' platforms.").

the augmented response "satisf[ies] the . . . need"[366] for the original work will not pay to obtain it in the marketplace.

2. Market Dilution

A number of commenters contended that courts should consider the harms caused where a generative AI model's outputs, even if not substantially similar to a specific copyrighted work, compete in the market for that type of work.[367] Pointing to copyright's underlying goals of incentivizing creation, the Copyright Alliance argued that "with generative AI, the harm is often to a creator's overall body of work or even the market more broadly. These harms all impact the creator's incentives, and they should be considered under a factor-four analysis."[368] Professor David Newhoff stated, "[G]enerative AI—if it does not produce market substitutes—primarily represents potential harm to authors and future authorship. . . . [T]he consideration in the context of 'training' should be expansive and doctrinal—namely that a potential threat to 'authorship' cannot, by definition, 'promote the progress' of 'authorship.'"[369] And the Association of American Publishers asserted that "[i]f a copyrighted work is reproduced to train a Gen AI model that will generate works that compete in the market with the copyrighted work, it will clearly reduce the value of that copyrighted work."[370]

Other commenters argued that the fourth factor analysis considers only harm to markets for the specific copyrighted work.[371] In the words of one, "if the [fourth factor] inquiry were to

---

[366] *Google Books*, 804 F.3d at 223.

[367] *See, e.g.*, ASCAP Initial Comments ("[A] single model like GPT has unprecedented potential to replace a wide range of copyrighted content by numerous creators spanning various industries. . . . [A]ny analysis for replacement effect under the fourth factor should not be limited to a single piece of copyrighted work or an artist, but also the market for that type of copyrighted work in general."); Music Workers Alliance Initial Comments at 4 ("[O]utputs are generally a blend of a variety of works, no one of which may be immediately recognizable as deriving from a particular work. In this environment, the only test that makes any sense is to explore the impact on a broad category of works."); UMG Initial Comments at 48.

[368] Copyright Alliance Initial Comments at 68. *See also* Kate Barsotti Initial Comments ("I will have no financial incentive to take classes, hone my craft, buy supplies, or spend years improving my work if it can be stolen and distributed without recourse.").

[369] David Newhoff Initial Comments at 1; *see also* Johan Brandstedt Initial Comments at 24 ("Only in rare cases does [generative AI] impact the market value of individual works. But . . . it cannibalizes individual artists" by targeting their portfolios.).

[370] AAP Initial Comments at 20.

[371] *See* Meta Initial Comments at 18 ("[W]hile it is possible (at least in theory) for Generative AI to create works 'of the same type' that compete in the overall market with the originals, this is not the kind of substitution that implicates the fourth fair use factor, which does not punish uses that 'simply enable[] the copier to enter the market for works of the same type as the copied work.'" (quoting *Sega Enters. Ltd. v. Accolade*, 977 F.2d at 1523)); Authors Alliance Initial Comments at 13 ("We can think of no fair use case that has ever assessed market harm by adopting such a broad approach to market harm."); Google Initial Comments at 10–11.

extend to whether the AI system competes in the market for a general class of works, it could have unintended and potentially detrimental consequences. This broader scope would potentially stifle innovation and creativity in AI development, as it could effectively ban the use of the technology altogether."[372]

While we acknowledge this is uncharted territory, in the Office's view, the fourth factor should not be read so narrowly. The statute on its face encompasses any "effect" upon the potential market.[373] The speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data.[374] That means more competition for sales of an author's works and more difficulty for audiences in finding them. If thousands of AI-generated romance novels are put on the market, fewer of the human-authored romance novels that the AI was trained on are likely to be sold. Royalty pools can also be diluted. UMG noted that "[a]s AI-generated music becomes increasingly easy to create, it saturates this already dense marketplace, competing unfairly with genuine human artistry, distorting digital platform algorithms and driving 'cheap content oversupply' - generic content diluting human creators' royalties."[375]

Market harm can also stem from AI models' generation of material stylistically similar to works in their training data. As the Office noted in Part 1 of this Report, many commenters raised concerns about AI outputs that imitate a creator's style, which copyright does not protect

---

[372] Scenario, Inc. Initial Comments at 14.

[373] 17 U.S.C. § 107. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. at 590 ("[The fourth factor] requires courts to consider not only the extent of market harm caused by the particular actions of the alleged infringer, but also whether unrestricted and widespread conduct of the sort engaged in by the defendant would result in a substantially adverse impact on the potential market for the original.") (cleaned up).

[374] *See* Science Fiction and Fantasy Writers Association Initial Comments at 6 ("The harm creators and audiences are already experiencing is a flood of trash, directly enabled by generative AI with no restrictions on output. . . . AI-generated material . . . literally crowds human writers out."); Boni Alimagno Reply Comments at 7 ("Focusing on specific copyrighted works neglect stylistic elements that recur throughout a body of work, which through publicity become how an artist creates market value for their style. AI art generators instead allow an artist's own body of work to face competition from a too similar body of work—driving the monetary value of their uniqueness downward. An artist is in competition with themselves.").

[375] UMG Initial Comments at 12–13. In a real-world example of generative AI's potential to distort royalties, a recently indicted individual earned more than $10 million in royalty payments from fraudulently streaming thousands of AI-generated songs across several music platforms. This diminished royalties for the other works streamed by those platforms—a clear economic impact on individual authors. *See* Press Release, U.S. Attorney's Office, Southern District of New York, North Carolina Musician Charged with Music Streaming Fraud Aided by Artificial Intelligence (Sept. 4, 2024), https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence. *See also* Compl. at 5, UMG Recordings Inc. v. Suno, Inc., No. 24-cv-11611 (D. Mass. June 24, 2024) (alleging that some outputs from defendant's music generative AI company "amass[ed] upwards of 2,000,000 streams," with some "finding their way onto the major streaming services . . . compet[ing] with the copyrighted sound recordings that enabled their creation").

as a separate element.[376] Even when the output is not substantially similar to a specific underlying work, stylistic imitation made possible by its use in training may impact the creator's market. In the words of the Writers Guild of America, because AI systems can be prompted to imitate a writer's style, applying fair use would force writers "to compete with AI-generated scripts trained on their work, without their authorization, and without fair compensation."[377] This threat is more acute because of the technology's ability to produce works so similar in style "that the average person cannot discern a difference in the marketplace[,] . . . creat[ing] direct competition with the creators whose works have been used to train the model."[378]

3. Lost Licensing Opportunities

Lost revenue in actual or potential licensing markets can also be an element of market harm. Because, in theory, copyright owners could accept payment for any uses of their works,[379] the relevant markets are those that are "traditional, reasonable, or likely to be

[376] *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 1: DIGITAL REPLICAS 53–56 (2024). There may, however, be cases where the replication of "style" does capture protectible elements of an original work of authorship. *See id.* at 55. *See generally* Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38 HARV. J.L. & TECH. 49 (2024).

[377] WGA Initial Comments at 2. *See also* Center for Art Law Initial Comments at 5 ("Generative AI tools, particularly AI image generators, are often prompted to output works that are "in the style of" specific artists, producing works designed to directly compete with that artist's work. . . . The advent of generative AI may result in a renewed interest in market competition and market impact in future fair use cases involving the technology and its outputs, potentially shifting the focus on whether AI-generated content replaces and competes with the original work.").

[378] CISAC Reply Comments at 3. In one highly publicized example, an AI image generator now allows users to generate images in the style of a popular Japanese animation studio, resulting in "a tsunami of images. Eve Upton-Clark, *OpenAI's Studio Ghibli-style Images Renew the Debate Over AI and Copyright*, FAST COMPANY (Mar. 28, 2025), https://www.fastcompany.com/91308222/openais-studio-ghibli-style-images-renew-the-debate-over-ai-and-copyright; *see also* Tor Constantino, *The Studio Ghibli Dilemma – Copyright in the Age of Generative AI*, FORBES (May 6, 2025), https://www.forbes.com/sites/torconstantino/2025/05/06/the-studio-ghibli-dilemma--copyright-in-the-age-of-generative-ai/. The result could undermine licensing opportunities for the studio. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. at 593–97 (noting that "the licensing of derivatives is an important economic incentive to the creation of originals" when remanding for development of the record as to the market for rap derivatives).

[379] *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 929 n.17 ("[A] copyright holder can always assert some degree of adverse affect on its potential licensing revenues as a consequence of the secondary use at issue simply because the copyright holder has not been paid a fee to permit that particular use. . . . Thus, were a court automatically to conclude in every case that potential licensing revenues were impermissibly impaired simply because the secondary user did not pay a fee for the right to engage in the use, the fourth fair use factor would always favor the copyright holder." (citations omitted)); *see also* Katherine Lee et al. Initial Comments at 100 ("Whether there is a licensing market for generative-AI models . . . is circular because the existence of a licensing market counts in favor of the copyright owner under the fourth factor but if this copying is a fair use, then no such market can develop." (citations omitted)).

developed."[380] A licensing market need not be long-standing or exhaustive, however, to be cognizable.[381]

Licensing is core to the business model of many content industries, and several industry representatives professed their willingness and ability to license works for AI training.[382] Many commenters stated that individual and collective licenses for AI use were already in existence or under development.[383] As of the end of 2023, they reported that AI developers were licensing copyrighted works in a number of sectors, including music,[384] vocal recordings,[385] and news reports.[386] Commenters highlighted public licensing deals between OpenAI and the Associated Press (news) and Shutterstock (images), Getty Image's collaborations with Nvidia and Bria, and

[380] *Texaco*, 60 F.3d at 930; *Princeton Univ. Press v. Michigan Document Servs., Inc.*, 99 F.3d 1381, 1387 (6th Cir. 1996). *See also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 592 (1994) ("that creators of original works would in general develop or license others to develop").

[381] *See Texaco*, 60 F.3d at 929 n.16, 930 (describing participation in licensing through the Copyright Clearance Center as limited, but concluding that it is a workable market that should be considered).

[382] *See, e.g.*, A2IM-Recording Academy-RIAA Joint Reply Comments at 14 ("Today, content licensing is at the core of record companies' businesses. From a record company perspective, generative AI simply represents a new potential market for licensing uses of their sound recordings."); N/MA Initial Comments at 58.

[383] *See, e.g.*, CCC Initial Comments at 12 ("CCC already offers market-based, global non-exclusive voluntary licenses to support AI in the commercial research, schools, and education technology sectors."); Letter from CCC, Summary of *Ex Parte* Meeting on Apr. 29, 2024 Regarding the Office's AI Study, to U.S. Copyright Office 2 (Apr. 22, 2024) ("CCC anticipates providing additional information on the development of licensing solutions within the coming months. CCC is also looking at additional transactional and external licensing options."); Getty Images Initial Comments at 19–20 ("[T]here is an established path for licensing visual data for use in training, and there are already fully licensed, non-infringing products in the marketplace."); NMPA Initial Comments at 23 ("AI system developers seeking to train on musical works can license directly from copyright owners or their agents, and this licensing process is already underway."); Recording Academy Initial Comments at 7 ("The company or individual behind the AI model should contact the copyright owner (or their designee) and obtain a license. Many AI models are already obtaining licenses this way and it has been the norm across many other examples within the music distribution ecosystem."); Shutterstock Initial Comments at 3 ("Shutterstock has built robust demand for ethically sourced AI training data. It has partnered with multiple companies that are interested in training their AI models on licensed data from Shutterstock, including LG and Meta."); Software & Information Industry Ass'n Initial Comments at 3 ("Many of our members already license their works for use as AI training data.").

[384] *See, e.g.*, NMPA Initial Comments at 19 ("The market is not merely a 'potential' or theoretical market the existence or feasibility of which is open to debate; it is an actual market, with great potential for growth. Music companies are currently licensing works for use in training AI models.").

[385] *See, e.g.*, FTC Initial Comments, Attach. at 18 (Tim Friedlander, President, National Association of Voice Actors) ("I personally am working on a synthetic voice that I have consent, compensation, and control for.")

[386] *See, e.g.*, AP Initial Comments at 3 ("[N]ews publishers have already developed a licensing market for machine learning. For example, in July 2023, the AP announced a licensing deal allowing OpenAI to train its AI models on portions of the AP's text archive of news stories.").

the collaboration between vAIsual and music/audio broker Rightsify.[387] They suggested that further licensing was expected, particularly in sectors well-positioned to accommodate expanded voluntary licensing, like music[388] and academic publishing.[389]

Since the comments were submitted, considerable activity has taken place. Recent public reporting reflects AI licensing for images[390] and audio-visual works,[391] academic and

[387] *See* Copyright Alliance Initial Comments at 29; Copyright Licensing Agency Initial Comments at 10; MPA Initial Comments at 29–30.

[388] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 23, 25 ("[T]he recorded music industry has all the necessary systems and infrastructure already in place to make obtaining advance consent demonstrably feasible."); ASCAP Initial Comments at 40 ("[B]ased on our experience, direct voluntary licensing is well suited for generative AI, and has worked successfully with respect to public performance rights of musical works in the U.S. over the past century and through many technological developments."); Nashville Songwriters Association International Initial Comments at 6 ("The music industry has been successfully licensing musical works for synchronization uses in the free market for decades and that same model can be applied for AI training uses"); UMG Initial Comments at 68 ("[T]he music industry is adept at licensing its content in huge quantities for countless different uses, so UMG does not anticipate legal, technical, or practical barriers to granting licenses for use of its content for training purposes.").

[389] *See* AAP Initial Comments at 24 ("Yes, direct voluntary licensing is feasible and is certainly the case for the publishing industry. Professional and scholarly publishers already employ licensing arrangements to facilitate access to their databases, whether for non-commercial research purposes or for commercial use."); Scientific Technical Medical Publishers Initial Comments at 13 ("Speaking for the academic/scientific/medical publishing sector, direct voluntary licensing is not only feasible but already pervasive in our sector for a variety of types of uses.").

[390] *See, e.g.*, Brody Ford, *Shutterstock's AI-Licensing Business Generated $104 Million Last Year*, YAHOO! FINANCE (June 4, 2024), https://finance.yahoo.com/news/shutterstock-ai-licensing-business-generated-120000890.html ("Demand for this data has opened up a new opportunity for Shutterstock, whose traditional business of licensing media to advertising firms and creative artists has slowed down in recent years. Many of the companies that licensed data from Shutterstock. . . . wanted to have images that were legally obtained and contained good-quality descriptions, which assists in the training process, [the Shutterstock CEO] said.").

[391] Etan Vlessing, *Lionsgate CEO Says AI Deal Promises "Transformational Impact" on Studio*, HOLLYWOOD REP. (Nov. 7, 2024), https://www.hollywoodreporter.com/business/business-news/lionsgate-ai-deal-runway-1236055999/ (indicating that the AI model could only be used by Lionsgate and its designees); Dashveenjit Kaur, *Content creators strike gold in AI content licensing boom*, TECHHQ (Jan. 15, 2025), https://techhq.com/2025/01/content-creators-strike-gold-in-ai-content-licensing-boom/ ("The AI content licensing landscape is shifting as significant technology companies compete to acquire exclusive video content from creators, offering substantial payouts for previously unused footage.").

nonfiction publishing,[392] and news publishing,[393] as well as various content aggregators offering or facilitating collective licensing of training materials.[394]

A number of commenters disputed that current licensing activity demonstrates the feasibility of broad implementation of voluntary licensing.[395] They argued that licensing cannot provide the quantity, diversity, or type of data that many AI systems require; that licensing such data would be prohibitively expensive and available only to certain developers and for certain copyrighted works; and that the practical challenges of identifying and contacting copyright owners would make full licensing impossible.[396]

---

[392] *See* Matilda Battersbu, *Wiley set to earn $44m from AI rights deals, confirms 'no opt-out' for authors*, THE BOOKSELLER (Aug. 30, 2024), https://www.thebookseller.com/news/wiley-set-to-earn-44m-from-ai-rights-deals-confirms-no-opt-out-for-authors (academic publishers Wiley and Taylor & Francis have licensed academic works to AI companies for use in training LLMs); Alice Robb, *How Much Should Authors Get Paid to License Books to AI?: Essay*, BLOOMBERG (Feb. 7, 2025), https://www.bloomberg.com/news/articles/2025-02-07/how-much-should-authors-get-paid-to-license-books-for-ai-training.

[393] *See* Bill Rosenblatt, *The Media Industry's Race to License Content for AI*, FORBES (July 18, 2024), https://www.forbes.com/sites/billrosenblatt/2024/07/18/the-media-industrys-race-to-license-content-for-ai/ ("The most active area for individual deals right now by far—judging from publicly known deals—is news and journalism."). *See id.* (listing publicly reported deals with news publishers); Todd Spangler, *Condé Nast Inks Pact With OpenAI, Latest Media Company to License Content to Generative AI Platform*, VARIETY (Aug. 20, 2024), https://variety.com/2024/digital/news/conde-nast-openai-licensing-deal-1236112556/ (reporting that "articles from [Condé Nast's] titles," including the New Yorker, Vanity Fair, and Wired, "would be incorporated into OpenAI products, which would credit the original publication as the source material").

[394] *See, e.g.*, Ed Nawotka, *CCC Launches Collective Licensing for AI*, PUBLISHERS WEEKLY (July 16, 2024), https://www.publishersweekly.com/pw/by-topic/digital/copyright/article/95512-ccc-launches-collective-licensing-for-ai.html (announcing collective licensing solution aimed at internal use). *See also* CREATED BY HUMANS, https://www.createdbyhumans.ai/ (describing their service as an "AI licensing platform for creators," noting that "[w]e negotiate the details of the license, and you track payments"); Audrey Schomer, *Training AI with TV & Film Content: How Licensing Deals Look*, VARIETY (Aug. 6, 2024), https://variety.com/vip/training-ai-tv-film-content-how-licensing-deals-look-1236096126/ (identifying Calliope Networks as a significant aggregator of high-quality and diverse film and television works "engaging in deal talks about licensing its catalog with several AI companies building video generation models"); Press Release, Dataset Providers Alliance, *Announcing the Launch of the Dataset Providers Alliance (DPA)* (June 26, 2024), https://www.thedpa.ai/post/leading-dataset-licensors-unite-to-launch-the-dataset-providers-alliance-dpa.

[395] *See, e.g.*, Pamela Samuelson et al. Initial Comments at 27 ("Media reports indicate several examples of companies like Reuters and Shutterstock entering into licensing deals with AI developers, but the feasibility of such direct licensing depends on the nature of the works and the concentration of rights in the relevant market. In many instances, transaction costs are likely to be high."); Hugging Face Initial Comments at 11 ("[I]t is not currently feasible to seek opt-ins for already published data—especially as the majority of data under copyright on the web does not have an easily identifiable rights holder.").

[396] *See, e.g.*, Anthropic Initial Comments at 9 ("[A] regime that always requires licensing . . . would, at a minimum, effectively lock up access to the vast majority of works, since most works are not actively managed and licensed in any way."); R Street Initial Comments at 5 ("[T]he costs associated with obtaining these licenses could make AI projects excessively expensive, thus impeding innovation and hindering industry growth. This approach may render

Although licensing markets are still developing and factual contexts vary, available information shows that markets exist or are "reasonable" or "likely to be developed,"[397] for certain copyright sectors, types of training or uses, and models. Direct licensing is most common and most promising with respect to corporate entities with catalogs of high-quality and easily identifiable content.[398] For example, content controlled by large stock photography companies, national news outlets, and major record companies or film studios may be more easily licensable. Such content likely has a higher training value because it is high-quality and curated, and the centralization of rights makes it easier to license without incurring substantial volume-related transaction costs.[399]

Yet, it is also unclear that markets are emerging or will emerge for all kinds of works at the scale required for all kinds of models. There are copyright sectors where licensing infrastructure does not yet exist and may be difficult to build, and the amount of training data needed to produce state-of-the-art models may vary by content type or type of training.[400] Administrative or transactional costs can pose particular challenges when works are created outside of professional creative industries or are not intended to be monetized,[401] or when

many AI-driven projects unattainable, particularly for smaller entities or researchers with limited resources."); CCIA Initial Comments at 15 ("Much of the material on which generative AIs are trained may lack any identified or identifiable author from whom to obtain a license. Even where an author might be identified, contacting them might be difficult or impossible."); EFF Initial Comments at 4 ("It would not be feasible to seek authorization from every copyright owner, particularly since the elimination of formalities means that copyright attaches at fixation to all sorts of amateur creations not part of any market."); BigBear AI Initial Comments at 22; Lee Hollaar Initial Comments at 4; Microsoft-Github Joint Initial Comments at 9; OpenAI Initial Comments at 13.

[397] *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d at 930.

[398] *See infra* Section V.A.1.

[399] *See id.*

[400] *See id*.

[401] For instance, "vernacular works"—content created and posted online by members of the public without the expectation of monetization—may be particularly difficult to license. These may include social media posts, individual blogs or user comments or reviews, or personal photographs or videos. Meta Initial Comments at 17 ("[I]t would be impossible for AI developers to license the rights to other critical categories of works—like internet reviews and other examples of casual, vernacular text—both because it would be impossible to locate the owners of such works, and administratively impossible to negotiate licenses with each of them.").

ownership is diffuse.[402] Transaction costs in some cases might exceed the value of the works for training and render direct licensing infeasible.[403]

As both the creative industries and AI technologies develop further, data needs and licensing markets will continue to evolve.[404] Where licensing markets are available to meet AI training needs, unlicensed uses will be disfavored under the fourth factor. But if barriers to licensing prove insurmountable for parties' uses of some types of works, there will be no functioning market to harm and the fourth factor may favor fair use.

4. Public Benefits

As part of the fourth factor,[405] some courts have evaluated the public benefits that the defendant's use is likely to produce, considering how these benefits relate to the goals of copyright and their relative importance.[406]

---

[402] Some of those administrative concerns could be allayed if the platforms hosting such content licensed it collectively, which Ben Sobel has suggested would build on existing markets for bulk user data. *See* Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 75–78 (2017). Of course, platforms engaging in such activity would need to obtain appropriate authorization from users.

[403] *See White v. W. Pub. Corp.*, 29 F. Supp. 3d 396, 400 (S.D.N.Y. 2014) (finding "that no potential market exists because the transactions costs in licensing attorney works would be prohibitively high"). As a result, there may be no traditional, reasonable, or likely to be developed market for such works. *See Monsarrat v. Newman*, 28 F.4th 314, 324 (1st Cir. 2022) (affirming conclusion that there was no market for plaintiff's copyrighted post to an online comment thread); *Swatch Group Mgmt. Servs. v. Bloomberg L.P.*, 756 F.3d at 91 (concluding that "hypothesized market for audio recordings of earnings calls convened by foreign companies that are exempt from Regulation FD cannot meet [the *Texaco*] standard").

[404] As one commenter noted, while "[i]t is true that, in some modalities (e.g. text), you still need a very large amount of data to train the best models . . . it is by no means certain that this will always be the case." Ed Newton-Rex Initial Comments at 2–3. *See* Meta Reply Comments at 5–7 ("[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output."). In addition to voluntary licensing, different industries and types of works may also be differently suited to alternate licensing models, like compulsory licensing or ECL. *See infra* Section V.B.

[405] Public benefits are also accounted for in the analysis of the four statutory factors as a whole. For example, transformative uses are often described as adding value for the public. *See* Pierre N. Leval, Commentaries, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990); *see also Perfect 10 Inc. v. Amazon.com, Inc.*, 508 F.3d at 1166 (examining the extent to which an image search engine "promotes the purposes of copyright and serves the interests of the public," finding "the significantly transformative nature of [the] search engine, particularly in light of its public benefit, outweighs [its] commercial uses").

[406] Public benefits should be "related to copyright's concern for the creative production of new expression." *See Google LLC v. Oracle Am., Inc.*, 593 U.S. at 35–36 ("Are those benefits, for example, related to copyright's concern for the creative production of new expression? Are they comparatively important, or unimportant, when compared with dollar amounts likely lost (taking into account as well the nature of the source of the loss)?"). *See also Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th at 195; *Perfect 10 Inc. v. Amazon.com, Inc.*, 508 F.3d at 1166; *see also* Amanda

A number of commenters identified public benefits from unlicensed generative AI training. OpenAI, for example, stated that generative AI promises to "augment human capabilities, thereby fostering human creativity."[407] Meta has asserted in litigation that its open-source models enable "platforms built on Llama, to bring innovative and, in some cases, potentially life-saving services and technologies to market."[408] Several commenters maintained that limiting training content would negatively affect model performance, leading to bias and inaccuracy.[409]

On the other hand, others asserted that unlicensed use of copyrighted works to train AI injure the public by impeding the growth of the creative economy and authors' ability to earn livelihoods.[410] DCN stated that generative AI systems' use of news articles appropriates their value and "may make it impossible for publishers to continue to create, develop, and publish new articles and other materials, which is surely not in the public interest."[411] Others maintained that the benefits of high-quality AI could be achieved with fully-licensed datasets. Commenters cited several examples of AI tools trained on licensed or public domain content, such as Adobe's Firefly (an image generator), Boomy (a music generator), Getty Images' AI image generator, and Stability AI's Stable Audio (a music generator).[412]

Levendowski, *Fairer Public Benefit in Copyright Law*, 47 CARDOZO L. REV. 1 (forthcoming 2025), https://ssrn.com/abstract=5080208 (examining thirty-eight US copyright cases raising "whether a secondary use serves a public benefit").

[407] OpenAI Initial Comments at 2–4. A few commenters also stressed AI's potential to maximize production of new expressive materials. *See, e.g.*, Scenario Initial Comments at 11; TechNet Initial Comments at 3; Van Lindberg Initial Comments at 29.

[408] Def.'s Notice of Mot. and Mot. for Partial Summ. J. and Opp. to Mot. for Partial Summ. J., Kadrey v. Meta Platforms, Inc., No. 23-cv-3417 (N.D. Cal. Mar. 24, 2025), ECF No. 489.

[409] *See, e.g.*, Meta Reply Comments at 1, 4–5; CCIA Initial Comments at 14, 16; Duolingo Initial Comments at 2; Project LEND Initial Comments at 12; R Street Initial Comments at 4; Microsoft-Github Joint Initial Comments at 9; Anthropic Initial Comments at 9; Stability AI Initial Comments at 15; Copia Institute Reply Comments at 3. Meta provided several hypotheticals of how it believed that limiting training pools can lead to low quality, including: a "model trained on public domain books" that fails "to understand modern customs, language, and values" and retains "the discriminatory biases inherent in texts published in the late 19th and early 20th centuries." Meta Reply Comments at 4–5. *Cf.* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 589 (2018).

[410] *See, e.g.*, Center for Art Law Initial Comments at 2 ("[T]he nature of the output produced by AI, as well as the unprecedented scale, threatens the livelihood and rights of human copyright holders."); Digital Context Next ("DCN") Initial Comments at 3; Recording Academy Initial Comments at 3 ("The potential for generative AI music to act as a market substitute against music created by humans could chill the growth and prospects of the creative workforce.").

[411] DCN Initial Comments at 3.

[412] *See, e.g.*, Creative.ai Reply Comments at 3; Copyright Alliance Initial Comments at 76–77; ASCAP Reply Comments at 2 n.1; Ed Newton-Rex Reply Comments at 2. Unlicensed training could also lead to less access to

In the Office's view, there are strong claims to public benefits on both sides. Many applications of generative AI promise great benefits for the public,[413] as does the production of expressive works. While the sheer volume of production itself does not necessarily serve copyright's goals,[414]commenters identified a wide range of potential benefits weighing in favor and against training on unlicensed copyrighted works. With regard to the fair use analysis, however, the Office cannot conclude that unlicensed use of copyrighted works for training offers copyright-related benefits that would change the fair use balance, apart from those already considered.

* * *

The copying involved in AI training threatens significant potential harm to the market for or value of copyrighted works. Where a model can produce substantially similar outputs that directly substitute for works in the training data, it can lead to lost sales. Even where a model's outputs are not substantially similar to any specific copyrighted work, they can dilute the market for works similar to those found in its training data, including by generating material stylistically similar to those works.

The assessment of market harm will also depend on the extent to which copyrighted works can be licensed for AI training. Voluntary licensing is already happening in some sectors, and it appears reasonable or likely to be developed in others—at least for certain types of works, training, and models. Where licensing options exist or are likely to be feasible, this consideration will disfavor fair use under the fourth factor.

---

information online. AI researchers have documented a rise in website settings restricting the crawling of data across the internet, and absent mechanisms to protect against undesired training, they expect "further decreases in the open web." Shayne Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons* at 4, ARXIV (July 24, 2024), https://arxiv.org/abs/2407.14933.

[413] *See* OpenAI Initial Comments at 4 (describing use of LLMs for, among other things, improvements in health, medicine, agriculture, and the preservation of language); Chamber of Progress Initial Comments at 3–4 (describing use of generative AI for medical research and vehicle safety). This is to say nothing of the benefits of *non-generative* AI systems, which have already produced miracles in scientific and medical research. *See, e.g.*, *Stopping Malaria in Its Tracks*, GOOGLE DEEPMIND (Oct. 13, 2022), https://deepmind.google/discover/blog/stopping-malaria-in-its-tracks/; BSA Initial Comments at 4 (describing uses for the diagnosis, prevention, and treatment of disease).

[414] *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 2: COPYRIGHTABILITY 36 (2025) ("If a flood of easily and rapidly AI-generated content drowns out human-authored works in the marketplace, additional legal protection would undermine rather than advance the goals of the copyright system."); Ben Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 90 (2017) ("The value in human authorship flourishes still further when it is consumed, appreciated, and transformed by other humans. This cycle of creation and engagement is what the law clumsily tries to protect and propagate. Indeed, copyright places special value on human creativity and human reading; it 'protects humans writing for humans.'").

E. Weighing the Factors

It is for the courts to weigh the statutory factors together "in light of the purposes of copyright,"[415] with no mechanical computation or easy formula. How much each factor adds to the balance, and in which direction, will depend on the facts and circumstances of the particular case.

We observe, however, that the first and fourth factors can be expected to assume considerable weight in the analysis. Different uses of copyrighted works in AI training will be more transformative than others. And given the volume, speed and sophistication with which AI systems can generate outputs, and the vast number of works that may be used in training, the impact on the markets for copyrighted works could be of unprecedented scale.

As generative AI involves a spectrum of uses and impacts, it is not possible to prejudge litigation outcomes. The Office expects that some uses of copyrighted works for generative AI training will qualify as fair use, and some will not. On one end of the spectrum, uses for purposes of noncommercial research or analysis that do not enable portions of the works to be reproduced in the outputs are likely to be fair. On the other end, the copying of expressive works from pirate sources in order to generate unrestricted content that competes in the marketplace, when licensing is reasonably available, is unlikely to qualify as fair use. Many uses, however, will fall somewhere in between.

F. Competition Among Developers

Some commenters and scholars have raised concerns about how the application of fair use will affect the competitive ecosystem. In the words of the Federal Trade Commission ("FTC"), "the evolution of the [fair use] doctrine could influence the competitive dynamics of the markets for AI tools and for products with which the outputs of those tools may compete."[416] They warn that requiring AI companies to license copyrighted works for use in training would entrench power in the largest and best-resourced companies and content owners.[417] Andreessen Horowitz asserted that "treating AI model training as an infringement of copyright would inure to the benefit of the largest tech companies—those with the deepest

---

[415] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 578 (1994).

[416] FTC Initial Comments at 5 (footnote omitted).

[417] *See, e.g.*, a16z Initial Comments at 9 ("[T]he cost of paying to license even a fraction of the content needed to properly train an AI model would be prohibitive for all but the deepest-pocketed AI developers, resulting in dominance by a few technology incumbents. This would undermine competition by the technology startups which are the source of the greatest innovation in AI."); Anthropic Initial Comments at 10; Engine Initial Comments at 4; Arun Sundararajan Initial Comments at 9; Pamela Samuelson et al. Reply Comments at 5; BigBear.ai Initial Comments at 22; CCIA Initial Comments at 14; Ad Hoc Group of Developers and Users Initial Comments at 2; R Street Institute Initial Comments at 7.

pockets and the greatest incentive to keep AI models closed off to competition."[418] R Street similarly contended that if training is not fair use, "[o]nly large entities, like tech giants, that have the resources to navigate the licensing landscape or have already amassed vast amounts of data might be able to compete effectively in the AI space."[419]

Other commenters disagreed. ASCAP argued that AI training licensing "need not pose an insurmountable obstacle to smaller AI developers" and can be "accomplished in numerous ways—e.g., grants or public funding—that do not exploit individual creators."[420] Ed Newton-Rex suggested "a revenue share between the content rights-holder and the AI provider, which can be achieved without any upfront payment," adding that "small teams and small companies are already putting in place such models, disproving the argument that they will be shut out by licensing."[421]

While concerns about the effects of licensing on competition among AI companies should not be discounted, we do not believe they alter the fair use analysis. Licensing will always be easier for those with deeper pockets, and the more works to be licensed, the greater the effect.[422] To the extent broader competition issues are at stake, they can more appropriately be dealt with by antitrust laws and the agencies empowered to enforce them. As the FTC acknowledged, "conduct that may be consistent with the copyright laws nevertheless may violate Section 5 [of the Federal Trade Commission Act]," including actions taken by large companies to entrench their positions in AI markets.[423]

---

[418] a16z Initial Comments at 8.

[419] R Street Institute Initial Comments at 7. *See also* Regulosity and Pangea Initial Comments at 12 ("While mid to large-size businesses have the financial means and workforce to hire legal teams to track down and obtain copyright use permissions . . . [e]ntrepreneurs, start-ups, and small businesses do not have the financial means or workforce to obtain permission from copyright owners.").

[420] ASCAP Reply Comments at 3 ("Licensing models are not one-size-fits-all: for instance, ASCAP's licensing system is sophisticated and flexible enough to accommodate music users of every size, ranging from the largest streaming services on the planet to mom-and-pop neighborhood businesses.").

[421] Ed Newton-Rex Reply Comments at 2.

[422] Licensing may not even be the most significant cost, as smaller players will have to pay for other resources as well, such as computing power.

[423] FTC Initial Comments at 6. *Cf.* Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1125–26 (1990) ("Additional considerations that I and others have looked to are false factors that divert the inquiry from the goals of copyright. They may have bearing on the appropriate remedy, or on the availability of another cause of action to vindicate a wrong, but not on the fair use defense.").

G. International Approaches

Other countries are also grappling with the legal issues surrounding use of copyrighted works to train AI models.[424] Several have enacted exceptions allowing for text and data mining ("TDM") that are potentially applicable to AI training.[425] TDM methods predate the current forms of generative AI. They are not necessarily "generative" in the sense of producing new expressive material but involve some of the same steps, particularly in the creation and curation of datasets. Jurisdictions with specific TDM exceptions include the European Union (EU), Japan, and Singapore.

In the EU, the 2019 Directive on Copyright in the Digital Single Market (DSM Directive) directs member states to provide exceptions for "reproductions and extractions" of copyrighted material for use in TDM, in certain circumstances.[426] Article 3 of the DSM Directive applies only to TDM activities by "research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access."[427] Article 4 is broader and applies to TDM activities by any

[424] Many countries have begun AI consultations or studies or have introduced or enacted AI specific legislation. *See* South Korean AI Basic Law (Dec. 12, 2024), https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_R2V4H1W1T2K5M1O6E4Q9T0V7Q9S0U0; Brazil Draft Bill 2338/2023; *A Consultation on a Modern Copyright Framework for Artificial Intelligence and the Internet of Things*, GOV'T OF CAN. (July 2021), https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/copyright-policy/consultation-modern-copyright-framework-artificial-intelligence-and-internet-things-0; Select Committee on Adopting Artificial Intelligence (AI), *Final Report*, PARLIAMENT OF AUSTL. (Nov. 2024), https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Adopting_Artificial_Intelligence_AI/AdoptingAI/Report. *See generally* Information Session on Copyright and Generative Artificial Intelligence - SCCR 46 Day 4 Afternoon, at 3:26L14 (Apr. 10, 2025).

[425] Text and data mining has been defined as an "automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations." Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 2019 O.J. (L. 130/92). *See also* U.S. COPYRIGHT OFFICE, SECTION 1201 RULEMAKING: SIXTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION, RECOMMENDATION OF THE REGISTER OF COPYRIGHTS 103–04 (2021) ("2021 SECTION 1201 RECOMMENDATION"), https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf.

[426] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 4, 2019 O.J. (L. 130/92). *See also* 2021 SECTION 1201 RECOMMENDATION at 103.

[427] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 3, 2019 O.J. (L. 130/92).

actor for any purpose, but conditions the availability of the exception on lawful access[428] and respecting opt-outs by copyright owners.[429]

In 2024, the EU adopted the Artificial Intelligence Act ("EU AI Act"), which references the DSM Directive's TDM exceptions in the context of generative AI. Recital 105 acknowledges that TDM techniques "may be used extensively in [the context of training AI models] for the retrieval and analysis of such content, which may be protected by copyright and related rights."[430] Article 53 obligates AI model providers to establish policies for complying with Union law and to identify and comply with copyright owner opt-outs under the DSM Directive's Article 4 TDM exception.[431]

There continues to be controversy, however, over how the TDM exceptions apply to uses involving generative AI and whether and how the opt-out provision will work. [432]

---

[428] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 4(3), 2019 O.J. (L. 130/92) ("Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.").

[429] Directive EU 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market and Amending Council Directives 96/9/EC and 2001/29/EC, art. 4(3), 2019 O.J. (L. 130/92) ("The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.").

[430] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), recital 105, 2024 O.J. (L. 2024/1689). The EU AI Act includes transparency requirements, a topic which will be further discussed in Part 4 of the Report.

[431] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), art.53, 2024 O.J. (L. 2024/1689).

[432] The issue of applicable law will also be important as training may take place in one country and deployment in another. The AI Act requires companies seeking to deploy their AI systems within EU borders to comply with EU rules on training. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Article 2, 1(c), 2024 O.J. (L. 2024/1689) (This Regulation applies to . . . . providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union). It is unclear whether other countries will decide to follow suit—and if so, what the impact would be on international commerce in AI products.

Discussions continue at both the EU level and in member states,[433] and so far there is little case law on point.[434] At this stage, it remains to be seen how that opt-out provision will be implemented by individual EU member states.

In other jurisdictions as well, various limitations or conditions have been included in TDM exceptions. Singapore's version requires lawful access to the work and limits the use of copies to the purpose of computational data analysis.[435] Copies may only be supplied to others for the purposes of verifying results or collaborative research.[436]

Japan's TDM exception allows the use of a copyrighted work for AI development or other forms of data analysis as long as the use is not to "personally enjoy…the thoughts or sentiments expressed in that work."[437] The exception does not apply if "the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation."[438] In its 2024 AI guidelines, Japan's Copyright Office explained that "enjoyment" refers to "the act of obtaining the benefit of having the viewer's intellectual and emotional needs satisfied through using the copyrighted work," citing examples such as reading literary works, appreciating musical works, and executing works of computer programming.[439] Generating material similar to the original works can be "for enjoyment," and if a user's purpose is even partly for enjoyment, the exception does not apply.[440] Similarly, "reproducing a copyrighted database work for the purposes of data

---

[433] UK Copyright and Artificial Intelligence Consultation, GOV.UK (Dec. 17, 2024), https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence. Academics have also weighed in. *See* Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training* 74 EMORY L.J. (forthcoming 2025), https://ssrn.com/abstract=4976393; Tim W. Dornis, *The Training of Generative AI Is Not Text and Data Mining*, EUROPEAN INTELL. PROP. REV. (forthcoming 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782; Tim W. Dornis & Sebastian Stober, *Copyright Law & Training of Generative AI – Technological and Legal Foundations* (2024), https://urheber.info/media/pages/diskurs/ai-training-is-copyright-infringement/e8fab9ab59-1725460935/executive-summary_engl_final_29-08-2024.pdf (English translation of executive summary), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4946214.

[434] One German court held that Germany's TDM exception for scientific research applied to a non-profit organization's reproduction of a photographer's work in the LAION dataset. *Kneschke v. LAION, LG Hamburg*, Urteil vom 27.09.2024 - 310 O 227/23, https://openjur.de/u/2495651.html.

[435] Copyright Act of 2021, div. 8, § 244.

[436] *Id.*

[437] Copyright Act, Act. No. 48 of 1970, art. 30-4, amended up to July 19, 2024.

[438] *Id.*

[439] Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* (May 2024), https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.

[440] *Id.* In a recent presentation, the Director-General of the Agency for Cultural Affairs of Japan elaborated that using a small dataset of a creator's works or style might not be allowed. *See generally* Information Session on Copyright and

analysis, such as AI training for which licenses for data analysis are available in the marketplace," is not covered.[441]

UK law contains a narrower exception, dating back to 1988, that permits copying to "carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose," but only if the copier has lawful access to the work.[442] As part of its recent consultation on Copyright and Artificial Intelligence, the government has inquired into the application of this exception to AI and sought comments on introducing a TDM exception subject to copyright owner opt-outs, similar to the approach in the EU.[443] This proposal has proved quite controversial, with commenters warning that it would impose burdensome transaction costs for both copyright owners and AI developers.[444]

Other countries have approached the legal status of AI training through the lens of fair use. In Israel, the copyright law includes a provision closely modeled on section 107 of the U.S. Copyright Act. In December 2022, the Ministry of Justice released an Opinion on the uses of copyrighted materials for machine learning,[445] concluded that the use of copyrighted materials in machine learning datasets and training process is, in most but not all cases, fair use.[446] It

---

Generative Artificial Intelligence - SCCR 46 Day 4 Afternoon, at 3:26, 14 Apr. 10, 2025; *see also* Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* 6 (May 2024).

[441] Legal Subcommittee under the Copyright Subdivision of the Cultural Council, *General Understanding on AI and Copyright in Japan* 6 (May 2024).

[442] Copyright, Designs and Patents Act 1988, § 29A.

[443] UKIPO, *Consultation of the Intell. Prop. Office on Copyright and Artificial Intelligence*, ¶¶ 67–74, https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence.

[444] *See* Dan Milmo, *Why Are Creatives Fighting UK Government AI Proposals on Copyright?*, THE GUARDIAN (Feb. 24, 2025), https://www.theguardian.com/technology/2025/feb/25/why-are-creatives-fighting-uk-government-ai-proposals-on-copyright; Jennifer Hahn, *Copyright Exemption Plans for AI Are "Nothing Less than Vandalism" Says UK Architects and Designers*, DEZEEN (Apr. 1, 2025), https://www.dezeen.com/2025/04/01/uk-copyright-ai-exemption-letter/; Sam Tabahriti, *Musicians Release Silent Album to Protest UK's AI Copyright Changes*, REUTERS (Feb. 25, 2025), https://www.reuters.com/lifestyle/musicians-release-silent-album-protest-uks-ai-copyright-changes-2025-02-25/; Joseph Bambridge, *OpenAI, Google Reject UK's AI Copyright Plan*, POLITICO (Apr. 3, 2025), https://www.politico.eu/article/openai-google-reject-uks-ai-copyright-plan/.

[445] Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for Machine Learning* (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf.

[446] The Ministry explained its conclusion as follows: "the *purpose and character of the use* is typically transformative and done for a worthy cause, albeit sometimes commercial; the *character of the work* differs from one case to another, and cannot be categorically addressed; the *scope of use* points in the direction of fair use in most cases, in particular when despite the reproduction of the work in full, the learning is done from its noncopyrighted parts; and the *impact on the market* of the work is negligible at best, both based on the present situation and in light of a structural analysis of the content markets in the online arena." Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for*

cautioned, however, that the Opinion "does not apply to [machine learning]-based products, but only to the learning process itself. The infringing status of the product will be examined ad-hoc based on extant copyright rules and standards, and this Opinion does not grant products an a-priori safe harbor."[447]

In Korea, the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission in 2023 released *A Guide on Generative AI and Copyright*.[448] The guide recognizes that there is "an ongoing debate within academia on the applicability of the fair use rule"[449] and observed that until "several related court precedents accumulate," the "applicability of the fair use defense will remain unclear," leaving open the possibility that "using a work for AI training without permission from the copyright holder" may constitute infringement.[450]

Approaches to generative AI and copyright matters in the People's Republic of China are developing, and it is not yet clear how the use of copyrighted works in training will be treated. The Copyright Act does not have an express exception for text and data mining activities or AI training.[451] Article 24 of the Act contains a list of enumerated exceptions,[452] including a new open-ended exception covering "other circumstances as provided in laws and administrative regulations."[453] With respect to litigation, one recent case held an AI platform provider contributorily liable for infringements occurring when users uploaded protected

---

Machine Learning, at 21–22 (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf. The decision notes that "[t]he exception is nondiverse datasets, such as ones that are designed to mimic the style a single author." *Id.* The Office's fair use analysis under U.S. law differs from the Ministry's views in a number of respects. *See supra* Section IV.A–IV.E.

[447] Ministry of Justice, State of Israel, *Opinion: Uses of Copyrighted Materials for Machine Learning*, at 8 (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf.

[448] A Guide to Generative AI and Copyright, Korean Ministry of Culture, Sports and Tourism (Dec. 27, 2023), https://www.korea.net/Government/Briefing-Room/Press-Releases/view?articleId=391&insttCode=A260123&type=N.

[449] Ministry of Culture, Sports and Tourism & Korea Copyright Comm'n, A Guide on Generative AI and Copyright, at 16 (2023), https://www.copyright.or.kr/eng/doc/etc_pdf/Guide_on_Generative_AI_and_Copyright.pdf.

[450] Ministry of Culture, Sports and Tourism & Korea Copyright Comm'n, A Guide on Generative AI and Copyright, at 17 (2023).

[451] Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. (forthcoming 2025). *See also* Prof. Wang Quin, WIPO Conversation on Intellectual Property and New Technologies, Apr. 23, 2025, at 1:13:35–1:19:33, https://webcast.wipo.int/video/WIPO_IP_CONV_GE_25_2025-04-23_AM_124875.

[452] Copyright Law of the People's Republic of China (promulgated by the Standing Comm. Nat'l People's Cong., Sept. 7, 1990, amended Nov. 11, 2020, effective June 1, 2021), art. 24. *See, e.g.*, Jie Hua, *Copyright Exceptions for Text and Data Mining in China: Inspiration from Transformative Use*., 69 J. COPYRIGHT SOC'Y 123 (2022). These statutory exceptions in Article 24 are sometimes colloquially referred to as 'fair use' exceptions but they are not structured like the U.S. doctrine with its four factors.

[453] Copyright Law of the People's Republic of China (promulgated by the Standing Comm. Nat'l People's Cong., Sept. 7, 1990, amended Nov. 11, 2020, effective (June 1, 2021), art. 24(13).

content into models available via the platform, which generated infringing copies.[454] While there have been other cases involving infringing output,[455] it appears that courts have yet to consider a copyright infringement claim against a foundation model developer based on the use of copyright protected works to train a foundation model.[456] Meanwhile, press reporting on the annual work report from the Supreme People's Court indicates that the issue of intellectual property and AI is an area of ongoing attention.[457] China has also issued at least two administrative measures providing guidance on generative AI services, including compliance requirements for training data.[458] Avenues for supporting and developing the AI sector were topics receiving significant press coverage in relation to the March 2025 National People's Congress.[459]

---

[454] On appeal, the Hangzhou Intermediate People's Court considered under what circumstances a service provider might need to delete not just infringing outputs but also the model that produced them. Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号)) upholding *SCLA v. Hangzhou AI Company* [2024] Hangzhou Internet Court (2024) Zhe 0192 Min Chu No.1587. (2024浙0192民初1587号) ("[A]n intelligent technology company in Hangzhou should delete the allegedly infringing Ultraman LoRA model, and should stop providing the release and application services of the relevant Ultraman LoRA model" (machine translation)). The Court further distinguished the training activities targeting the Ultraman IP from other types of training activities that, in the courts view, do not have a purpose of reproducing original expression. *Id*. *See* Song, Seagull et al. *AI-Generated Content and Copyright (China),* PRACTICAL LAW (Mar. 8, 2025), uk.practicallaw.thomsonreuters.com/w-042-2994 (citing *SCLA v Hangzhou AI Company* [2024] Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号)) upholding *SCLA v. Hangzhou AI Company* [2024] Hangzhou Internet Court (2024) Zhe 0192 Min Chu No.1587. (2024浙0192民初1587号).

[455] *See SCLA v AI Company* [2024] Guangzhou Internet Court (Yue 0192 Min Chu No. 113) (（2024）粤 0192 民初 113 号).

[456] For example, in *SCLA v. Hangzhou AI Company* the defendant was an AI service provider who interfaced with a third-party AI model. *SCLA v Hangzhou AI Company* [2024] Hangzhou Intermediate People's Court (Zhe 01 Min Zhong No.10332) ((2024)浙01民终10332号).

[457] Meredith Chen, *China's supreme court puts AI protections on its 2025 agenda*, SOUTH CHINA MORNING POST (Mar. 12, 2025), https://www.scmp.com/news/china/politics/article/3301639/chinas-supreme-court-puts-ai-protections-its-2025-agenda.

[458] *See* Interim Measures for the Management of Generative Artificial Intelligence Services, https://www.chinalawtranslate.com/en/generative-ai-interim/; Basic Safety Requirements for Generative Artificial Intelligence Services, https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf. *See also* Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. (forthcoming 2025); Prof. Wang Quin, WIPO Conversation on Intellectual Property and New Technologies, Apr. 23, 2025, at 1:18:30 – 1:19:45. https://webcast.wipo.int/video/WIPO_IP_CONV_GE_25_2025-04-23_AM_124875.

[459] *China says it will increase support for AI, science and tech innovation*, REUTERS (Mar. 4, 2025), https://www.reuters.com/technology/china-says-it-will-increase-support-ai-science-tech-innovation-2025-03-05/; *DeepSeek-Fueled AI Fever Injects New Energy Into China's NPC*, BLOOMBERG NEWS (Mar. 11, 2025), https://www.bloomberg.com/news/articles/2025-03-11/deepseek-fueled-ai-fever-injects-new-energy-into-china-s-npc.

Finally, a few countries are considering statutory approaches to compensation.[460] In Brazil, a pending bill would require AI companies to compensate rightsholders for the use of their works in training.[461] The draft directs the parties to discuss compensation in a manner that allows rightsholders to negotiate effectively either directly or collectively, calculate compensation that reasonably and proportionally considers the AI agent's size and the potential competition impacts; and preserves freedom of agreement.[462] In 2024, Spain opened public commentary on a Draft Royal Decree which would establish an extended collective licensing mechanism for the mass exploitation of protected works in the development of AI models,[463] although the proposal was subsequently withdrawn.[464]

In the NOI, we asked "[a]re there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?" A number of commenters suggested that harmonization would be valuable to AI developers and copyright owners.[465] Several addressed AI legislation elsewhere, particularly regarding TDM, transparency, and permissions signaling,

---

[460] *See generally* Christopher Geiger & Vincenzo Iaia, *The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI*, COMPUT. LAW & SEC. R. Vol. 52 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4594873 (advocating a compulsory license in the European Union "to address copyright issues related to Generative AI in a fundamental rights-compliant manner. Indeed, it enhances a complementary dialogue between the imperative of access to in-copyright works (which in this case is a technical necessity for the development of AI systems) and the protection of the moral and material interests of creators.").

[461] *See* Article 65 of Bill 2338/2023, https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. The bill was approved by the Senate in December 2024 and as of March 2025 is under consideration in the Chamber of Deputies.

[462] *See* Article 65 of the Bill 2338/2023.

[463] Draft Royal Decree Regulating the Granting of Extended Collective Licenses for the Mass Exploitation of Works and Subject-matter Protected by Intellectual Property Rights for the Development of Artificial Intelligence Models for General Use https://www.cultura.gob.es/servicios-al-ciudadano/informacion-publica/audiencia-informacion-publica/cerrados/2024/concesion-licencias-colectivas.html.

[464] IPA Editor, *Madrid Withdraws the Royal Decree on AI Licenses*, INT'L PUBLISHERS ASS'N (Feb. 25, 2025), https://internationalpublishers.org/madrid-withdraws-the-royal-decree-on-ai-licenses/; *SPAIN: EWC supports its Spanish Members against abusive AI training*, EUROPEAN WRITER'S COUNCIL (Feb. 12, 2025), https://europeanwriterscouncil.eu/spain-ewc-supports-its-spanish-members-against-abusive-ai-training/.

[465] Copyright Alliance Initial Comments at 19 ("U.S. rightsholders are not isolated or unaffected by international developments, and so it is vital that international approaches to AI and copyright are harmonized in that they respect and uphold the copyright of human creators and copyright owners."); Stability.ai Initial Comments at 8 ("A patchwork of different copyright laws governing model development could impede AI innovation around the world.").

but they did not call for the United States to emulate these approaches.[466] Meta reported that "[c]ountries around the world have adopted express and broad text- and data-mining (TDM) or fair use exceptions, creating similarly enabling environments for technological advancement and investment."[467] UMG noted that the TDM exceptions in Japan and Singapore were enacted before the rise of generative AI, observing that "[w]hatever their historical merit, generative AI poses threats that render them obsolete and damaging for the creative community, the music industry, and the general integrity of intellectual property law."[468]

A number of commenters discussed the EU framework, particularly to criticize its opt-out provisions.[469] Some stressed that copyright is by its nature fundamentally an opt-in system of exclusive rights, or asserted that requiring opt-outs would be burdensome.[470] NMPA cautioned against the creation of "a patchwork of international exemptions with varying opt-out requirements" which would be "difficult if not impossible for most rightsholders to navigate."[471] Others raised concerns about the persistence of opt-outs given the frequency of metadata stripping and their limited usefulness when works are obtained from unauthorized sources.[472] One commenter noted that the feasibility of opt-out regimes may vary by model or type of work.[473]

Additionally, some commenters argued that the United States is treaty-bound to prohibit the unlicensed use of copyrighted works for AI training.[474] CISAC, for example,

---

[466] ASCAP Initial Comments at 10; ImageRights International Reply Comments at 2; Patrun, Inc. Initial Comments at 3; Copyright Clearance Center Initial Comments at 3–5; Conal Osfield Initial Comments.

[467] Meta Initial Comments at 19.

[468] UMG Initial Comments at 17; *see also* AAP Reply Comments at 13.

[469] For further discussion on the EU opt-out in the context of its text and data mining exception, *see infra* Section V.B.3.

[470] *See* European Writers' Council Initial Comments at 12; NMPA Initial Comments at 6.

[471] NMPA Initial Comments at 6.

[472] *See* Association of Medical Illustrators Initial Comments at 3 ("In practicality, the opt-out approach is a red herring because metadata is easily removed, and artists will never have access to header code or .htaccess files on websites where their copyrighted works appear. The biggest reason opt-outs will never work for copyright owners is that online piracy, shadow libraries, and dark web image banks are rampant — providing an endless supply for AI bot crawlers."); Copyright Alliance Initial Comments at 87.

[473] Stability AI Initial Comments at 15.

[474] *See, e.g.*, CISAC Initial Comments at 3–4; International Authors Forum Initial Comments at 1 ("As a party to the Berne Convention, the USA must also ensure that the three-step test is upheld for exceptions, especially in the protection of works by text authors. Any attempt to define uses such as copying for the development of AI models without permission as fair use would be unfair, and in violation of the three-step test, this must be made clear."); Professional Photographers of America Initial Comments at 10. ("PPA considers any exception that broadly allows

maintained that extending fair use to cover generative AI training "violates the 'three-step test'"[475] in various copyright treaties to which the United States is a party.[476] Another stakeholder argues that an opt out-based exception is unworkable and inconsistent with treaty obligations.[477]

These are still early days, and it remains to be seen how exceptions elsewhere will be applied or what new ones will be developed. Already, however, a few common elements can be observed. Governments and courts are endeavoring to differentiate among the different acts involved in assembling data, training models, and producing outputs. Many of the relevant provisions distinguish between uses for scientific, analytical, or educational purposes and other uses, notably for enjoyment purposes. And several condition eligibility for exceptions on lawful access to works in the training data.

As other countries determine their approaches to generative AI training, the Copyright Office will continue to monitor developments to assess the implications for U.S. copyright policy.

scraping of copyrighted works without the authorization of the copyright owner to violate all three steps of the three-step test that governs permissible exceptions to copyright in international instruments.").

[475] CISAC Initial Comments at 3–4.

[476] The three-step test requires that copyright exceptions be limited to special cases that do not conflict with normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author. *See* Berne Art. 9(2); *see also* WIPO Copyright Treaty art. 10, Dec. 20, 1996, 36 I.L,M. 65 (1997); WIPO Performances and Phonograms Treaty arts. 16, Dec. 20, 1996, 36 I.L.M. 76 (1997).

[477] Nicholas Caddick, Opinion provided to the Publishers Association on the Proposed UK Exception for Text and Data Mining, Feb. 21, 2025, https://www.publishers.org.uk/wp-content/uploads/2025/03/Legal-Opinion-of-Nicholas-Caddick-KC-Berne-Convention.pdf.

V. LICENSING FOR AI TRAINING

To the extent that some uses of copyrighted works to train AI models will require licensing, what forms of licensing can best accommodate the interests of both copyright owners and AI companies? This section sets out different options and considers their benefits and challenges.

The NOI asked several questions on this topic, including whether direct or collective voluntary licensing is feasible in some or all creative sectors, what legal, technical, or practical issues there might be, and whether Congress should consider establishing a compulsory licensing or extended collective licensing (ECL) system.[478] Commenters provided extensive information in response, with a range of views. Below we first discuss voluntary licensing issues and then the possibility of government intervention.

A. Voluntary Licensing

Voluntary licenses, negotiated in the free market, enable parties to set terms tailored to the specific uses of the works. These agreements can be negotiated on an individual (direct) or collective basis. Collective voluntary licensing agreements are often administered by third-party organizations (typically called "collective management organizations" or "CMOs"), authorized by multiple copyright owners to negotiate on their behalf and collect and distribute royalties.[479]

As discussed above, voluntary licensing of copyrighted works for use in AI training is increasingly taking place. As of the end of 2023, commenters reported that AI developers and copyright owners had entered into license agreements in several sectors,[480] and more individual and collective licensing has occurred since.[481] But questions remain about the extent to which voluntary licensing is feasible for different types of works and fully able to meet the needs of the AI industry.[482]

---

[478] See NOI Questions 6.2, 9–9.1, 10–10.14.

[479] By acting as a central clearinghouse, CMOs can enable transactions that might not otherwise take place. A CMO's centralized infrastructure can also provide for streamlined transactions and efficient ongoing licensing administration, reducing overall costs for owners and users alike. In this Report, we refer to "CMOs" broadly to include any entity or organization that aggregates and licenses the rights of multiple copyright owners. In addition to including traditional "collectives," we are also including entities that are more commonly referred to as rights "administrators." Examples include Artists Rights Society (visual artworks), ASCAP (performances of musical works), CCC (textual works), Harry Fox Agency (music works), and Merlin (digital music licensing). In many countries, CMOs must be authorized by the government, but most American CMOs are private entities.

[480] *See supra* text accompanying notes 383–387.

[481] *See id.*, text accompanying notes 390–394.

[482] *See id.*, text accompanying notes 395–396

Apart from the impact on the actual or potential market for copyrighted works, discussed above in the context of fair use, commenters focused on three main topics: (1) the feasibility of voluntary licensing; (2) the ability to provide meaningful compensation; and (3) possible legal impediments to collective licensing.

1. Feasibility of Voluntary Licensing

Many commenters, generally representing technology interests, expounded upon logistical, financial, and other challenges involved in voluntary licensing, including whether a sufficient quantity and variety of works can be licensed at the scale necessary to train high-quality models. They asserted that the cost of licensing copyrighted works for AI training would create an insurmountable obstacle.[483] For example, a16z stated that, "under any licensing framework that provided for more than negligible payment to individual rights holders, AI developers would be liable for tens or hundreds of billions of dollars a year in royalty payments," which would serve as a barrier to AI development and innovation.[484] Several commenters expressed concern about the financial impact of a licensing requirement on researchers in particular, including those "who want to try to solve the many problems associated with AI (such as detecting 'deep fakes,' preventing 'hallucinations,' 'unlearning' information, and reducing computing's energy demands)."[485] Meta also pointed to the potential impact on open-source licensing of AI models, arguing that "no company could afford to pay

---

[483] *See, e.g.*, BigBear AI Initial Comments at 22 ("Licensing requirements can lead to increased costs for AI developers and organizations. They may need to pay for licenses, royalties, and legal services, potentially raising the barrier to entry for smaller players and startups."); Hugging Face Initial Comments at 11 ("An outcome where licensors pay millions of dollars to train on hundreds of thousands or millions of works under copyright would constitute a 'worst of both worlds' outcome in our assessment, as such a deal would be costly enough to exclude any but the very largest companies from training new models, while still providing negligible additional income to the original data creators."); R Street Initial Comments at 5 ("[T]he costs associated with obtaining these licenses could make AI projects excessively expensive, thus impeding innovation and hindering industry growth. This approach may render many AI-driven projects unattainable, particularly for smaller entities or researchers with limited resources.").

[484] a16z Initial Comments at 10–11.

[485] *See, e.g.*, ACM Tech Policy Committee Reply Comments at 1–2 (a direct licensing requirement could be especially challenging "for academic researchers and institutions since it is unlikely that funding agencies, such as the National Science Foundation, would underwrite the time, effort, and expense of contacting every copyright owner" such these researchers "might have to do their research using limited training material not representative of the real world or be unable to do research at all"); *see also* Project Lend Initial Comments at 12 ("Any opt in/out regime or voluntary licensing scheme could exacerbate this effect and have the added consequence of pricing out those who cannot afford the licensing fees, halting many uses, including research and scholarship."); Anthropic Initial Comments at 10 ("Efforts to research the safety and interpretability of these models would be particularly undermined, and likely result in only the most highly resourced entities being able to advance research in this space, as our empirical work shows that research on the largest and most capable systems is qualitatively different than for small models."); PIJIP Initial Comments at 6 ("The creation of a licensing requirement at this early stage would limit research, not-for-profit uses, and would lock in advantages for large commercial actors, who can negotiate licensing agreements."); Van Lindberg Initial Comments at 36 ("A licensing requirement for AI systems would stop most AI research and development in the United States.").

licensing fees based on third-party uses of that company's models, and even tracking how models were used would be impracticable."[486]

Commenters also cited practical challenges in securing licenses for the volume and variety of works potentially needed for AI training.[487] R Street stated that "[t]he process of identifying, negotiating and securing licenses for every individual piece of content in a dataset would be resource-intensive. These increased costs could be passed on to consumers or could deter companies from pursing certain AI-driven projects altogether."[488] According to several commenters, these problems would be compounded by the difficulty in determining ownership of many of the works in training datasets,[489] a necessary predicate to entering into licensing negotiations. For example, Meta contended that "it would be impossible for AI developers to license the rights to other critical categories of works—like internet reviews and other examples of casual, vernacular text—both because it would be impossible to locate the owners of such works, and administratively impossible to negotiate licenses with each of them."[490] It asserted that even collective licensing would create "massive administrative problems."[491]

---

[486] Meta Initial Comments at 17.

[487] *See supra* text accompanying notes 399–403 (discussing administrative and transactional costs of voluntary licensing); *see also* CCIA Initial Comments at 14 ("Especially in the digital age, when large volumes of work are produced and published online each day, it is dubious that any licensing process will be able to keep up with non-AI innovation, calling into question the technology's utility."); Lee Hollaar Initial Comments at 4 ("It may be an insurmountable task to obtain 'affirmative consent from a large number of copyright owners."); Microsoft Initial Comments at 9 ("Any requirement to obtain consent for accessible works to be used for training would chill AI innovation. It is not feasible to achieve the scale of data necessary to develop responsible AI models even when the identity of a work and its owner is known."); OpenAI Initial Comments at 13 ("The diversity and scale of the information available on the internet is thus both necessary to training a 'well-educated' model . . . and also makes licensing every copyrightable work contained therein effectively impossible.").

[488] R Street Initial Comments at 5.

[489] *See, e.g.*, CCIA Initial Comments at 15 ("Much of the material on which generative AIs are trained may lack any identified or identifiable author from whom to obtain a license. Even where an author might be identified, contacting them might be difficult or impossible."); Anthropic Initial Comments at 9 ("However, a regime that always requires licensing for use of material in training would be inappropriate; it would, at a minimum, effectively lock up access to the vast majority of works, since most works are not actively managed and licensed in any way."); Hugging Face Initial Comments at 11 ("While opting into the use of work as training material may be a medium to long-term goal, it is not currently feasible to seek opt-ins for already published data—especially as the majority of data under copyright on the web does not have an easily identifiable rights holder."); *see also* EFF Initial Comments at 4 ("It would not be feasible to seek authorization from every copyright owner, particularly since the elimination of formalities means that copyright attaches at fixation to all sorts of amateur creations not part of any market.").

[490] Meta Initial Comments at 17.

[491] Meta Initial Comments at 20; *see also* Engine Initial Comments at 8 ("The combination of the need for diverse data sets that could contain anything in the universe of expressive material eligible for copyright protection and the indirect—and even diminishing—value of each individual piece of data that an AI model is trained on, means that no

Commenters representing copyright owner and creator interests, on the other hand, argued that the costs or difficulty of obtaining licenses for the volume of works required for AI training is not an excuse for failing to do so.[492] They contended that obtaining licenses is simply a cost of doing business,[493] and one that AI companies can afford,[494] especially where their commercial products depend on the use of copyrighted works.[495] Authors Guild stated,

---

existing model for large scale licensing can be easily applied to AI training and development. Existing collective licensing mechanisms have been most successful in the context of homogeneous transactions among repeat players with similar preferences, which does not describe the way AI models interact with copyrighted material or the state of the AI ecosystem.") (internal quotation marks omitted); *cf.* CCIA Initial Comments at 11 ("[w]hile obtaining permission from, e.g., songwriters may be viable through existing collective licensing groups, training data in less common languages or from various subcultures are far less likely to be organized and the appropriate entity to contact for permission may even be impossible to determine").

[492] *See, e.g.*, AAP Initial Comments at 23 ("The claim that the volume of works used for training makes it burdensome for a Gen AI systems developer to seek permission is not an excuse for infringing on the copyrights and livelihoods of the thousands of authors, publishers, and other artists."); Graphic Artists Guild Initial Comments at 16 ("The critical issue in obtaining licenses for generative AI for images is the sheer volume of licenses that are required for training. However, licenses should be obtained as licenses have always been obtained – by negotiating with the visual artist, the visual artist's agent, or an entity empowered to negotiate on the visual artist's behalf."); Monotype Reply Comments at 3 ("The volume of material ingested to train AI models is (or can be) massive . . . ; however, the volume of ingested works should not negate the responsibility of developers of AI models to respect the copyrights of others. Just because it's a big job doesn't mean it shouldn't be done."); Getty Images Reply Comments at 10 ("A number of commenters have complained that licensing the use of copyrighted works in training sets would be either impossible, impractical, or unduly expensive because of the sheer number of works some model developers would like use for training purposes. The scope of infringement in which an infringer would like to engage hardly excuses the infringement."); Authors Guild Reply Comments at 4 ("It would turn copyright law on its head to hold that a party can avoid liability as long as its infringements are too numerous to account for.").

[493] *See, e.g.*, Copyright Alliance Initial Comments at 72 ("Licensing copyrighted works is a normal cost of doing business, and licenses are entered into across the spectrum of copyright industries."); MPA Initial Comments at 34 ("Transaction costs in the area of intellectual property are a routine cost of doing business, particularly for access to a large amount of content. Those costs are neither new nor unique in the context of training AI models."); N/MA Reply Comments at 22–23 ("Especially in light of the tremendous economic benefits [AI] companies and their backers are poised to enjoy, they should be required to factor content acquisition costs into their models, just like any other cost of doing business."); UMG Initial Comments at 70 ("Simply asking the AI community to take the time necessary to license music is an appropriate and necessary 'cost' that far outweighs the appropriation of an entire artform without permission or compensation. Other legitimate businesses that use copyrighted works en masse bear those licensing costs and have each negotiated agreements that fit their particular needs. It would be unjust to relieve the AI industry from that same responsibility.").

[494] *See, e.g.*, Getty Images Initial Comments at 21 ("The multi-billion-dollar scale of investment that leading technology companies have made in developing AI Systems and AI Models accommodates the cost of obtaining licenses and there is no reason to believe that respect for copyright laws in the context will inhibit innovation."); UMG Initial Comments at 70 ("[M]any of the key players in the AI industry are huge companies that should have little difficulty absorbing this necessary expense.").

[495] *See, e.g.*, Copyright Alliance Reply Comments at 28 ("Quite simply, an AI tool has no value without the copyrighted materials on which they are trained, and the AI tool operators should not profit at the expense of the copyright owners whose valuable content is an essential part of the value of the AI tool."); AAP Reply Comments at 2

"Arguments that it is too expensive do not justify the use [without permission]. AI companies are spending millions and even billions on development and computing power. Why should the authors' contribution be free for the taking when generative AI is nothing without the works it is trained on?"[496] In the Copyright Alliance's view, "[t]he idea that just because it may be harder to get consent from copyright owners when large volumes of works are being used, it is therefore not infringement, would simply incentivize infringers to illegally copy more as a means for avoiding infringement—that cannot possibly be the law."[497]

These commenters also disputed the factual premise that voluntary licensing is infeasible. Getty Images asserted that "[l]icenses to scaled quantities of content and metadata required to train Generative AI Models are already readily available," and "[t]he claim by some developers that there is no way to get consent from copyright holders given the quantity of materials needed to train AI Models is simply untrue."[498] It stated that "[t]here is an established market for training data, and there is a growing body of high-quality Generative AI Models that have been trained on content licensed for that purpose."[499]

Commenters also pointed out that AI licensing deals are already occurring,[500] pointing to a growing number of examples of fully licensed models in certain sectors and for certain purposes. Some AI developers describe their companies, products, and models as relying exclusively on owned or licensed data,[501] and at least one organization, Fairly Trained, has

("It is deeply ironic that these billion-dollar companies bemoan the financial burden they would face if they were required to pay reasonable license fees to the copyright owners whose works are the very building blocks of Gen AI and whose livelihoods are threatened by the same systems.").

[496] Authors Guild Reply Comments at 4.

[497] Copyright Alliance Initial Comments at 72.

[498] Getty Images Initial Comments at 20.

[499] *Id.* at 20; *see also, e.g.*, MPA Initial Comments at 30 ("MPA speaks only on behalf of its members, but the fact that some individual copyright owners and AI companies already are engaged in licensing on an individual basis suggests that voluntary licensing is feasible in various creative sectors."); Copyright Alliance Initial Comments at 75 ("Yes, voluntary licensing is feasible, as evidenced by existing agreements between AI developers and copyright owners for generative AI training (and other previous technological innovations in the way copyrighted content is used and distributed) and licenses that are being developed by rights owners.").

[500] *See supra* text accompanying notes 383–387

[501] *See, e.g.*, Getty Images Initial Comments at 5 (describing its product "Getty AI by Getty Images" as a text-to-image tool "trained exclusively on licensed content," which would provide recurring compensation to copyright owners whose content was used in training); *Bria AI Accountability Framework: Being a Responsible AI Developer*, BRIA.AI, https://bria.ai/responsible-ai-policy ("We use only commercially licensed data explicitly authorized for training of generative AI models. . . By successfully building high-quality models through sustainable data partnerships, Bria demonstrates that responsible innovation and respect for intellectual property are not only possible, but commercially viable. This evidence is crucial for policymakers, regulators and courts, showing there is no need to choose between fostering AI progress and protecting creators.").

established mechanisms to certify such claims.[502] Fully licensed training datasets have supported the production of AI models and products capable of producing text,[503] images,[504] and music.[505] Of these, music models are the most common to be certified by Fairly Trained.[506]

AI companies and supporters stressed that current licensing activity does not demonstrate the feasibility of voluntary licensing at scale across all contexts. For example, a16z stated that "[t]he fact that large rights owners are willing to strike deals is irrelevant, as such deals would only permit use of a small amount of the content needed to adequately train AI systems."[507] Meta asserted that "it would be impossible for any market to develop that could enable AI developers to license all of the data their models need," noting that "[g]enerative AI models need not only a massive *quantity* of content, but also a large *diversity* of content," and deals with individual rightsholders "would provide AI developers with the rights to only a

---

[502] *See Fairly Trained Certified Models*, FAIRLY TRAINED, https://www.fairlytrained.org/certified-models. This organization, founded by Ed Newton-Rex, provides certification for "any generative AI model that doesn't use any copyrighted work without a license," and "exists to make it clear which companies take a more consent-based approach to training, and are therefore treating creators more fairly." *See About*, FAIRLY TRAINED, https://www.fairlytrained.org/about.

[503] *See* Press Release, 273 Ventures, *Meet KL3M: the first Legal Large Language Model* (Feb. 20, 2024), https://273ventures.com/kl3m-the-first-legal-large-language-model/ ("The genesis of KL3M lies in our Kelvin Legal DataPack, a proprietary dataset that now contains over two trillion tokens of legal, financial, and general domain text. Our DataPack is the first large-scale, commercially-available dataset collected with clear provenance and legal permissibility for training commercial models."); Michael J Bommarito II, Julian Bommarito, and Daniel Martin Katz, *The KL3M Data Project: Copyright-Clean Training Resources for Large Language Models*, ARXIV 23 (Apr. 9, 2025), https://arxiv.org/abs/2504.07854 ("the dataset provides a comprehensive foundation for small or domain-specific model pre-training that can be supplemented with other appropriately licensed datasets. . . . We believe that the KL3M Data Project has empirically demonstrated that large-scale, high-quality data collection can successfully operate within established legal and ethical boundaries.").

[504] *See, e.g.*, *Firefly*, ADOBE, https://www.adobe.com/products/firefly.html ("Trained on content we have permission to use, like Adobe Stock, Firefly is designed to be safe for commercial use."); Press Release, Getty, *Getty Images Launches Commercially Safe Generative AI Offering* (Sept. 25, 2023), https://newsroom.gettyimages.com/en/getty-images/getty-images-launches-commercially-safe-generative-ai-offering; *AI Image Generator for Enterprise*, SHUTTERSTOCK, https://www.shutterstock.com/business/generative-ai.

[505] *Ethical AI in Music: Navigating Copyright Concerns*, SOUNDRAW (Aug. 6, 2024), https://blog.soundraw.io/post/ethical-ai-in-music ("Unlike platforms that train on copyrighted material without permission, we use music entirely produced in-house. This ensures that every track you generate is free from copyright infringement."); *AI Music Generator for Commercial Use with Rightsify's Hydra*, RIGHTSIFY, https://rightsify.com/hydra/ ("Rightsify is committed to respecting copyright, and the Hydra dataset is limited to Rightsify's data to ensure the uniqueness and legality of the generated music."); Ashley King, *Music AI, Creator of Moises, Raises $40 Million in Series A Funding – With a Mission to Build the Future of Ethical AI in Music*, DIGITAL MUSIC NEWS (Jan. 22, 2025) ("The company is committed to developing ethical AI solutions strictly trained on fully licensed content").

[506] *See Fairly Trained Certified Models*, FAIRLY TRAINED, https://www.fairlytrained.org/certified-models.

[507] a16z Initial Comments at 9.

miniscule fraction of the data they need to train their models."[508] Meta also disputed the viability of fully licensed models, contending that "there is no evidence that licensed or public domain data is sufficient to build a useful state-of-the-art Generative AI model capable of competing with available alternatives."[509] It noted, however, that "[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output."[510]

Some commenters stressed that voluntary licensing would be especially challenging for smaller stakeholders on both sides. Daniel Gervais stated that "[i]t is simply not reasonable to expect a user, especially a smaller one, to identify every right holder in every copyrighted work they want to use (even assuming they can determine what is and is not a protected work) and then locate and contact those rightsholders one by one. Nor does it make business sense for even large rightsholders to have an army of licensing agents dealing with potentially thousands of small-scale users around the world, not to mention currency and linguistic barriers."[511] Others expressed concern that smaller copyright owners would have reduced bargaining power and would either be overlooked in licensing deals or would receive substandard terms.[512]

A number of commenters supported voluntary collective licensing as a way of reducing transaction costs and facilitating bulk licensing.[513] SGA called collective licensing "the most

---

[508] Meta Initial Comments at 17.

[509] Meta Reply Comments at 5–7.

[510] *Id.*

[511] Daniel Gervais Initial Comments at 4.

[512] *See, e.g.*, CCIA Initial Comments at 14 ("[I]t is unlikely that developers will expend the resources to enter into licensing agreements with less prominent creators, resulting in an undiversified dataset composed primarily of work from the largest (and likely, the most litigious) copyright holders."); Brooklyn Law Incubator & Policy Clinic Initial Comments at 12 ("Inequitable bargaining is also commonplace in voluntary licensing regimes, where one party has (i) access to better alternatives (ii) more market share or (iii) more knowledge power."); Graphic Artists Guild Initial Comments at 5–6 ("AI image generator platforms have indicated that licensing from individual visual artists is difficult if not impossible, considering the high volume of images they require. This puts individual visual artists at a disadvantage in competing for licensing agreements against entities with large libraries of licensed imagery, such as publishing houses, media companies, and stock image agencies.").

[513] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 25 ("Voluntary collective licensing that happens in the free market, without any government mandate or intervention, can be both desirable and feasible, as exemplified by the success of the digital rights agency Merlin."); AAP Initial Comments at 24 ("Voluntary collective licensing is consistent with the exclusive rights of copyright owners and may prove to be a feasible approach alongside direct licensing."); ASCAP Initial Comments at 4–5, 41–44; Artists Rights Society Reply Comments at 3–4; ASCRL Initial Comments at 4–5; CCC Initial Comments at 12, 15; Copyright Alliance Initial Comments at 77–78; Copyright Licensing Agency Initial Comments at 11; Prof. Daniel Gervais Initial Comments at 4; European Writers' Council Initial Comments at 13; Graphic Artists Guild Initial Comments at 6, 15; Music Workers Alliance Initial Comments at 5; National Writers Union Initial Comments at 16–17.

cost-effective and efficient manner of authorizing the ingestion of copyrighted works into generative AI systems."[514] Authors Guild opined that "collective licensing could solve the problem of how to license a mass number of works to AI developers for AI training on behalf of individual creators and small business on an industry-by-industry basis."[515] News/Media Alliance asserted that "[w]hile collective licensing should not be required, and individual licensing always permitted, voluntary collective licensing may well prove useful by providing the ability to aggregate smaller publishers, thereby reducing transaction costs and facilitating more efficient licensing and distribution for a greater number of licensors."[516] And Recording Academy said that while "direct licensing should be the default approach," "where direct licensing is inefficient or inaccessible with respect to independent songwriters and artists who lack the resources and leverage to successfully enter into such agreements," "voluntary collective licensing may prove beneficial."[517]

Commenters largely agreed that the quantity, quality, and type of data needed will vary among AI models, depending on their structure and intended use. And the industries from which copyrighted works are drawn reflect varied market realities, each with different licensing customs and practices. For example, while "[i]t is true that, in some modalities (e.g. text), you still need a very large amount of data to train the best models . . . it is by no means certain that this will always be the case."[518]

2. Ability to Provide Meaningful Compensation

Commenters were divided as to whether or not copyright owners can be compensated meaningfully for licensing their works for AI training. Some contended that it would not only be cost prohibitive for AI developers to pay copyright owners in the aggregate, but that

---

[514] SGA et al. Reply Comments at 10.

[515] Authors Guild Initial Comments at 25–26.

[516] N/MA Initial Comments at 14–15, 56–57.

[517] Recording Academy Initial Comments at 8. Commenters that questioned CMOs' capabilities limited their discussion to the music industry, and argued that direct licensing is feasible and preferred in that context. *See, e.g.*, Rightsify Group Reply Comments at 7 (in the music sector, "direct licensing is feasible and the better option," because "traditional [CMOs] would not be able to accurately account and pay royalties for AI licenses," as "[t]his is a new licensing model that requires high quality metadata for every musical work which is something [CMOs] have struggled with"); Music Reports Initial Comments at 4 (asserting that in the music sector, "CMOs are so poorly suited by their history and infrastructure to deal with the volume of today's digital media services that they are typically forced to outsource their administrative functions to third party providers"); UMG Initial Comments at 66 ("Generative AI requires the creativity, rapid response, and adaptability inherent in free market licensing. UMG can best meet these interests through direct licensing, rather than delegating those licensing duties to a [CMO].").

[518] Ed Newton-Rex Initial Comments at 2–3. *See* Meta Reply Comments at 5–7 ("[u]ltimately, whether it is possible to train a competent Generative AI model using only public domain or licensed data will depend on a number of fact-specific considerations, including the medium of the model's output.").

compensation to any individual copyright owner would be negligible due to the volume of works typically used for training.[519] Hugging Face deemed this a "worst of both worlds" scenario, stating that "such a deal would be costly enough to exclude any but the very largest companies from training new models, while still providing negligible additional income to the original data creators."[520]

On the other side, commenters argued that these statements ignore the value of compensation accrual over time, which can add up to meaningful amounts.[521] In the words of the Copyright Alliance, "[t]he notion that licensing should not be required because these royalties may be small would turn copyright, and many other licensing models, on its head."[522] These commenters asserted that AI training can have a positive economic impact on copyright owners,[523] motivating the creation of new works,[524] with one declaring that "[t]he economic consequences of requiring licenses will be to bolster creators, the U.S. economy, and our culture."[525] Another suggested that if AI companies struggle to compensate rightsholders in the

---

[519] *See, e.g.*, Meta Initial Comments at 20 ("[A]ny fair royalty due would be incredibly small in light of the insignificance of any one work among an AI training set."); a16z Initial Comments at 10 ("Again, a staggering quantity of individual works is required to train AI models. That means that, under *any* licensing framework that provided for more than negligible payment to individual rights holders, AI developers would be liable for tens or hundreds of billions of dollars a year in royalty payments.").

[520] Hugging Face Initial Comments at 11.

[521] *See, e.g.*, Copyright Alliance Reply Comments at 27 ("Moreover, with sufficient volume, even low 'per use' royalty rates can add up to considerable money."); N/MA Reply Comments at 25–26 ("Aggregating smaller amounts of revenue over time is a standard and typical foundation for internet, media, and other digital business models (e.g., subscription, advertising, or as-a-service models). The power of these business models is demonstrated throughout the economy, including in media publishing, which depends on subscription and advertising revenue over time, cloud computing, and music and video streaming.") (citations omitted).

[522] Copyright Alliance Reply Comments at 27; *see also, e.g.*, ASCAP Reply Comments at 3–4 ("[A] supposed concern that licensing will not compensate creators enough cannot justify a refusal to compensate them at all—as numerous AI developers flagrantly continue to do. The AI industry, like any other, must compensate creators for its use of their protected content.").

[523] *See, e.g.*, National Writers Union Initial Comments at 19 ("The economic impact of implementing opt-in licensing requirements for generative AI system training data would be a net positive. It would stimulate the economy significantly by providing marketplaces with ample opportunities for creative workers to license existing works and produce new content for training purposes."); Image Rights Int'l Reply Comments at 5 (favoring requiring copyright owners' consent because doing so "ensures that creators are fairly compensated for their work, especially when it's being used for profit-making purposes"); Recording Academy Initial Comments at 8 ("Regardless of whether it is a sound recording or a musical work, voluntary direct licensing is the preferred regime and the only proven approach to fairly compensate all artists, songwriters, and studio professionals.").

[524] *See, e.g.*, Digital Context Next Initial Comments at 4–5 ("[L]icensing would help maintain the incentive for publishers to continue creating quality new content."); AAP Initial Comments at 27 ("A licensing requirement . . . would have a positive economic impact on the development and adoption of Gen AI systems, as well as the continued creation and distribution of high-quality works by the creative sector.").

[525] CCC Initial Comments at 15.

near-term, rightsholders can negotiate licenses that forgo up-front payments or traditional royalties in exchange for later shares in revenues as the companies grow.[526]

The quality of training data may also affect potential compensation, and some have observed that quality for training purposes may correspond with works' commercial value in other contexts.[527] Licensors touted their products as attractive to AI companies because they can provide data that is newly released, high-quality, curated, and clean.[528] An AI developer might, for example, use licensed material because it is "diverse and high quality [and] long-context" and give it higher weight in training than other data.[529] Because data quality and model quality are correlated, AI firms seeking to offer higher model quality than their competitors may turn to licensing; this has resulted in what some have described as a multi-billion dollar race.[530]

3. Possible Legal Impediments to Collective Licensing

Some commenters raised concerns that copyright owners banding together to negotiate collective licenses could have antitrust implications.[531] One contended that "collective licensing

[526] CCC Initial Comments at 15; N/MA Reply Comments at 25 ("[L]icensing valuations do not need to be the same for all types of content, nor would all permissive uses be expected to be royalty bearing. . . . [V]enture capital values generative AI companies based on projections that revenue will accrue over time," up to "$1.4 *trillion* market by 2024, mainly due to incremental value projections.").

[527] *See* George Wukoson & Joey Fortuna, The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs at 7–8, ZIFF DAVIS (Nov. 4, 2024) ("LLM company statements made over the past year about licensing deals with commercial web publishers indicate that the need for high-quality training text data has only grown more acute as developers compete to keep scaling."). Wukoson and Fortuna further conclude that "LLM company training data disclosures—largely dating to earlier, pure-research periods of the technology's evolution—and analysis of public training datasets show long-running exploitation of high-quality publisher content (extremely lucrative for the LLM companies) and imply lost licensing revenue from some of the world's most highly-valued companies." *Id*. at 16.

[528] Manifesto, CREATED BY HUMANS, https://www.createdbyhumans.ai/manifesto, Dataset Providers Alliance, *Shaping the Future of AI Data – The Dataset Providers Alliance Position Paper* 2, https://www.thedpa.ai/ai-data-licensing-position-paper.

[529] *Apple Intelligence Foundation Language Models* at 4–5.

[530] Katie Paul & Anna Tong, *Inside Big Tech's Underground Race to Buy AI Training Data*, REUTERS (Apr. 5, 2024), https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/ (describing dealmaking and discussions involving various categories of works).

[531] In the music context, two CMOs that license public performances of musical works—ASCAP and BMI—are subject to longstanding antitrust consent decrees overseen by the Department of Justice. *See* U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 34–42 (2015) (discussing the history and scope of the consent decrees). Two other music CMOs—SESAC and GMR—have settled private actions alleging antitrust violations. *See* Final Order Dismissing Case With Prejudice, *Radio Music License Committee, Inc. v. SESAC, Inc.*, No. 13-cv-05807 (E.D. Pa. 2015); *Meredith Corp. v. SESAC, LLC*, 87 F. Supp. 3d 650 (S.D.N.Y. 2015) (approving a settlement "modeled on the terms of the ASCAP and BMI consent decrees"); Stipulation of Voluntary Dismissal With Prejudice, *Radio Music License Committee, Inc. v. Global Music Rights, LLC*, No. 19-cv-03957 (C.D. Cal. 2022).

is inherently anticompetitive and existing [CMOs] for music have repeatedly demonstrated their tendency to use their collective power to the detriment of both their licensees and their constituent authors."[532]

To avoid such concerns, several commenters urged adoption of an antitrust exemption allowing collective licensing of copyrighted works for AI training.[533] Others believed that statutory change was premature,[534] or suggested first seeking guidance from the Department of Justice.[535]

B. Statutory Approaches

There was little support among commenters for statutory approaches to licensing, whether compulsory licenses or ECL.

1. Compulsory Licensing

Compulsory licenses are established by law and allow use of a copyrighted work without the consent of the copyright owner. They apply to specific uses, users, and works, and require compliance with certain statutory and regulatory requirements, such as making royalty payments and related filings.

Compulsory licenses in the United States have in the past been adopted where Congress determined that the free market was incapable of supporting effective or efficient voluntary

---

[532] *See* Music Reports Initial Comments at 3.

[533] *See, e.g.*, ASCRL Initial Comments at 5–6; Authors Guild Initial Comments at 10, 26 ("What stands in the way of collective licensing is that antitrust laws impose risks to forming CMOs that set rates on behalf of their members."); Letter from Authors Guild, Summary of *Ex Parte* Meeting on May 6, 2024 Regarding the Office's AI Study, to U.S. Copyright Office at ex. A (May 10, 2024) (proposing bill text); European Writers' Council Initial Comments at 13; Graphic Artists Guild Initial Comments at 8, 15; National Writers Union Initial Comments at 10, 16–17 ("[E]xisting organizations are chilled by fear of possible antitrust enforcement, which impedes efforts to organize creative workers into [CMOs]."); N/MA Initial Comments at 58 (stating that "it is possible that legislation, such as antitrust exceptions, to augment existing abilities to negotiate collectively could be helpful" even though "it is not clear that such legislation is actually necessary given that many collective licensing entities . . . currently operate in accordance with antitrust laws without the need for legislative exceptions"); SGA et al. Reply Comments at 10.

[534] *See, e.g.*, AAP Initial Comments at 25 ("We believe it is currently premature to consider any statutory or other changes to facilitate negotiation of collective licenses."); MPA Initial Comments at 31 ("At this time, MPA does not believe there is a need for any statutory changes (such as an antitrust exemption)."); STM Initial Comments at 13; Music Reports Initial Comments at 3–4 ("Congress emphatically should not consider statutory or other changes—especially not an antitrust exemption—that would facilitate or prioritize collective licenses.").

[535] *See, e.g.*, Copyright Alliance Initial Comments at 80 (noting that "[m]any CMOs already operate without an antitrust exemption" and suggesting that a possible approach could be for the Department of Justice to provide antitrust guidance through a Business Review Letter); Getty Images Initial Comments at 20 ("[I]t would be helpful for the appropriate anti-trust authorities to issue guidance regarding the level of collaboration amongst copyright holders who wish to license collectively in this context that is permitted under existing anti-trust laws.").

licensing.[536] Because such licenses obviate the need to engage in negotiations, they can be an efficient mechanism in situations with high transaction costs to permit a publicly beneficial use of copyrighted works while providing remuneration to copyright owners.[537]

At the same time, they generally require a substantial administrative apparatus. Rate setting and distribution proceedings involve significant sums and are often contentious. Participants may spend large amounts on legal fees and proceedings can take years to reach final resolution. Many licenses have also required the promulgation of voluminous and complex regulations.

The Office has historically been wary of compulsory licenses as "a derogation of the author's right to control the use and distribution of his or her work,"[538] urging that they "should be enacted only in exceptional cases, when the marketplace is incapable of working."[539] As we have previously observed, "once a compulsory license is implemented it becomes deeply embedded in industry practices and—even when its original rationale is lost in time—is

---

[536] *See, e.g.*, H.R. REP. NO. 94-1476, at 89 (1976) (regarding the Section 111 license, concluding that "it would be impractical and unduly burdensome to require every cable system to negotiate with every copyright owner whose work was retransmitted by a cable system"); *id.* at 117–18 (regarding the Section 118 license, explaining that "public broadcasting may encounter problems not confronted by commercial broadcasting enterprises"); S. REP. NO. 115-339, at 4 (2018) (regarding a new blanket Section 115 license, explaining that "[s]ong-by-song licensing negotiations increase the transaction costs to the extent that only a limited amount of music would be worth engaging in such licensing discussions"); H.R. REP. NO. 100-887, pt. 2, at 15 (1988) (regarding the Section 119 license, referring to it being "a temporary, transitional statutory license to bridge the gap until the marketplace can function effectively"); S. REP. NO. 60-1108, at 6–9 (1909) (regarding the predecessor to the Section 115 license, concluding that its adoption was needed to prevent a monopoly from forming in the player piano roll market); *see also* H.R. REP. NO. 100-887, pt. 1, at 15 (1988) ("Congress should impose a compulsory license only when the marketplace cannot suffice."). Compulsory licenses are best understood as legislative compromises, historically accompanying an expansion of copyright owners' rights by Congress. *See* Barbara A. Ringer, *Copyright in the 1980's (*1976), https://www.copyhype.com/2023/01/barbara-a-ringer-copyright-in-the-1980s-1976.

[537] *See, e.g.*, Yafit Lev-Aretz, *The Subtle Incentive Theory of Copyright Licensing*, 80 BROOK. L. REV. 1357, 1378 (2015); Kristelia A. García, *Private Copyright Reform*, 20 MICH. TELECOMM. & TECH. L. REV. 1, 39 (2013).

[538] *See* U.S. COPYRIGHT OFFICE, A REVIEW OF THE COPYRIGHT LICENSING REGIMES COVERING RETRANSMISSION OF BROADCAST SIGNALS 32 (1997); *see also* U.S. COPYRIGHT OFFICE, SATELLITE TELEVISION EXTENSION AND LOCALISM ACT § 302 REPORT 1 (2011) ("[B]y their nature, statutory licenses are exceptions under copyright law and a limitation on the fundamental principle that authors should enjoy exclusive rights to their creative works, including for the purpose of controlling the terms of public dissemination."); U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 148 (2015).

[539] *Music Licensing Reform: Hearing Before the Subcomm. on Intell. Prop. of the S. Comm. on the Judiciary*, 109th Cong. (2005) (statement of Marybeth Peters, Register of Copyrights); *see* U.S. COPYRIGHT OFFICE, U.S. COPYRIGHT OFFICE ANALYSIS AND RECOMMENDATIONS REGARDING THE SECTION 119 COMPULSORY LICENSE 7 (2019) ("[T]he Copyright Office's long-held view [is] that a compulsory license should be utilized only if compelling reasons support its existence.") (internal quotation marks omitted); U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 163 (2015) ("[C]ompulsory licensing should exist only when clearly needed to address a market failure."); U.S. COPYRIGHT OFFICE, SATELLITE HOME VIEWER EXTENSION AND REAUTHORIZATION ACT SECTION 109 REPORT 78 (2008); U.S. COPYRIGHT OFFICE, A REVIEW OF THE COPYRIGHT LICENSING REGIMES COVERING RETRANSMISSION OF BROADCAST SIGNALS iv, 12 (1997).

difficult to undo. That alone should counsel caution in all but the most manifest instances of market failure."[540] Compulsory licenses "should be provided only if shown to be required by a clear public interest outweighing the reasons for protecting the author's rights" and "should not go any further than is shown to be necessary in the public interest."[541] Congress has expressed similar views.[542]

Most commenters who addressed this issue opposed or raised concerns about the prospect of compulsory licensing.[543] Those representing copyright owners and creators argued that the compulsory licensing of works for use in AI training would be detrimental to their ability to control uses of their works, and asserted that there is no market failure that would justify it.[544] A2IM and RIAA described compulsory licensing as entailing "below-market royalty rates, additional administrative costs, and . . . restrictions on innovation."[545] The Copyright Alliance said that it "undermines the Constitutional purposes and goals of federal copyright law and destroys the existing incentives for copyright owners to create and

---

[540] U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 168 (2015), https://www.copyright.gov/policy/musiclicensingstudy/copyright-and-the-music-marketplace.pdf.

[541] COPYRIGHT LAW REVISION PART 6: SUPPLEMENTARY REPORT OF THE REGISTER OF COPYRIGHTS ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW: 1965 REVISION BILL 14, 35 (Comm. Print 1965).

[542] *See, e.g.*, S. REP. NO. 106-42, at 10 (1999) ("[T]he Committee is aware that in creating compulsory licenses, it is acting in derogation of the exclusive property rights granted by the Copyright Act to copyright holders, and that it therefore needs to act as narrowly as possible to minimize the effects of the Government's intrusion on the broader market in which the affected property rights and industries operate."); H.R. REP. NO. 100-887, pt. 1, at 15 (1988) ("Congress should impose a compulsory license only when the marketplace cannot suffice.").

[543] *See, e.g.*, A2IM-Recording Academy-RIAA Joint Reply Comments at 28 (noting "virtually no support from would-be licensees" and "broad opposition of both the tech sector and copyright owner groups" to compulsory licensing); ASCAP Initial Comments at 41–43; AMI Initial Comments at 6; Authors Guild Initial Comments at 27; Directors Guild Reply Comments at 3; Graphic Artists Guild Initial Comments at 15; Independent Film & Television Alliance Reply Comments at 8; Music Workers Alliance Initial Comments at 5; NSAI Initial Comments at 2–3; Recording Academy Initial Comments at 8; STM Initial Comments at 13; CCIA Initial Comments at 15; TechNet Initial Comments at 9–10. *But see, e.g.*, ImageRights International Reply Comments at 6 ("Establishing a compulsory licensing regime could be considered, but it should be carefully structured to respect the rights of creators and the diverse nature of works."); BigBear.ai Initial Comments at 18–19 (asserting that a compulsory license "is worthy of consideration").

[544] *See, e.g.*, AAP Initial Comments at 25–26; Getty Images Initial Comments at 21 ("[C]ompulsory . . . licensing schemes are not desirable when a marketplace for direct licensing already exists, which is the case with the licensing visual works and metadata to use in connection with the training and development of AI Models."); Jennifer Unruh Initial Comments at 5 ("Compulsory licensing of visual artwork would seriously undermine the expressive rights of the originating artists, including their right to not speak and to not make derivative works or reproductions."); MPA Initial Comments at 28–32 ("Market-based licensing for training AI models is feasible and preferrable to a compulsory licensing regime."); N/MA Initial Comments 53–56; SONA et al. Initial Comments at 5–6 ("[C]ompulsory licensing is stifling to a creator's livelihood and creativity."); UMG Initial Comments at 66–67.

[545] A2IM-RIAA Joint Initial Comments at 25.

disseminate a diverse array of creative works to the public."[546] And NMPA saw it as "an extreme remedy that deprives copyright owners of their right to contract freely in the market, and takes away their ability to choose whom they do business with, how their works are used, and how much they are paid."[547] Moreover, in the view of Authors Guild, "there is no indication that AI licensing markets have failed or are likely to do so."[548]

Commenters from the technology sector asserted that AI training is a noninfringing use and should not be subject to *any* licensing regime, whether voluntary or compulsory.[549] As with voluntary licensing, they argued that it is not logistically feasible[550] and would result in only meager royalty payments[551] due to the volume of works used. For example, a16z contended that a compulsory licensing scheme "would prove administratively impossible to implement" largely due to "scale," noting that "[f]or a very significant portion of those ["billions of pieces of text from millions of individual websites" used for training], it is essentially impossible to identify who the relevant rights holders are, and thus there would be no viable way to get statutory royalties to the proper parties."[552] Authors Alliance added that compulsory licensing

[546] Copyright Alliance Initial Comments at 80–82.

[547] NMPA Initial Comments at 24.

[548] Authors Guild Initial Comments at 27.

[549] *See* a16z Initial Comments at 9 ("[S]uch legislation would effectively require AI developers to remunerate rightsholders for a use that falls squarely within the protections of the fair use doctrine."); CCIA Initial Comments at 15 ("There is no principled basis for establishing [a compulsory licensing] regime. Just as a reader does not need to pay for learning from a book, an AI system should not have to pay for learning from content posted on a website."); Engine Initial Comments at 8 ("[S]tartups should not need licenses to train their AI models on copyrighted materials, both because that should be considered a noninfringing use under the law and, if it were to be considered a use, it would be protected by fair use.").

[550] *See, e.g.*, Meta Initial Comments at 20; TechNet Initial Comments at 9–10 ("[A]ny statutory licensing scheme would be impossible to administer."); Van Lindberg Initial Comments at 33–34 ("Compulsory licensing is not feasible given that the majority of AI training inputs are (and will likely continue to be) anonymous, pseudonymous, and unregistered.").

[551] *See, e.g.*, a16z Initial Comments at 8–11; Engine Initial Comments at 8; Meta Initial Comments at 20 ("[A]ny fair royalty due would be incredibly small in light of the insignificance of any one work among an AI training set."); TechNet Initial Comments at 9–10 (explaining that either "[a]ny licensing framework that provided any significant compensation to individual authors would impose a massive and insurmountable barrier to AI development, as it requires tens of billions of individual works—and, accordingly, tens of billions of individual royalty payments—to train an effective model," or "any statutory licensing scheme that imposed a less crippling financial obligation on the next generation of AI developers would mean that the resulting payments to individual authors would be miniscule," such that "[s]uch a scheme, with its attendant inefficiencies, neither benefits creators nor promotes the progress of science and the useful arts"); Van Lindberg Initial Comments at 33–34.

[552] a16z Initial Comments at 8–11.

is "logistically infeasible because of the scale and complexity of the training datasets needed to train AI models."[553]

Some cautioned that compulsory licensing is inflexible and "will not be able to keep up with the pace of development of generative AI, and may end up hurting both copyright holders and AI developers alike."[554]

2. Extended Collective Licensing

ECL is another approach, which has been adopted in some European countries in other contexts.[555] ECL typically involves a CMO being authorized to license all copyrighted works within a particular class of works for specific uses, binding all copyright owners in that class unless they opt out and choose to negotiate separately. This permits users to license numerous disparate works by copyright owners (including individual authors or small businesses) who have not affirmatively joined a CMO.

To obtain such authorization, the CMO usually must demonstrate that it represents a substantial number of copyright owners of works in that class and may also be required to satisfy other criteria. Unlike compulsory licenses, with rates and terms set by the government, the licenses issued by a CMO under an ECL system are negotiated with users in the free market. In this way, an ECL system functions like voluntary collective licensing, but with the government regulating the overall system and exercising some degree of oversight.[556]

---

[553] Authors Alliance Initial Comments at 14–18.

[554] ASCAP Initial Comments at 43 ("[Voluntary collective licensing] can better adapt to the evolving needs of copyright holders and AI developers."); *see also, e.g.*, David Newhoff Initial Comments at 2 ("Legislation of this nature is likely to be short-sighted and may lock in regimes that fail to serve authors."); NSAI Initial Comments at 7–8 ("[A]compulsory license envisioned today would be obsolete before it could even be implemented. Free market licensing is the only way to allow the music industry to keep pace with the rapid development of generative AI.").

[555] *See* Bingbin Lu, *The Orphan Works Copyright Issue: Suggestions for International Response*, 60 J. COPYRIGHT SOC'Y 255, 279–80 (2013).

[556] In its initial form, ECL covered only narrow types of works or uses, such as the use of published works for educational and scientific purposes, or the reproduction of works within an organization solely for internal use. U.S. COPYRIGHT OFFICE, LEGAL ISSUES IN MASS DIGITIZATION: A PRELIMINARY ANALYSIS AND DISCUSSION DOCUMENT 36 (2011). However, a few countries have adopted ECL programs with wider scopes. *See* U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION 83 (2015). The EU's 2019 Copyright in the Digital Single Market (DSM) Directive includes a provision permitting member states to authorize CMOs to provide ECLs subject to certain safeguards. Article 12 and its accompanying recitals emphasize that such an ECL should only be used "within well-defined areas of use, where obtaining authorisations from rightholders on an individual basis is typically onerous and impractical to a degree that makes the required licensing transaction unlikely." Art. 12, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Additionally, as mentioned above, in November 2024 the Spanish government issued a now-withdrawn decree seeking public comment on an ECL for AI model training on copyright-protected works. *See supra* text accompanying notes 463–464.

The ECL option received more support from commenters than a compulsory license, although views were mixed. Supporters generally envisioned ECL only for specific types of works, and not as a solution for all AI training.[557] Several suggested that ECL could be well-suited to the needs of visual artists.[558] Authors Guild proposed a twofold ECL system, distinguishing between past and future uses, and between professional creatives and other members of the public.[559]

Opposition came primarily from copyright owners who favored a purely voluntary licensing approach,[560] but also from commenters who opposed all licensing obligations.[561] Some viewed ECL as presenting similar concerns to compulsory licensing[562] or practically infeasible due to scale.[563] Others confined their opposition to the works in their own sectors on the grounds that a voluntary licensing market already exists.[564]

---

[557] *See, e.g.*, Copyright Alliance Initial Comments at 69–70 (stating that it would not oppose an ECL "if (i) there exists a general consensus of organizations and individual creators within a particular industry (for example, the book publishing industry) who are willing to accept 'opt outs' solely in the context of enacting an [ECL] provision; (ii) such provision is narrowly targeted to a particular industry and a particular type of work(s); and (iii) such license would not directly or indirectly affect (through inadvertent consequences or otherwise) those industries and works not intended to be covered by the license"); Daniel Gervais Initial Comments at 5; IT for Change Initial Comments at 6.

[558] *See* ASCRL Initial Comments at 5; Graphic Artists Guild Initial Comments at 15–16; ImageRights International Initial Comments at 6.

[559] Authors Guild Initial Comments at 22–27. For copyrighted works already used to train AI systems, "an ECL system would give rightsholders the opportunity to receive compensation for this prior unauthorized use," with those who decline to participate being able to opt out and "preserve their right to sue." *Id.* Because "technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained," the proposal targets compensation, transparency, and accountability for past uses, rather than permission. *Id.* at 22, 24. ECL would cover future uses of works which are not typically monetized, but "professional creators" would be represented only by existing CMOs and organizations best equipped to reach and represent those groups. *Id.* at 27. The proposal also discusses the need for enabling legislation and a robust authorization system to be managed by the Copyright Office.

[560] *See, e.g.*, Digital Media Licensing Ass'n Initial Comments at 13; News/Media Alliance Initial Comments at 53–54 n. 169; Recording Academy Initial Comments at 8; Scientific Technical Medical Publishers (STM) Initial Comments at 13; Dina LaPolt Initial Comments at 6; Rightsify Initial Comments at 8.

[561] *See* Authors Alliance Initial Comments at 15–18; CCIA Initial Comments at 11–15.

[562] *See, e.g.*, ASCAP Initial Comments at 43–44; AAP Initial Comments at 26 ("Because [ECL] also acts in derogation of the exclusive rights of copyright owners, it raises many of the same concerns as compulsory licensing"); MPA Initial Comments at 33 ("As with compulsory licensing, extended collective licensing also risks tipping the marketplace scales between copyright owners and those who exploit their works."); Music Reports Initial Comments at 5.

[563] *See, e.g.*, Authors Alliance Initial Comments at 15–18.

[564] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 26 (sound recordings); Getty Images Initial Comments at 21 (visual works).

3. Opting Out

A number of commenters addressed the possibility of a statutory "opt-out" mechanism, allowing copyright owners to signal the withholding of their works from AI training. Such an approach has been adopted in the EU as part of its text and data mining exception, as described above.[565]

Copyright owners rejected the idea of any opt-out approach. They asserted that it would be antithetical to current law,[566] unduly burdensome,[567] impossible to utilize after training occurs,[568] and difficult to implement.[569] News/Media Alliance stated that "existing law is 'opt in'" and that "[c]hanging this presumption under U.S. law would require the adoption of an additional exception under the law, a major undertaking that is not warranted under present circumstances.[570] And National Writers Union contended that "[a]n opt-out approach is not a feasible option for some creative workers and copyright owners," as "[t]ools like technical flags

---

[565] *See supra* Section IV.G.

[566] *See, e.g.*, A2IM-RIAA Joint Initial Comments at 21 ("The Copyright Act establishes an opt-in, permissions-based regime. . . . There is no basis in law or policy for imposing an opt-out regime."); DMLA Initial Comments at 11; Authors Guild Initial Comments at 22; CCC Initial Comments at 9; Getty Images Initial Comment at 17; UMG Initial Comments at 57; Artists Rights Society Reply Comments at 3–4.

[567] *See, e.g.*, Recording Academy Initial Comments at 6 ("[An opt out] shifts the burden of responsibility to the author, many of whom are at a stark disadvantage to handle such a responsibility. Opt-out approaches are a one-way path to creating an imbalance within the creative ecosystem between the haves and the have-nots."); CCC Initial Comments at 10 ("Placing the burden of asserting rights on the copyright holders [to opt out] is inequitable, burdensome, and largely impractical. Only those making copies know what they are copying in the first instance and thus the copyright owners are not in a position to opt out."); ImageRights International Reply Comments at 11 ("An opt-out system would impose an unreasonable burden on creators, obliging them to vigilantly track every instance where their work is used to train models, merely to exercise their right to opt out."); Graphic Artists Guild Initial Comments at 12; MPA Initial Comments at 26–27.

[568] *See, e.g.*, IT for Change Initial Comments at 5 ("[A]n opt-out measure merely guarantees that the relevant work will not be used in future training datasets. This provides no protection against existing AI models that have been trained on the relevant work."); Copyright Alliance Initial Comments at 70 ("To the best of our knowledge, technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained—though they might be coming. Some indicate that untraining models is challenging. Others indicate that it can be done, but it could be expensive. In the event an AI model cannot practically be retrained or a particular ingested work cannot practically be 'forgotten,' that serves as further evidence of why an opt-out system would not work since the harm caused to the copyright owner cannot be undone once the work has been ingested (and many of the biggest models in current use have already been built).") (citations omitted).

[569] *See, e.g.*, CEDRO Initial Comments at 10 ("In the case of the authorization (opt out), it poses more difficulties, for example, how to exercise it . . . where to exercise it, if the work is disseminated on the Internet, should it be exercised in all copies? As can be observed, it is difficult to determine."); Epidemic Sound Reply Comments at 2 ("We believe that it will be very difficult in practice to implement any such opt out system that is effective and not too burdensome for right holders. Also for these reasons, opt in systems are preferred."); MPA Initial Comments at 26.

[570] N/MA Initial Comments at 10.

and metadata can be prohibitive for those unfamiliar with digital technologies and people with impairments that impact their ability to utilize these tools."[571]

Commenters also discussed a variety of potential opt-out methods, such as using metadata, databases, watermarking, technical flags, and website terms of service.[572] While some in the technology sector identified certain approaches as "effective," "simple," or "ideal,"[573] many raised concerns, pointing to the ease with which metadata can be removed or the inability of copyright owners to use a platform-level flag, like robots.txt, if they do not control the platform.[574] Copyright Alliance further asserted that robots.txt "has significant limitations because it is only effective to the extent it is recognized and respected, and it was not designed to be targeted to scraping for generative AI ingestion."[575] Moreover, it said that robots.txt "would also prevent a search engine from scraping and categorizing the work," and that "[a] copyright owner may want their work to be scraped for search engine purposes—so they can be found on the internet—but not for AI ingestion."[576]

Those commenters with a positive view of opt outs said they could be beneficial to "support[ing] open development of generative AI datasets and pre-trained models by a broader range of actors," "foster[ing] international consistency with regimes such as the EU directive on Copyright in the Digital Single Market and proposed AI Act,"[577] and empowering creators to share their works freely without fear of objectionable use, while creating "a default of permissiveness that promotes an overall more open creative environment."[578] Several asserted

---

[571] National Writers Union Initial Comments at 14.

[572] *See, e.g.*, MPA Initial Comments at 26; ImageRights International Reply Comments at 5; BigBear AI Initial Comments at 17; Copyright Alliance Initial Comments at 70–71; Committee for Justice Initial Comments at 7; European Writers Council Initial Comments at 12.

[573] *See, e.g.*, Digimarc Initial Comments at 4; OpenAI Initial Comments at 10 ("OpenAI has implemented an easy means for websites to exclude their content from being accessed by OpenAI's 'GPTBot' web crawler. This simple opt-out mechanism is built on the well-established robots.txt standard that has been used for nearly 30 years. Adoption metrics suggest that this option is now well known and has been broadly embraced."); CCIA Initial Comments at 12 ("[A]n enhanced robots.txt would be an ideal way to achieve [opt-outs] for Web data.").

[574] *See, e.g.*, Copyright Alliance Initial Comments at 71; Microsoft Initial Comments at 9; Digimarc Initial Comments at 3; UMG Initial Comments at 59; Getty Images Reply Comments at 11; MPA Initial Comments at 26–27; Association of Medical Illustrators Initial Comments at 3.

[575] Copyright Alliance Initial Comments at 71.

[576] *Id.*

[577] Hugging Face Initial Comments at 2.

[578] Public Knowledge Initial Comments at 9.

that voluntary measures adopted by AI companies allowing copyright owners to opt out of training have merit, but did not advocate for an opt-out system to be established by law.[579]

C. Analysis and Recommendations

In assessing any form of licensing, it is important to recognize the wide variations in works and uses involved in AI training. Feasibility will depend on the types of works needed, the licensing practices of the relevant industries, the design of the AI system, and its intended uses. For instance, licensing a music model that can produce rudimentary jingles is different from licensing a state-of-the-art LLM that can compete on advanced reasoning benchmarks. And sophisticated commercial entities will be easier to find and negotiate with than individual non-professionals.

As discussed above, a number of voluntary direct and collective licensing agreements for using copyrighted works in AI training have emerged over the past several years, with others in development.[580] Some AI systems have now been trained exclusively on licensed or public domain works.[581] These developments demonstrate that voluntary licensing may be workable, at least in certain contexts—particularly where training is focused on valuable content that can be licensed in relatively high volumes (*e.g.*, popular music and stock photography), or in fields where the number of copyright owners is limited. The Office recognizes, however, that practical challenges remain in many areas. The growing licensing market does not itself establish that voluntary licensing is feasible at scale for all AI training needs. To the extent that the remaining gaps cannot reasonably be filled, alternative solutions may be needed.

As to compensation, further market developments may provide more insight on the extent to which licensing agreements can effectively compensate copyright owners for the use of their works in AI training. The agreements that already exist indicate that mutually agreeable compensation terms can be negotiated in some situations, although it remains to be seen how they scale. Compensation structures based on a percentage of revenue or profits, without large up-front cash outlays, may be an attractive alternative for smaller developers looking to enter the market. As to concerns voiced by commenters about the affordability for academic researchers, we note that the research projects they identify may well qualify as fair use and therefore would not require licenses.[582] And the amount of monetary compensation

---

[579] *See, e.g.*, Internet Archive Initial Comments at 9; BSA Initial Comments at 9 ("We support further voluntary conversations between creators and AI developers and deployers to arrive at effective, consensus technical mechanisms.").

[580] *See supra* Section IV.D.3.

[581] *See supra* text accompanying notes 502–506.

[582] *See supra* Sections IV.A.2.c; IV.A.3.

that some copyright owners will accept may depend on contractual conditions regarding control of the use of their works.

As discussed above, there appears to be strong interest among those representing copyright owners and creators in developing voluntary collective licensing for the AI context.[583] Collective licensing can play a significant role in facilitating AI training, reducing what might otherwise be thousands or even millions of transactions to a manageable number. The aggregation of rights could be mutually beneficial, such as where transaction costs might otherwise exceed the value of using a work or where copyright owners might be difficult to find. Although collective licensing presents its own logistical and organizational challenges, it affords copyright owners and licensees flexibility to tailor agreements to their needs. Multiple CMOs can each license different types of copyrighted works on terms that make sense for that particular creative industry and AI model.

As to antitrust concerns, courts have found that there is nothing intrinsically anticompetitive about the collective, or even blanket, licensing of copyrighted works, as long as certain safeguards are incorporated—such as ensuring that licensees can still obtain direct licenses from copyright owners as an alternative.[584] Although antitrust law is beyond the scope of the Office's expertise, we believe that greater clarity would be valuable. We encourage the Department of Justice to provide guidance, including on the benefit of an antitrust exemption in this context.

We agree with commenters that a compulsory licensing regime for AI training would have significant disadvantages. A compulsory license establishes fixed royalty rates and terms and can set practices in stone; they can become inextricably embedded in an industry and become difficult to undo.[585] Premature adoption also risks stifling the development of flexible

---

[583] *See supra* Sections IV.D.3; V.A.1.

[584] *See Broadcast Music Inc. v. CBS Inc.*, 441 U.S. 1, 18–25 (1979) (finding that a CMO offering a blanket license to perform all of the musical works in its catalog was not a per se antitrust violation); *Columbia Broadcasting System Inc. v. ASCAP*, 620 F.2d 930, 935–39 (2d Cir. 1980) (upholding the same blanket license under antitrust law's "rule of reason," explaining that it did not unreasonably restrain competition because licensees could still feasibly obtain direct licenses from copyright owners); *see also* U.S. DEP'T OF JUST. & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR THE LICENSING OF INTELL. PROP. 30 (2017) (explaining that while "pooling arrangements can have anticompetitive effects in certain circumstances," they "are often procompetitive").

[585] For example, during the process that led to the 1976 Act, "it became apparent that record producers, small and large alike, regard the [predecessor to section 115] as too important to their industry to accept its outright elimination," and "while still opposing the provision in principle, some copyright owners implied that ultimately there might be advantages in ameliorating the harsh and burdensome effects of the compulsory license rather than doing away with it altogether." COPYRIGHT LAW REVISION PART 6: SUPPLEMENTARY REPORT OF THE REGISTER OF COPYRIGHTS ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW: 1965 REVISION BILL 53–54 (Comm. Print 1965) (observing that the predecessor to section 115 "had a profound effect upon the development of the American record industry, and that many of the present practices in the industry are directly related to [it]").

and creative market-based solutions. Moreover, compulsory licenses can take years to develop, often requiring painstaking negotiation of numerous operational details.[586]

For those sectors where voluntary licensing may prove unworkable or infeasible, ECL would be a less intrusive approach. It would permit copyright owners to choose to license separately, while enabling full coverage of the entire sector for AI training. Allowing authorized CMOs to negotiate rates and terms and establish policies and procedures, subject to government oversight would provide flexibility, rather than freezing rates in the statute or setting them through judicial or administrative proceedings.[587]

As to the possibility of an opt-out mechanism, the Office agrees that requiring copyright owners to opt out is inconsistent with the basic principle that consent is required for uses within the scope of their statutory rights. But to the extent that Congress may consider an exception or limitation for AI training in the future, the ability to opt out could preserve some ability to block unwanted uses or negotiate terms. Nevertheless, significant concerns have been raised about the effectiveness and availability of opt-outs , which would need to be addressed.[588]

Finally, we note that the law, technology, and markets for training are relatively nascent, and there is a dynamic interplay between them. To begin with, the current licensing market may be distorted by the unsettled legal questions about fair use. While some AI companies may have licensed works for training to avoid uncertainty or obtain access to high-quality or otherwise-unavailable materials, other licensing activities may be inhibited by reliance on fair use. As courts begin to resolve pending cases, greater legal clarity may lead to greater collaboration on technical and market-based solutions. Similarly, new model architectures and techniques may be developed to facilitate training using fewer unlicensed works without sacrificing quality. Whether companies devote resources toward such solutions may in turn be influenced by the shifting incentives created by legal and licensing developments.

---

[586] While few commenters discussed such details, questions to be addressed would include: What should be the scope of the license? Who should be eligible to obtain the license? What should the royalty rate be or how should rates be set going forward? How should works be valued and royalties allocated? How should royalties be collected and distributed? Should new CMOs be established by law or existing ones designated to administer the compulsory license? What kind of reporting should be required of licensees?

[587] In similar circumstances involving numerous or difficult to locate copyright owners, the Office has in the past suggested an ECL solution. In a 2015 report, we concluded that ECL "is the best answer to solving the mass licensing that is inherent to mass digitization." U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION 83 (2015). *See also* Letter from Karyn Temple, Acting Register of Copyrights and Director, U.S. Copyright Office, to Senators Grassley and Feinstein (Sept. 29, 2017).

[588] An opt-out mechanism should be simple and straightforward enough that individual copyright owners lacking legal or technological expertise can use it. A system-by-system or company-by-company opt-out would be burdensome to monitor and implement. Nor should it be sufficient for AI companies to merely honor platform-level flags, like robots.txt, because in many cases copyright owners have no control over the platforms where their works appear—whether a legitimate or a pirate site. At the same time, the mechanism must also be reasonable for AI companies (including small startups) to operationalize. Unless they can ascertain which works are subject to an opt out at any given point in time, the system will be ineffective.

In light of the foregoing, at this point in time, the Office recommends allowing the licensing market to continue to develop without government intervention. If market failures are shown as to specific types of works in specific contexts, targeted intervention such as ECL should be considered.

VI. CONCLUSION

Throughout its history, copyright law has adapted to new technology, furthering its progress while preserving incentives for creative activity. This has enabled our nation's creative and technology industries to become global leaders in their fields. While the use of copyrighted works to power current generative AI systems may be unprecedented in scope and scale, the existing legal framework can address it as in prior technological revolutions. The fair use doctrine in particular has served to flexibly accommodate such change. We believe it can do so here as well.

In applying current law, we conclude that several stages in the development of generative AI involve using copyrighted works in ways that implicate the owners' exclusive rights. The key question, as most commenters agreed, is whether those acts of prima facie infringement can be excused as fair use.

The fair use determination requires balancing multiple statutory factors in light of all relevant circumstances. Although it is not possible to prejudge the result in any particular case, precedent supports the following general observations:

Various uses of copyrighted works in AI training are likely to be transformative. The extent to which they are fair, however, will depend on what works were used, from what source, for what purpose, and with what controls on the outputs—all of which can affect the market. When a model is deployed for purposes such as analysis or research—the types of uses that are critical to international competitiveness—the outputs are unlikely to substitute for expressive works used in training. But 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.

For those uses that may not qualify as fair, practical solutions are critical to support ongoing innovation. Licensing agreements for AI training, both individual and collective, are fast emerging in certain sectors, although their availability so far is inconsistent. Given the robust growth of voluntary licensing, as well as the lack of stakeholder support for any statutory change, the Office believes government intervention would be premature at this time. Rather, licensing markets should continue to develop, extending early successes into more contexts as soon as possible. In those areas where remaining gaps are unlikely to be filled, alternative approaches such as extended collective licensing should be considered to address any market failure.

In our view, American leadership in the AI space would best be furthered by supporting both of these world-class industries that contribute so much to our economic and cultural advancement. Effective licensing options can ensure that innovation continues to advance without undermining intellectual property rights. These groundbreaking technologies should benefit both the innovators who design them and the creators whose content fuels them, as well as the general public.

Finally, as in prior Parts of this Report, the Office recognizes that facts on the ground are evolving at a rapid pace. We will continue to monitor developments in technology, case law, and markets, and to offer further assistance to Congress as it considers these issues.