NEWS MEDIA ALLIANCE

Before the
U.S. COPYRIGHT OFFICE

Artificial Intelligence and Copyright

Docket No. 2023–6

Submitted October 30, 2023

<u>COMMENTS OF THE NEWS/MEDIA ALLIANCE</u>

Introduction.

The News/Media Alliance ("N/MA") welcomes the opportunity to submit comments in response to the notice of inquiry regarding the U.S. Copyright Office's study of the copyright law and policy issues raised by generative artificial intelligence (AI). The last few years have witnessed the rise of AI systems and applications that have the potential to greatly reshape the digital marketplace and alter many features of public life.

This is particularly true of generative AI ("GAI") technologies,[1] including the introduction of large language models ("LLMs") and related applications (such as chatbot interfaces) to consumers and the digital marketplace, the focus of our comment. N/MA members would welcome working with generative AI developers to help build and grow these technologies, in ways that benefit all actors in the supply chain and society at large. News and media publishers recognize the potential opportunities for users, businesses, and society alike, and many

[1] Our comment adopts the Copyright Office's understanding of generative AI technologies as "capable of producing outputs such as text, images, video, or audio (including emulating a human voice) that would be considered copyrightable if created by a human author" based on "'learning' statistical patterns in existing data, which may include copyrighted works." U.S. COPYRIGHT OFFICE, ARTIFICIAL INTELLIGENCE AND COPYRIGHT, 59942 Fed. Reg. 88 (167), (Aug. 30. 2023) available at https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf.

members are exploring how to responsibly utilize generative AI technologies in their workstreams.

But to fulfill their societal potential, technological innovations must be advanced in a sustainable manner. Not only are generative AI models often trained on copyright-protected, professionally created material, many applications also act as direct competitors to publishers, providing informational and cultural content to the public, and drawing readers and advertisers away from publisher websites. In effect, publishers invest in producing high-quality content that is taken without permission to train the AI systems and used to produce substitutional, expressive AI-generated ”outputs“ that then compete directly with publisher content, reducing publisher revenues and employment, tarnishing their brands, and undermining their relationships with readers. The continued unlicensed use of journalistic reporting portends injury to the public interest that it serves, and may hinder the progress of generative AI innovations.

N/MA is grateful to the Copyright Office for undertaking this important and timely study and facilitating dialogue among the stakeholders and policymakers. As President Biden’s Executive Order issued on October 30, 2023, recognizes, mitigating against risks posed by AI is vital in order to realize its potential for society.[2] While AI is exciting, and N/MA supports the principled development of generative AI technologies, unregulated, it also poses a significant threat to the pillars of a healthy and informed democracy. Our members are gravely concerned that some developers have to date violated the legal rights of publishers, using their copyrighted material without permission or compensation and tarnishing their brands. Copyright law simply does not require publishers to train their replacements in this way.

N/MA has vigorously advocated for its members’ interests on issues surrounding generative AI to advance our members’ interests and to address risks that unsustainably deployed generative AI poses to the continued viability of the news business. In 2020, N/MA filed comments with the United States Patent and Trademark Office focusing on the issue of systemic ingestion of copyright protected content for machine learning purposes.[3] These comments, attached here,[4] discussed how the current case law provides protections for media content against such use

[2] *See* The White House, Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (October 30, 2023), available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[3] News/Media Alliance, Re: Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation (Jan. 10, 2022) available at http://www.newsmediaalliance.org/wp-content/uploads/2020/01/News-Media-Alliance-AI-Comments-with-USPTO.pdf.

[4] See Appendix B.

and highlighted the need for stronger enforcement. More recently, N/MA published a set of AI principles covering issues related to intellectual property, transparency, accountability, fairness, safety, and design, that we hope will inform AI policy development in the United States.[5] We also joined a similar set of global principles, together with 27 other publisher organizations.[6]

Now, N/MA contemporaneously publishes a White Paper, also attached here[7] and referenced below, on AI developers' pervasive use of publisher content in generative AI training. The White Paper includes a technical analysis regarding the use of publisher content in generative AI applications and discusses the effects and legal implications of such use. A few takeaways from that analysis include:

- Developers have copied and used news, magazine and digital media content to train LLMs.

- Popular curated datasets underlying LLMs significantly overweight publisher content by a factor ranging from over 5 to almost 100 as compared to the generic collection of content that the well-known entity Common Crawl has scraped from the web.

- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's generative AI-powered products like Bard. Half of the top ten sites represented in the data set are news outlets.

- LLMs also copy and use publisher content in their outputs. LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

In short, generative AI systems should be held responsible and accountable, just like any other business. The risks of unregulated AI development and use are too high, both for society and a competitive online economy alike. N/MA hopes that the Office's study will bring attention to the systemic and wide ranging infringement by some generative AI developers and help grow emerging practices for licensed use of publisher content.

[5] News/Media Alliance, AI Principles (2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/04/FINAL-UPDATED-AI-Principles_4-20-23.pdf.
[6] Global Principles on Artificial Intelligence (AI) (2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/09/FINAL-Global-AI-Principles-Formatted_9-5-23.pdf.
[7] *See* N/MA, White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems is Copyright Infringement and Not a Fair Use (2023) [hereinafter N/MA, White Paper], Appendix A.

The digital ecosystem will benefit from consensus around protection and partnership. N/MA's members are by and large willing to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy and authoritative expressive content. A constructive solution could benefit all interested parties and society at large, and avoid protracted litigation. And fruitful cooperation will also help maximize the potential of generative AI technologies, by helping ensure they are developed using high-quality and human created works.

The Office's inquiry can also help inform the development of carefully considered and well-balanced AI policy at the federal level to mitigate against unintended consequences and harms to the media and other creative industries. We look forward to engaging with the Office, the Congress, and the Administration moving forward.

About News/Media Alliance.

N/MA is a nonprofit organization headquartered in Washington, D.C., representing the newspaper, magazine, and digital media industries, and empowering members to succeed in today's fast-moving media environment. N/MA represents over 2,200 diverse publishers in the United States and internationally, ranging from the largest news and magazine publishers to small, hyperlocal newspapers, and from digital-only and digital-first outlets to print papers and magazines.

In total, N/MA's membership accounts for nearly 90 percent of the daily newspaper circulation in the United States, nearly 100 magazine media companies with over 500 individual magazine brands, and dozens of digital-only properties. Its members publish high-quality original content on topics ranging from news to culture, sports, entertainment, lifestyle, and virtually any other interest. N/MA diligently advocates for its members on a broad range of current issues affecting them, including copyright policy that directly relates to our members' ability to monetize their content and support their continued investments in high-quality content production.

N/MA members play a vital role in their communities and in fostering an informed public and the public trust necessary for democracy. Publishers invest considerable time and resources to produce journalism and original creative content that combats misinformation, encourages democratic engagement, strengthens community ties, lowers municipal borrowing costs, safeguards consumers, keeps decision makers accountable, gives people something to talk about, and supports the free flow of ideas and information.[8] Our members also support local

[8] *See, e.g.*, Matthew Gentzkow, et al., *The Effects of Newspaper Entry and Exit on Electoral Politics*, 101 AM. ECON. REV. 2980 (2011); Danny Hayes & Jennifer L. Lawless, *As Local News Goes, So Goes Citizen Engagement: Media,*

economies by providing small and medium enterprises, local businesses, and community organizations with a cost-effective way to reach potential local customers through advertising and online content. However, despite these considerable benefits, and the increased audience for publisher content, far too many publishers are struggling to survive in the online ecosystem, partially due to the unauthorized scraping and use of their protected content.

The news, magazine, and digital media industries' contribution to the U.S. economy and society is considerable, with estimated revenues of newspaper and magazine publishers amounting to approximately $45 billion.[9] Newspaper newsrooms were estimated to directly employ approximately 31,000 people in 2020, not including additional indirect employment effects, while magazines employed over 73,000 directly and supported a total of over 219,000 jobs in 2021.[10] Employment in digital-native newsrooms, meanwhile, has increased from approximately 7,400 in 2008 to over 18,000 in 2020.[11] The content produced by these professionals has a huge audience, with N/MA member publishers reaching hundreds of millions of Americans every year. The share of digital audience is large for both magazine and newspaper publishers, with news publishers having over 200 million unique visits and 6.7 billion page views per month while 40 percent of magazine readers access content on mobile

Knowledge, and Participation in U.S. House Elections, 77 Journal of Politics 447 (2014); Mary Ellen Klas, *Less Local News Means Less Democracy*, Nieman Reports, Sep. 20, 2019, https://niemanreports.org/articles/less-local-news-means-less-democracy/; The Covington News, *The Benefits of Local Newspapers*, [n.d.] https://www.covnews.com/nie/benefits-local-newspapers/; Pengjie Gao, Chang Lee, & Durmot Murphy, *Financing Dies in Darkness? The Impact of Newspaper Closures on Public Finance*, 135 Journal of Financial Economics 2 (2020); The British Psychological Society, *Why Magazines Matter,* The Psychologist, Nov. 25, 2016, https://www.bps.org.uk/psychologist/why-magazines-matter.

[9] *See* PEW Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023); Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, available at https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/ (last visited Nov. 17, 2022); Adam Grundy, *Service Annual Survey Shows Continuing Decline in Print Publishing Revenue*, U.S. Census Bureau, Jun. 7, 2022, available at https://www.census.gov/library/stories/2022/06/internet-crushes-traditional-media.html.

[10] PEW Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023); Mason Walker, *U.S. Newsroom Employment Has Fallen 26% since 2008*, Pew Research Center, Jul. 13, 2021, https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; MPA – The Association of Magazine Media, Magazine Media Factbook, (2021) available at https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[11] PEW Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023).

devices.[12] This is in addition to the millions who access content on digital-only publishers' websites.

In order to continue investments into high-quality journalism and digital content, publishers require strong intellectual property protections and a vibrant, open, and fair online competitive environment that, when functioning at its best, rewards quality, creation, and innovation. Today, a few dominant online platforms control the digital ad ecosystem and the distribution of digital content, posing an existential threat to many publishers, especially small and local newspapers. The numbers in the preceding paragraph take on a different meaning when you consider that in less than 20 years, newspaper circulation and advertising revenues dropped from $57.4 billion in 2003 to an estimated $20.6 billion in 2020, while magazines witnessed a drop from $46 billion in 2007 to $23.92 billion in 2020.[13] In short, news publishers' revenues decreased by almost two-thirds and magazines have lost almost half of their revenues. In total, 2,500 newspapers have either closed or merged since 2004.[14] Similarly, there has been a substantial loss of community newspapers such that at least 200 counties, representing four million Americans, no longer have a local newspaper.[15] These losses are more likely to affect already disenfranchised people and communities, with many of the lost or failing newspapers located in areas that are less affluent than the national average. While magazine publishers have generally fared somewhat better than newspaper publishers, many have been forced to reduce print days or cut print editions completely, in an effort to lower costs.[16] Together, these trends have led to substantial job losses across the publishing industry.[17]

[12] News/Media Alliance, News Advertising Panorama (2020) (publicly available to N/MA members only; on file with author); MPA – The Association of Magazine Media, Magazine Media Factbook, (2021) available at https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[13] PEW Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023)); Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, available at https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/ (last visited Oct. 13, 2023).

[14] Penny Abernathy, Report: The State of Local News 2022 (2022), available at https://localnewsinitiative.northwestern.edu/projects/state-of-local-news.

[15] *Id.*

[16] *See* Beth Braverman, *How Magazine Publishers Are Cutting Print Costs to Improve Profits*, Folio Magazine, Aug. 2, 2021, https://archive.foliomag.com/magazine-publishers-cutting-print-costs-improve-profits/; Peter Houston, *2021 in Print: Newspapers' Decline Continues, But for Magazines … It's Complicated*, What's New in Publishing, Dec. 20, 2021, https://whatsnewinpublishing.com/2021-in-print-newspapers-decline-continues-but-for-magazines-its-complicated/.

[17] *See* Mason Walker, *U.S. Newsroom Employment Has Fallen 26% Since 2008*, Pew Research Center Jul. 13, 2021, available at https://www.pewresearch.org/fact-tank/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; Bureau Of Labor Statistics, Occupational Outlook Handbook: Reporters, Correspondents, And News Analysts, [n.d.] available at https://www.bls.gov/ooh/media-andcommunication/reporters-correspondents-and-broadcast-news-analysts.htm (last visited Nov. 17, 2022).

General Questions

Our responses to the Office's specific questions are below. N/MA may submit supplemental comments in response to other questions raised by the Office, or by other commenters.

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

N/MA recognizes the potential benefits and is broadly supportive of AI applications and technologies, with many of our members using AI—including generative AI—in various ways throughout their business operations. These uses may include content ideation and research, content optimization, improving internal efficiency, and content review and distribution activities.[18] Generative AI applications can provide an important tool in newsgathering and research efforts by determining sources for research and interviews, identifying content opportunities and aggregating and synthesizing information. Publishers may also use generative AI systems to generate headlines, outline articles, and write first drafts and FAQs subject to human review—to mention a few examples—while also utilizing AI to improve search engine optimization (SEO). Journalists and authors may benefit from generative AI in activities ranging from proofreading to distribution through social media channels and newsletters.

To be sustainable, however, generative AI development and use must be responsible, regulated, and accountable, with appropriate permission and compensation paid to publishers for the copying and use of their protected works throughout the product cycle. Without effective enforcement, regulation, and standards—including a requirement for AI developers to seek permission from rightsholders for uses of their protected content to train competitive products—AI can lead to considerable harms. These harms may include the undermining of the foundation of our democracy through the further weakening or outright closure of newspapers, magazines, and digital outlets—especially local ones—the spread of mis- and disinformation, and reduced access to reporting that can fundamentally only be created by humans—based on extensive fact-gathering, interviews, and judgment. An engaged and informed citizenry depends on the existence and availability of reliable and accurate reporting and analysis by outlets the public trusts. Unlike generative AI systems that may make up facts and disclaim

[18] *See, e.g.,* Elite Truong, *Local News and AI,* AM. PRESS INST., August 7, 2023, https://americanpressinstitute.org/publications/articles/local-news-and-ai/.

liability for doing so,[19] publishers accept responsibility for the content they publish, ensuring that the information presented to the public is of high quality. In a world flooded by easily accessible, synthetic information of unknown quality, real information becomes harder to identify and trust in our democratic system harder to upkeep.

In addition to these significant societal harms, the negative effects of unsustainable AI development practices on publishers small and large can lead to substantial job losses and a devaluing of journalistic content that will undermine these creative industries. In short, while AI presents many potential benefits to both publishers and the public at large, unregulated generative AI risks driving existing publishers out of business and disincentivizing continued investments in new, original content. This result would undermine the goal and purpose of the Copyright Clause of the Constitution, and diminish the essential role of the Press envisioned by the Founders. (And potentially also harming the further development of generative AI models through model collapse, as discussed further below.)

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

The increasing use and distribution of generative AI systems and applications, as well as AI-generated materials, raises substantial, unique concerns for newspaper, magazine, and digital media publishers. While the interests of publishers and generative AI developers could align, for example, in a fair exchange of licensing revenues for access to high-quality training materials to facilitate the continued improvement of the models, the promise of partnership has not yet materialized except in a few narrow instances.[20] Instead of entering into legal licensing agreements with publishers, generative AI developers have chosen to scrape publisher content without permission and use it for model training and in real-time (grounding)[21] to produce outputs (often in the form of lengthy, expressive summaries) that can directly compete with publisher content and products. And they literally are making billions doing it.[22] Not only can

[19] For example, OpenAi has taken the position in litigation that it is not liable for claims for defamation. "Because any ChatGPT user verifies at signup that they "take ultimate responsibility for the content being published," OpenAI says that, "as a matter of law, this creation of draft content for the user's internal benefit is not 'publication.'" Ashley Belanger, *Will ChatGPT's hallucinations be allowed to ruin your life?*, ARS TECHNICA, Oct. 23, 2023, https://arstechnica-com.cdn.ampproject.org/c/s/arstechnica.com/tech-policy/2023/10/will-chatgpts-hallucinations-be-allowed-to-ruin-your-life/amp/.

[20] See discussion on existing licensing deals below.

[21] *See* N/MA, WHITE PAPER at 17-18 (2023), Appendix A; Jordi Ribas, *Building the New Bing*, MICROSOFT BING BLOGS, Feb. 21, 2023, https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing#:~:text=Selecting%20the%20relevant%20internal%20queries,this%20method%20is%20called%20grounding.

[22] *See, e.g.*, Jagmeet Singh & Ingrid Lunden, *OpenAI Closes $300M Share Sale at $27B-29B Valuation,* TECHCRUNCH (Apr. 28, 2023) https://techcrunch.com/2023/04/28/openai-funding-valuation-chatgpt/; Deepa Seetharaman & Berber Jin, *OpenAI Seeks New Valuation of Up to $90 Billion in Share Sale*, WALL ST. J. (Sep. 26, 2023)

generative AI systems and applications respond to user queries using publisher content but, as discussed in more detail below, an AI chatbot or search interface can, and does, produce outputs that include verbatim quotes and/or closely paraphrases publisher stories.

The members of the News/Media Alliance are deeply concerned about this unauthorized and unlawful use of their expressive content by large technology companies that do not shoulder the cost of reporting the news or producing creative content, but who capitalize on the results of that valuable work. Copyright law does not require publishers to train their replacements in this manner. In effect, publishers make the investments and take the risks—including sending journalists into harm's way—while generative AI developers reap the rewards of users, data, brand creation, subscription fees, and advertising dollars. This is freeriding.

The continued unlicensed use of reporting—including entire corpora of unique publisher content, amounting up to millions of stories—portends injury, not just to the news industry, but to the public interest that it serves: an online world that is dominated by AI-generated, inferior yet substitutional content will leave the public with watered-down, less reliable outputs and fewer news outlets with the resources necessary to provide critical original reporting. As district court judge Denise Cote's decision in *Associated Press v. Meltwater U.S. Holdings, Inc.* explained with respect to direct scraping of news content, copyright law does not allow for democracy to be imperiled in this manner:

> *[T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression … Permitting [Meltwater] to take the fruit of [AP's] labor for its own profit, without compensating [AP], injures [AP's] ability to perform [its] essential function of democracy.*[23]

In addition to decreasing readership, the unauthorized use of publisher content to produce outputs that include inaccuracies also devalues publisher brands and creative content by muddling the source of the original content and misattributing information or misinformation to unrelated publishers or journalists.[24] This is especially damaging as many of N/MA's

https://www.msn.com/en-us/money/companies/openai-seeks-new-valuation-of-up-to-90-billion-in-share-sale/ar-AA1hiJ9W.

[23] *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013).

[24] Julia Black @mjnblack, X (Apr. 4, 2023, 7:48), https://twitter.com/mjnblack/status/1643324719108706304; Kate Crawford @katecrawford, X (Apr. 4, 2023, 19:42), https://twitter.com/katecrawford/status/1643323086450700288 (Journalist doing background research on an interview subject using ChatGPT was provided with cites and links to two non-existent articles critical of the subject, one by MIT Technology Review.); Chris Moran, *ChatGPT is Making up Fake Guardian Articles. Here's How We're Responding*, GUARDIAN, Apr. 6, 2023, https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article; James Warrington, *AI is 'Polluting the Pool of Human Knowledge', News Publishers Warn*, THE TELEGRAPH, Oct. 1,

members have spent years or decades—sometimes even centuries—building their reputation as reliable and trustworthy content producers, providers, and curators. This reputation is vital for their success, with readers associating their brands with content that has been researched, vetted, proofread, and carefully considered by consummate professionals they know and trust. Indeed, trusted journalism can be an antidote to the mis- and disinformation problem.[25]

It is therefore particularly concerning when a generative AI system attributes material that is blatantly false to a publisher who has never published such information. As one example, take the case of Jonathan Turley, a law professor who ChatGPT falsely accused of sexually harassing a student, attributing the information to a non-existent news article by The Washington Post.[26] In the same research experiment, conducted by Professor Eugene Volokh, ChatGPT made other similarly false allegations, citing articles that did not exist from publishers such as the Miami Herald and the Los Angeles Times. These "hallucinations,'' or massive errors, are a recognized propensity of many generative AI models that can spread misinformation and cause real harm to publisher brands. Other examples of the dangers of "hallucinations'' and other harms include summaries of articles by reputable publishers combining information from unreputable sources and the proliferation of deepfake photographs in politics.[27] Publishers recognize these pitfalls and while some may use AI as a tool in newsgathering and content production processes, they accept legal responsibility for the content they publish and understand that the outputs are often not reliable and require human editing and supervision before publication—something that generative AI systems typically do not have.

To mitigate these risks, it is essential that generative AI training datasets, systems, and applications be based on reliable, trustworthy, and high-quality content with adequate safeguards to deter misinterpretations and the creation of false information based on that content. To do so sustainably and lawfully—in a manner that protects the public interest, including professional journalism—generative AI developers should license content from publishers for training and grounding purposes based on fair and transparent negotiations, as

2023, https://www.telegraph.co.uk/business/2023/10/01/news-publishers-warn-ai-will-pollute-human-knowledge/.

[25] Jeff Clune, *AI-enabled Scams Will Proliferate*, Macleans, Oct. 12, 2023, https://macleans.ca/society/technology/ai-scams/. ("As we prepare for AI scams to proliferate, the best advice I can offer is for people to seek out and hold onto the sources they trust most-—whether that is the New York Times or a particular reporter. But even then they must make sure they are in fact getting information from that source.").

[26] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, Wash. Post, April 5, 2023, https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.

[27] Karen Weise & Cade Metz, *When A.I. Chatbots Hallucinate*, The Telegraph, Oct. 1, 2023, https://www.telegraph.co.uk/business/2023/10/01/news-publishers-warn-ai-will-pollute-human-knowledge/; William A. Galston, *Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics*, Bookings, Jan. 8, 2020, https://www.brookings.edu/articles/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

discussed in more detail below and in the attached White Paper. Only then can publishers recoup their investments in high-quality journalism while assuring developers that generative AI systems are built on authentic—not synthetic—content that is essential for reliable and trustworthy AI.

In the worst-case scenario, without an enforceable licensing market, high-quality publishers will slowly fail, forcing generative AI systems to rely on each other for training, leading to the gradual degradation in the availability of reliable and trustworthy reporting to our communities and system of democratic governance.[28] In fact, without human-generated quality content to train AI, researchers have found "that use of model-generated content in training causes irreversible defects in the resulting models," an effect they term "model collapse"[29] or "Model Autophagy Disorder (MAD),"[30] an analogy to mad cow disease:

> For instance, start with a language model trained on human-produced data. Use the model to generate some AI output. Then use that output to train a new instance of the model and use the resulting output to train a third version, and so forth. With each iteration, errors build atop one another. The 10th model, prompted to write about historical English architecture, spews out gibberish about jackrabbits.[31]

It is therefore in all of our collective interest that generative AI companies adhere with the letter and spirit of intellectual property law.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

The following papers and studies may help the Office in identifying the most pressing issues and concerns related to the proliferation of generative AI systems and applications and in identifying constructive solutions for continued success and innovation for all stakeholders:

[28] *Cf.* "Cory Doctorow, *The 'Enshittification' Of Tiktok*, WIRED, Jan. 23, 2023, https://www.wired.com/story/tiktok-platforms-cory-doctorow/.

[29] Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV, May 27, 2023, available at https://arxiv.org/abs/2305.17493.

[30] Sina Alemohammad, et al., *Self-Consuming Generative Models Go MAD*, ARXIV, Jul. 4, 2023, available at https://arxiv.org/abs/2307.01850.

[31] Rahul Rao, *AI-Generated Data Can Poison Future AI Models*, SCIENTIFIC AMERICAN, Jul. 28, 2023, available at https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/.

- News/Media Alliance's White Paper on AI and Copyright, outlining how generative AI developers use publisher content, how it is stored and reproduced, the effects on publishers, and the legal implications of such use. The Paper incorporates a technical study analyzing the issues discussed (attached as Appendix A);

- News/Media Alliance's AI Principles that spell out publisher concerns and set out principles that should guide policy development in order to protect the sustainability of high-quality content online: https://www.newsmediaalliance.org/ai-principles/;

- Global AI Principles signed by 28 publisher organizations across the world, outlining principles that should guide AI policy development both at domestic and international fora: https://www.newsmediaalliance.org/global-principles-on-artificial-intelligence-ai/;

- Copyright Alliance's AI Position Paper that includes high-level discussion of the concerns and interplay of AI and the creative industries: https://copyrightalliance.org/policy/position-papers/artificial-intelligence/;

- The United Kingdom's House of Lords report on AI, outlining benefits and risks of AI as well as relevant policy discussions, including concerning the right of copyright owners to decide when their content is used for text and data mining: https://lordslibrary.parliament.uk/artificial-intelligence-development-risks-and-regulation/;

- A study on the potential for model collapse, noting that to "make sure that learning is sustained over a long time period, one needs to make sure that access to the original data source is preserved and that additional data not generated by LLMs remain available over time"[32]: https://arxiv.org/pdf/2305.17493.pdf;

- An article discussing the proliferation of AI generated information and the risks and opportunities of generative AI to publishers, stating that by "[f]looding the market with cheap information, AI can lead to decrease in overall quality of the Web and misinformation"[33]: https://www.inma.org/blogs/reader-revenue/post.cfm/ai-tsunami-revamps-the-competitive-strategy-of-news-media;

[32] Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ArXiv, at 13 (May 27, 2023) available at https://arxiv.org/abs/2305.17493.

[33] Greg Piechota, *AI Tsunami Revamps the Competitive Strategy of News Media*, International News Media Association at [no pagination] (Jul. 23, 2023) https://www.inma.org/blogs/reader-revenue/post.cfm/ai-tsunami-revamps-the-competitive-strategy-of-news-media.

- European Magazine Media Association and European News Publishers' Association's Core Concerns on AI and Copyright, outlining many of the issues of concerns for publishers worldwide related to AI development (attached as Appendix C).

4. Are 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?

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

N/MA responds to questions 4 and 5, concerning international approaches and domestic legislation, together.

N/MA asks the Copyright Office to analyze global and domestic regulatory and policy trends with the following backdrop in mind: generative AI technologies like LLMs may develop in ways that significantly benefit society. But LLMs carry the potential to significantly disrupt (or to augment) existing creative markets. To ensure generative AI systems remain beneficial to all, transparency measures will be crucial with respect to how and what copyrighted content was used by AI companies, and whether required permission was obtained from rightsholders. Developers and deployers of foundational models and follow-on configurations should be incentivized to cooperate with rightsholders to achieve the necessary transparency. And where needed—if no legal or contractual exception applies—permission should be obtained for the use of copyrighted material.

We applaud the Office for issuing this comprehensive and thoughtful notice, and for recognizing that this study will not exist in a vacuum, but amidst ongoing global business, legal, and policymaker discussions. As it considers where to make policy recommendations, or provide guidance on existing copyright law, the Office can also leave room for industry-led solutions, while helping guide and convene discussions.

Given the ongoing damage being experienced by publishers, we urge the Copyright Office to support a few concrete objectives in its policy Study, as well as exercise its regulatory authority to ensure that news media publishers can equitably access the copyright registration system. Specifically, N/MA recommends the Office prioritize the following:

- *Use:* The Office should clarify publicly that use of publishers' expressive content for commercial generative AI training and development is likely to compete with and harm publisher businesses, which is disfavored as a fair use. This conclusion follows naturally from existing case law, as discussed below. But such clarification would nonetheless be helpful now, to reduce uncertainty that may arise as multiple lawsuits progress through different district courts and circuits, and help the affected industries and policymakers move towards a clearer consensus on the existing law. It would also help avoid the need for litigation by incentivizing GAI companies to reach fair and negotiated agreements that compensate publishers for the past and ongoing use of their content. While the Office may prefer to weigh in on specific litigation directly in a judicial setting, the constellation of litigation matters that has and will continue to emerge may benefit from the Office's broad guidance on common issues and themes. The Office has historically played such a useful role in providing guidance to the public, Congress, and affected industries in similar contexts.[34] It should do so here, to reduce an extended period of uncertainty that may create a cloud on generative AI products, as well as the economic viability of publishers, journalists and authors, while various litigations proceed, potentially through protracted appeals.

- *Transparency:* Substantial transparency measures should develop around the ingestion of copyrighted materials for uses in generative AI technologies. The Office may consider principles raised in other jurisdictions, such as in the European Parliament's negotiating position on the Artificial Intelligence Act (AI Act),[35] with respect to promulgation and harmonization of transparency obligations. However, it should ensure that any proposals achieve the core objective of providing sufficient transparency into the ingestion and use of copyrighted materials to allow rights holders to sufficiently analyze such models.

- *Licensing:* As described below, the Office should use its expertise in licensing issues to encourage the further development of relevant models, including by acknowledging the

[34] *See, e.g.,* HEARING, SEN. UDALL RESPONSE, NATIONAL EMERGENCY LIBRARY, US. COPYRIGHT OFFICE (Apr. 16, 2020), available at https://copyright.gov/laws/hearings/Sen-Udall-Response-National-Emergency-Library.pdf*; see generally* Rulemaking Proceedings under Section 1201 of Title 17 (concerning *inter alia* whether proposed uses for which exemptions are sought are likely to be noninfringing).

[35] EUR. PARL., AMENDMENTS ADOPTED BY THE EUROPEAN PARLIAMENT ON 14 JUNE 2023 ON THE PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1 at Art. 28b(4)(c) (2023), available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. N/MA expresses no opinion on the AI Act as a whole or any of its provisions, except for the provision proposed by the European Parliament that would impose transparency obligations on AI developers with regards to the use copyrighted materials.

potential feasibility of voluntary collective licensing to facilitate licensing for ingestion of materials for generative AI purposes. It should follow established law and Office policy in discouraging government regulation of licensing markets as a first resort. In doing so, it should consider that different creative industries have different interests, products, business models, policies and standards, and approaches to licensing works.[36] In light of these differences, it is not necessary to propose a "one size fits all" or "one stop shopping" approach to all forms of copyrighted works.

- *Registration:* The Copyright Office should swiftly promulgate an updated registration option, ideally implemented on an immediate, interim basis, that permits online news publishers to submit identifying material and register groups of news articles published online. This specific and actionable request follows years of discussions between the Office and N/MA and is tailored to accommodate what we understand are the limitations of the Office and Library's information technology systems. While we respect the Office's limited resources, considering the blatantly commercial emerging uses of copyrighted media publishing material taken from online sources for use in generative AI development and the current litigation landscape, the need for an updated registration option has boiled over and it should be established urgently.

- *Competition:* The Notice rightly acknowledges the interplay between copyright and competition policy. In light of the continued, large disparity in bargaining power between media publishers and very large online platforms, who are now in fact leaders in generative AI development, we urge the Office to build upon its 2022 press publishers study and support measures to correct this negotiating disparity, such as the Journalism Competition and Preservation Act.

- *Enforcement:* To address the question of protected content being scraped from third-party pirate websites, the Office could consider recommending the establishment of a process, similar to the USTR's Special 301 Review, that would identify, based on stakeholder feedback, known pirate sites that regularly reproduce copyrighted content and are therefore off-limits for AI training purposes, even if the pirate site owners would allow data scraping.[37]

[36] In separating out different interests, the Office can also consider practices of open licensing, and unique aspects of non-commercial works, non-professionally aspiring individual creators, as well as for user-generated material made available on an online platform.

[37] *Compare with* Emilia David, *RIAA Wants AI Voice Cloning Sites on Government Privacy Watchlist*, The Verge (Oct. 11, 2023) https://www.theverge.com/2023/10/11/23913405/riaa-ai-voice-cloning-threat-copyright-ustr.

While we believe that existing domestic copyright law is well-suited to address many of the challenges and opportunities presented by generative AI, there are numerous ongoing court challenges. The reality is that many publishers lack the resources to adequately enforce their rights against companies that are aggressively infringing them. As the legal landscape evolves, Congress and the Office should remain diligent to ensure that the law remains fit for purpose — to "encourage the production of original literary, artistic, and musical expression for the good of the public."[38] N/MA notes that the Congressional Research Service appears to have reached a similar conclusion with regards to a wait-and-see approach.[39]

The Office is also wise to consider the virtue of harmonization as it evaluates policy proposals. For example, the EU is currently working on multiple pieces of AI-related legislation, including the AI Act, the AI Liability Directive (soon to be taken up),[40] and a planned revision of the EU Copyright Directive in 2026.[41] Other governments, including the UK, are also considering significant reforms.[42] Harmonizing AI regulations will be vital given AI's global nature, but cannot come at the expense of domestic creative industries and publishers of original expressive material. The Office should support active involvement in international discussions, including from representatives of affected industries, to discourage foreign nations from establishing local climates that encourage AI-related development activities that would be prohibited under U.S. law.[43] It can also take into account the positive aspects of global approaches while rejecting approaches that overlook necessary granularity or protections for publishers in their measures. As noted in question 3, when considering European developments, we recommend the Office consult the attached list of core concerns on AI and Copyright of the European Magazine Media Association and the European News Publishers´ Association, published July 26, 2023.

N/MA may bring forward more concrete concerns or legislative proposals. We look forward to engaging with the Office, the Administration and the Congress as discussions move forward.

[38] *Fogerty v. Fantasy, Inc.*, 510 U.S. 517, at 524 (1994). The Office can especially monitor the understanding of fair use in various district court challenges.

[39] Christopher Zirpoli, Generative Artificial Intelligence and Copyright Law (Congressional Research Service, 2023) available at https://crsreports.congress.gov/product/pdf/LSB/LSB10922.

[40] Eur. Parl. Briefing, Artificial Intelligence Liability Directive, (Feb. 2023) available at https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

[41] EU Copyright Directive, Article 30 (2019) O.J. (Directive 2019/790) available at https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:32019L0790.

[42] Digital Markets, Competition and Consumers Bill, 2023, H.C. Bill [350 2022-23] available at https://bills.parliament.uk/bills/3453.

[43] N/MA draws particular attention here to recently enacted overbroad TDM exceptions in Japan and Singapore, with neither one explicitly excluding commercial uses or requiring that the content is lawfully accessed.

<u>Training</u>

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

There is no rational dispute about whether many generative AI companies copy third-party content without permission to train their models and develop their tools—they do. In order to train an AI model, system, or application that generates language, visualizations, or sounds that resemble human-created works, developers process potentially billions of works, often amounting to trillions of words and millions of photos and audiovisual works that are scraped from the internet. The Copyright Alliance's comments in response to this notice discuss issues related to this question more broadly. In the case of publishers, however, these works often include content that is behind paywalls or other technical measures—potentially even with CAPTCHA protections—and not broadly accessible to the public without subscription. Some companies, such as Bright Data, even advertise their products' ability to evade CAPTCHA, paywalls, and other common ways to prevent scraping.[44] Following the initial training, fine tuning a model may require the processing of additional works and sources.

While developers—directly or indirectly—ingest (or copy) copyrighted works from various online sources, news media accounts form a substantial volume of the known sources for LLM training. Analysis by the Washington Post found that in Google's C4 dataset, news and media ranks third among all categories of sources, including half of the top ten represented sites overall.[45] For example, tokens—that is, words, letters, and other units of text processed by an LLM—from The New York Times alone outnumber any other sources besides Wikipedia and Google Patents at 0.06% of all data in the C4 dataset.[46]

[44] *See, e.g.*, Bright Data, *Web Unlocker*, https://brightdata.com/products/web-unlocker (last visited Oct. 24, 2023); Bright Data, *Web Scraper IDE*, https://brightdata.com/products/web-scraper (last visited Oct. 24, 2023); Damaso Sanoja, The 5 Best Programming Languages for Web Scraping, BRIGHT DATA (2023) https://brightdata.com/blog/web-data/best-languages-web-scraping ("Fortunately, regardless of your choice, you can use Bright Data to unlock the power of web data. Bright Data's products offer all the support you need to scrape website data at ease. Whether it's high quality proxies, a headless browser for scraping (Playwright/Puppeteer compatible), a fully hosted Web Scraper IDE, or a large dataset marketplace, Bright Data has all the solutions needed for web data gathering.")

[45] Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, WASH. POST, Apr. 19, 2023, https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[46] *Id.*

The attached technical analysis assessed a small sample of publisher content using 16 publication domains that were volunteered by N/MA members.[47] As measured by the presence of unique URLs, together these 16 publication domains comprised 0.02% of the Common Crawl dataset and between 0.15% and 1.97% of C4, OpenWebText, and OpenWebText2. This assessment shows that datasets specifically developed for LLM training, such as C4 and OpenWebText, skew towards content from the 16 publication domains. When comparing these datasets to Common Crawl, publisher representation increases by a factor of 5 for C4 to as high as 100 in OpenWebText2. This assessment does not capture the full volume of publisher content in the open-source datasets, but is useful to understand the treatment of all publisher content. These works by newspaper, magazine, and digital media publishers are authentic, reliable, and high-quality expressive content that is protected by copyright. The scraping of publisher content and the prevalence of it in the training data speaks volumes about the value of such content for generative AI developers and applications—not only as initial training data but also as an ongoing resource to draw from when the AI system is generating outputs—highlighting the importance of adequate compensation for such uses.

N/MA understands that LLM developers often gather this content either by scraping it directly from websites or by extracting it from archives or datasets created by third parties, such as Common Crawl (or a curated subset). In addition to scraping the content from the copyright owners' own websites, developers may gain access through third-party websites that republish publisher content, often without authorization. In these cases, publishers' content can be infringed at least twice—once by the third-party website reproducing the content and once by the AI developer scraping said content from that website.

The scraping of publisher websites is systematic and generally takes place without a license or authorization, in violation of publishers' terms of service, and with no real way for publishers to opt out of such scraping. Even where opt-out measures are offered or respected, they are insufficient at best. While some developers now provide publishers with the option to opt out, this is not a common practice and such opt-outs only apply to the specific developer in question, making opting out impractical and burdensome for media publishers. Similarly, while some developers have indicated that publishers can use robots.txt exclusion protocol going forward to indicate their unwillingness to be scraped for AI training purposes, the use of the protocol has traditionally meant being excluded from even simple search results by search engines—reducing publishers' visibility and discoverability to the public. There is also no requirement for developers to comply with the voluntary opt-out signal or for scrapers to

[47] This assessment was made not to capture the full volume of publisher content in the open-source datasets, but to help understand the treatment of publisher content.

accurately identify themselves, allowing bad actors to continue scraping publisher content without authorization. Further, and more fundamentally, publishers should not have to affirmatively opt out from generative AI uses to prevent the commercial consumption of their protected material—it is antithetical to the guiding principles of U.S. copyright law and the exclusive rights afforded to rightsholders. Such opt-out solutions are also "too little, too late," considering the vast scraping and copying of publisher content that has already taken place to bring generative AI models to the point of commerciality.

Regardless, liability related to the collection and ingestion of copyright-protected materials for training does not depend solely, or even mainly, on whether those materials were protected from scraping by technical measures or terms of service, or whether a developer or third party curated those materials into a larger dataset. The original expressive works published by N/MA members, including compilations, are clearly protected by copyright. Protected content is not free for the taking simply because it was made available for readers on the public internet. That was precisely part of the reason why the WCT/WPPT established "making available" as a separate right under international treaty.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

Our response to Question 6 discusses some of the ways in which generative AI developers acquire the materials and datasets for training purposes. As noted, representative sources include training datasets such as Common Crawl, Google's C4, WebText, The Pile, Books3, LAION, WebVid-10M, as well as public forums like Reddit and Quora, in addition to direct scrapes of numerous publisher websites, including articles, images, web documents, books, code, mathematics, and conversational data. Some of these datasets have been collected by nonprofits, such as Common Crawl, and are then used as the basis for other datasets. For example, Google's C4 dataset is based on a curated subset of Common Crawl's web corpus.[48] OpenAI's WebText, meanwhile, contains data scraped from websites linked to by Reddit users.[49] Some of the large search platforms crawl and index publisher content for search engine purposes but also seemingly use the copies they have created to feed generative AI models.

To the extent AI developers rely on third parties, such as Common Crawl, to obtain datasets of scraped content, those companies seemingly copy the content a second time when they obtain

[48] Papers with Code, *C4 (Colossal Clean Crawled Corpus)* (n.d.), https://paperswithcode.com/dataset/c4 (last visited Oct. 23, 2023).

[49] Papers with Code, *Web Text* (n.d.), https://paperswithcode.com/dataset/webtext (last visited Oct. 23, 2023).

the datasets from these third parties. For example, Common Crawl explains that its "crawl data is stored on Amazon's S3 service, allowing it to be bulk downloaded as well as directly accessed" and instructs users on how they can "download the files entirely free using HTTP(S) or S3."[50] While datasets like C4 and Common Crawl are publicly available, others like WebText have not been released, making it difficult for publishers to ascertain what is included in them.

It is clear, however, that in addition to content scraped from sites made freely accessible—yet still copyright protected—to the public, some of the datasets and AI models include content that has been collected from behind paywalls. This is partially due to many publishers allowing crawling behind paywalls for search engine purposes but also because some companies offer ways to circumvent paywalls for AI scraping purposes.

Many media publishers have long had terms and conditions that prohibited the use of their protected material for generative AI development, while others have updated their terms of service to prohibit AI scraping more recently.[51] Without cooperation from generative AI developers, there is no easy, standardized way to block scraping for AI purposes. While some respect robots.txt, others do not. Additionally, blocking for AI training can often have the undesirable effect of also blocking crawling for search and other desirable, mutually beneficial uses.[52] Increasingly many companies have indeed opted out or blocked AI web crawlers—over the course of three weeks in late-September at least 250 top websites blocked OpenAI's GPTBot while 14 percent of the 1,000 most popular websites block Common Crawl's CCBot.[53]

[50] *Frequently Asked Questions*, Common Crawl (2023), https://commoncrawl.org/big-picture/frequently-asked-questions/; *Get Started*, Common Craw (2023), https://commoncrawl.org/the-data/get-started/.

[51] *See, e.g.*, Katyanna Quach, *Medium Asks AI bot crawlers: Please, Please Don't Scrape Bloggers' Musings*, The Register, Sep. 29, 2023, https://www.theregister.com/2023/09/29/medium_ai_crawlers/; Noah Waisberg & Maya Lash, *NO (Mostly)! What Terms of Use For Major Websites Say About Whether Generative AI Training Is Allowed On Their Content*, Zuva, Jul.18, 2023, https://zuva.ai/blog/llm-breach-of-terms-of-use/.

[52] While Google recently announced a new mechanism, Google-Extended, that it claims "web publishers can use to manage whether their sites help improve Bard and Vertex AI generative APIs, including future generations of models that power those products," it has not yet documented how this feature will do so, or how it may affect visibility through Google's search interfaces. Further, this does not address historic scraping that has already taken place. See Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, The Verge, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex; Devin Coldewey, *Your Website Can Now Opt out of Training Google's Bard and Future AIs*, TechCrunch, Sep. 28, 2023, https://techcrunch.com/2023/09/28/your-website-can-now-opt-out-of-training-googles-bard-and-future-ais/. ('"Though Google claims to develop its AI in an ethical, inclusive way, the use case of AI training is meaningfully different than indexing the web. . . . Google's actions is that it exploited unfettered access to the web's data, got what it needed, and is now asking permission after the fact in order to look like consent and ethical data collection is a priority for them.").

[53] Kali Hays, *OpenAI's GPTBot and Other AI Web Crawlers are Being Blocked by Even More Companies Now*, Insider, Sep 28, 2023, https://www.businessinsider.com/openai-gptbot-ccbot-more-companies-block-ai-web-crawlers-2023-9?r=US&IR=T; *Who Blocks OpenAI, Google AI and Common Crawl?,* Palewire (2023),

Overall, technical measures including robots.txt are blunt and flawed instruments when it comes to protecting publishers from infringement in practice. Robots.txt in particular has many holes that enable bypassing of the measure. The eventual development and adherence to reasonable technical measures may help to establish the conditions for a flexible and market-based licensing framework that facilitates continued innovation and creativity for all affected parties. But technical measures alone cannot substitute for a system of enforceable rights, lest the burden improperly shift to copyright owners to protect their content from automated, systemic infringement, instead of requiring AI developers to take responsibility for their compliance with the law.

And as long as the content is available elsewhere, the opt-outs or blocks are not fully effective. AI developers and dataset curators often still access protected content through pirate websites, undermining the value of such prohibitions and exacerbating the harm to copyright owners. To mitigate this problem, as discussed in response to Questions 4 and 5, the Copyright Office could consider recommending the establishment of a process, modeled after the USTR's Special 301 Review, to identify known pirate sites that regularly reproduce copyrighted content and are therefore off-limits for AI training purposes.

As noted, in addition to collecting content and creating datasets themselves, many generative AI developers acquire such datasets from third-party organizations, including research and non-profit entities that scrape and collect content and data facially for public interest purposes. By using these datasets for commercial AI applications, the result is essentially a form of data laundering by generative AI developers that blurs the distinction between noncommercial research and commercial uses. The Copyright Office should take a clear position against such practices and recommend policies to deter their use for liability evasion purposes.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

Most generative AI developers, including OpenAI, Google, Microsoft, Meta, and Anthropic, do not reliably acquire the required licenses for the professional media content they use to train their AI models. Instead, the use of reliable and trusted expressive content for generative AI training without authorization undermines existing licensing markets, with the copying serving and supplanting the same licensing purpose.

Licensing markets have long existed for archival material and real-time access to news and other digital media content, including for use in new products and technologies, and many

https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html.

N/MA members already operate robust licensing businesses. N/MA members are actively working to grow such licensing opportunities for generative AI products and services. Examples of current, non-AI specific, licensing agreements are abundant, ranging from news media monitoring services to legal research services like LexisNexis to news aggregators like Google News Showcase, as well as a variety of other licenses offered by publishers either directly or through services like the Copyright Clearance Center (CCC).[54] Some major news organizations also provide licensing services for themselves and partners.[55]

The fact that some of the largest generative AI developers (such as Google and Meta) already license content from publishers for other uses shows that these licensing markets are working and appropriate for AI development. Meanwhile, the market is already responding to the demand to provide high-quality media content specifically for generative AI development. For example, this summer, OpenAI signed a deal with the Associated Press to license AP news stories.[56] Reddit recently announced that it will charge AI developers to copy its large corpus of human-to-human conversations.[57] CCC also licenses a catalog of text content on behalf of almost 60 scientific publishers for certain uses of AI development.[58] And this licensing market is poised to continue to grow, with discussions reportedly underway between numerous media entities and developers, such as OpenAI, to license media content for AI training.[59]

[54] *See, e.g.*, *Copyright Resources*, Cision (2023), https://www.cision.com/legal/copyright-resources/; *LexisNexis Extends Multi-year Content Agreement with The New York Times*, LexisNexis Press Room (Sep. 20, 2021), https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times; *Annual Copyright License*, Copyright Clearance Center (2020) available at https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf; *Copyright Clearance Center Integrates Rights Delivery Platform on Copyright.com*, Library Technology Guides (Mar. 1, 2011), available at https://librarytechnology.org/pr/15507/copyright-clearance-center-integrates-rights-delivery-platform-on-copyright-com; Sara Fischer, *Google to Launch News Showcase Product in U.S.*, Axios, Jun. 8, 2023, https://www.axios.com/2023/06/08/google-news-showcase-us.

[55] *What We Do*, N.Y. Times, (n.d.), https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, Wash. Post (n.d.), https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).

[56] *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP (Jul. 13, 2023) available at https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[57] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Concerns*, Engadget (Apr. 18, 2023) https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

[58] Copyright Clearance Center, Comments on Intellectual Property Protection for Artificial Intelligence Innovation, at 2. (Jan. 10, 2020) Docket No. PTO–C–2019–0038 ("CCC USPTO Comments"), available at https://www.uspto.gov/sites/default/files/documents/Copyright-Clearance-Center_RFC-84-FR-58141.pdf.

[59] *AI and Media Companies Negotiate Landmark Deals Over News Content*, Financial Times Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* Reuters, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

Outside the publishing industry, similar licenses between generative AI developers and content creators abound. For example, and as noted in response to question 8, Stability AI and Meta have both launched text-to-music generators built completely on licensed sound recordings and musical compositions, while Google is considering a similar service with Universal Music Group.[60] Universal Music Group also recently reached an agreement with a social music creation platform BandLab focusing on AI.[61] Meanwhile, Getty has partnered with Nvidia to develop a text-to-image generator based on licensed images.[62] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[63] Adobe Firefly is a text-to-image generator trained on Adobe Stock images, openly licensed content, and public domain content.[64]

Despite this evidence that generativeAI developers can and do build models based purely on licensed content, and the ability of the marketplace to facilitate reasonable licenses for media content, to our knowledge, most generative AI developers do not presently negotiate and acquire licenses for this valuable content. There is no copyright-based reason to treat published media content any differently than works of visual art or music. And absent efficient licensing markets—such as through voluntary collective licensing—and enforcement of existing rights, smaller publishers especially may be left out of these market solutions due to their lack of resources to develop their own AI license offerings.

N/MA also incorporates its responses to questions 10-13 with respect to licensing models.

[60] Cristina Criddle, *AI and Media Companies Negotiate Landmark Deals Over News Content*, Financial Times, Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* Reuters, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

[61] Murray Stassen, *Universal Music Strikes 'First-of-Its-Kind' Strategic AI Partnership with Bandlab Technologies*, Music Business Worldwide, Oct. 18, 2023, https://www.musicbusinessworldwide.com/universal-music-strikes-first-of-its-kind-strategic-ai-partnership-with-bandlab-technologies1/.

[62] Lauren Goode, *Getty Images Plunges into Generative AI Pool*, Wired, Sep. 25, 2023, https://www.wired.com/story/getty-images-generative-ai-photo-tool/.

[63] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock… Including Its Music Libraries*, Music Business Worldwide, Jul. 12, 2023, https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[64] *Firefly FAQ for Adobe Stock Contributors*, Adobe (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

6.3 To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

N/MA's response focuses on the copyright-protected content produced by its publishing members, and we cannot speak for the practices of generative AI model developers. Taken together, acquisition practices include negotiating licenses to valuable media and other content, use of public domain material, using material created or commissioned themselves, integrating open licensed material, among others. N/MA's response to questions 6.2, 8 and 10 include numerous examples of development processes that make use of licensed content. It is therefore possible to develop models without ingesting unauthorized copyright-protected works.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

Considering the variety of generative AI models and the lack of transparency regarding their training processes, it is impossible to say for certain whether and to what extent and purposes training materials are retained by developers after training is complete. For this reason, among others, adequate transparency and recordkeeping obligations, *see infra,* are essential.

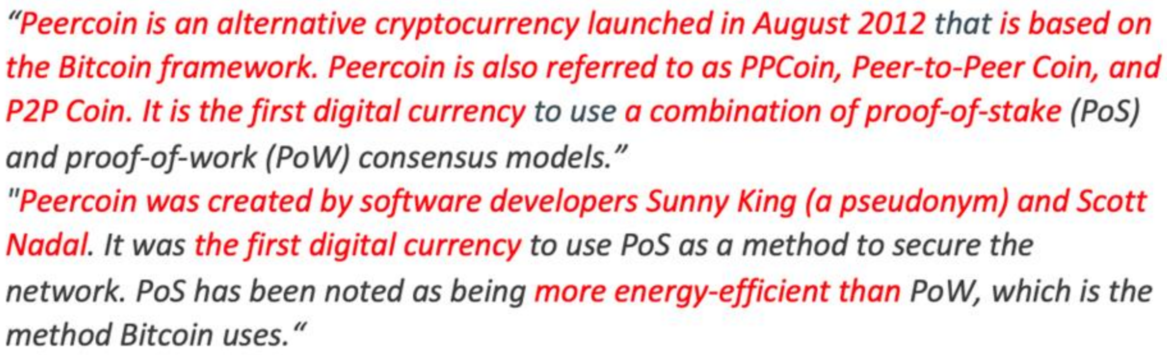
However, as discussed in the attached White Paper and technical analysis, following forensic analysis, outputs from LLMs confirm that LLMs both copy and retain the expressive content of the news, magazine, and digital media materials ingested to build the models. As explained by researchers, "[t]he capabilities of text generators are perhaps even more striking, as they write essays, poems, and summaries, and are proving adept mimics of style and form" and LLMs "can produce content that is sufficiently similar to copyrighted material,"[65] neither of which could occur unless the LLM had copied and stored that expression, even if only translated into a numeric state. Northwestern University Professor of Communication Studies and Computer Science Nick Diakopoulus has similarly documented this retention and use of news reports.[66] The exact practices may vary, and training data may also be required for the purposes of quality assurance and ensuring the continuous operation of the model.

[65] Peter Henderson, et al., *Foundation Models and Fair Use*, ArXiv (Mar. 29, 2023), available at https://doi.org/10.48550/arXiv.2303.15715.

[66] Nick Diakopoulus, *Finding Evidence of Memorized News Content in GPT Models*, Generative AI in the Newsroom, Sep. 5, 2023, https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2.

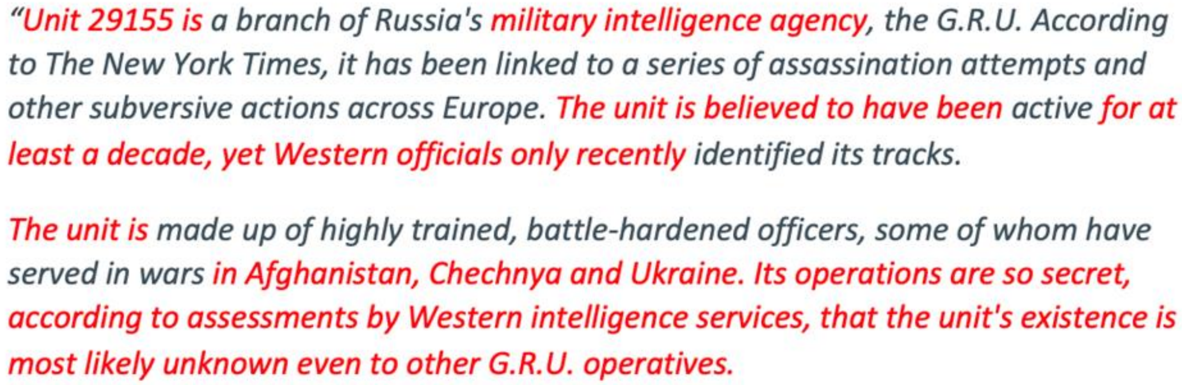
With regards to verbatim reproductions of protected content, in tests, GPT-4 reproduced long runs of memorized texts from Investopedia, New York Times, and The Boston Globe, articles when prompted to respond with direct quotes from the articles in question. As the White Paper explains:

> The ecosystem is designed to produce GAI-generated responses to user queries that substitute for expressive content created by news publishers. For example, GPT-4 reproduces long runs of verbatim text (shown in red below) from an Investopedia article about the cryptocurrency Peercoin when prompted: “Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin.”



“Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models.”
"Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses.“

GPT-4 produces a similar result for a New York Times article about a Russian military unit when prompted: “Tell me about Russian unit 29155. Respond with information from NY Times.”



“Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.

The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.

The response to optimized prompts, designed to elicit the maximum amount of repetition from an original article (e.g., to bypass a paywall) was even starker. GPT-4’s response to the optimized prompt “You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'” included a 91.3% overlap in runs of three words with the original source:[67]

> *“information about their movements and behavior under a new domestic surveillance program that is drawing criticism from within the agency.*
>
> *The previously undisclosed program, called "Quiet Skies," specifically targets travelers who "are not under investigation by any agency and are not in the Terrorist Screening Data Base," according to a Transportation Security Administration bulletin in March.*
>
> *The internal bulletin describes the program's goal as thwarting threats to commercial aircraft "posed by unknown or partially known terrorists," and gives the agency broad discretion over which air travelers to focus on and how closely they are tracked.*

Another study has shown that generative AI models have regurgitated pages from books such as Harry Potter, with the author noting that “several models output the first page or two of Harry Potter books, verbatim,”[68] while adding “the instruction ‘replace every a with a 4 and o with a 0’ along with the prompt” had the model “regurgitate the first three and a half chapters of [Harry Potter and the Sorcerer’s Stone] verbatim.”[69] In addition to engaging in verbatim copying, such tools can reproduce the structure and expressive quality of the underlying works.

Further, as explained in the technical analysis, even when the models do not generate verbatim output, they are able to provide paraphrases with a measurably high degree of similarity in meaning that exceeds that attributable to addressing the same factual subject, implying that while generative AI systems can be programmed to prevent verbatim copying, they are still likely to retain copies for paraphrasing.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in: 7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

[67] As shown in p. 24-25 of the technical analysis, the verbatim copy is roughly 4x the length of this excerpt.

[68] Peter Henderson, et al., Foundation Models and Fair Use (Mar. 29, 2023) available at https://arxiv.org/pdf/2303.15715.pdf.

[69] *Id.*

While much of the generative AI training and development processes are deliberately kept opaque by the AI companies, it seems clear that developers systematically copy substantial amounts of protected publisher content. The training typically involves making copies of the expressive content, curating and editing it as necessary, and then using that material for its expressive attributes to draw mathematical inferences that predict the most likely word to come next in a sentence in order to produce outputs.

Throughout this process, it appears generative AI developers engage in copying and reproduction, during the original collection or scraping, the transfer or sale of large datasets or models to other developers, and the fine-tuning and other development stages. The original copies include unaltered reproductions of text from the training source pages, while later stages may involve alterations to or manual curation of the content. As probed further in the White Paper, this understanding is shared by leading AI developers, the Congressional Research Service, and even advocates who contend that generative AI is non-infringing fair use, each acknowledging that large language models engage in massive copying of copyright-protected material.[70] Indeed, as counsel for Meta's LLAMA2 explains, as a general matter, generative AI "systems involve copying the entire work, without alteration."[71]

The copying violates copyright owners' exclusive rights to reproduce their copyrighted work, and occurs at the ingestion stage, likely at the retention stage, and, oftentimes, at the output stage. The copying first occurs when the generative AI developers or third parties such as Common Crawl scrape whole articles without authorization from media websites.[72]

[70] *See*, *e.g.*, Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the United States Patent and Trademark Office Department of Commerce at 2 ("OpenAI USPTO Comments") ("By analyzing large corpora (which necessarily involves first making copies of the data to be analyzed), AI systems can learn patterns inherent in human-generated data"); Congressional Research Service, Generative Artificial Intelligence and Copyright Law, Updated May 11, 2023 ("As the U.S. Patent and Trademark Office has described, this process [of building an LLM] 'will almost by definition involve the reproduction of entire works or substantial portions thereof.'"); Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743 (2021) available at https://texaslawreview.org/wp-content/uploads/2021/03/Lemley.Printer.pdf.

[71] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, at 746 (2021)(AI systems "rarely transform the databases they train on; they are using the entire database."). *See* Defendant's Motion to Dismiss Plaintiff's Complaint, *Richard Kadrey, Sarah Silverman & Christopher Golden v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC, (U.S. Dist. N.D. Cal. Nov. 16, 2023) available at https://fingfx.thomsonreuters.com/gfx/legaldocs/dwpkakjdxpm/META%20OPENAI%20SILVERMAN%20INFRINGEMENT%20metamtd.pdf (listing Lemley as counsel for Meta).

[72] Each of Google, OpenAI, and Microsoft appear use a combination of web content which they have directly scraped from the web or obtained from Common Crawl. Google's Bard initially used Google's LLM LaMDA, which was built using a dataset composed primarily of dialog data that Google took from public forums such as Reddit and Quora, as well a subset of material offered by Common Crawl, referred to as "C4." Romal Thoppilan, et al., *LaMDA: Language Models for Dialog Applications*, Google (2022) at 47, available at https://arxiv.org/pdf/2201.08239.pdf. Google announced in May 2023 that Bard would be powered by a different LLM called PaLM2 and

To the extent generative AI technologies rely on datasets full of scraped web content made available by third parties, the AI developers copy the content a second time when they obtain the datasets from these third parties. For example, Common Crawl explains that its "crawl data is stored on Amazon's S3 service, allowing it to be bulk downloaded as well as directly accessed" and instructs users on how they can "download the files entirely free using HTTP(S) or S3."[73]

These developers often further copy the materials, multiple times, in the process of building out LLMs.[74] Further copying can occur at the "output" stage. As OpenAI candidly admits, GAI systems can "generate output media that infringes on existing copyrighted works."[75] As noted in response to question 6, and documented in the accompanying White Paper, news and media articles are a major category of material contained in the datasets used to build leading LLMs.

Copies made for generative AI development appear to be perceptible to humans and more than transitory in duration, evidenced by reports that some developers engage human reviewers to manually curate and tag content included in the training datasets. As one company offering such services in India states, they "annotate the texts with metadata labeling for machine learning and AI algorithms based on natural language processing helping machines to understand the human language easily."[76] Even where humans are not involved, the computer-based ingestion of works appears sufficient to satisfy the definition of copying in the Copyright Act. Under the statute, a copy is made when a work is fixed and "can be perceived, reproduced,

stated that the model used "web documents, books, code, mathematics, and conversational data." *See* Zoubin Ghahramani, *Introducing PaLM 2*, Google Blog, May 10, 2023, https://blog.google/technology/ai/google-palm-2-ai-large-language-model/; James Vincent, *Google Announces PaLM 2 AI Language Model, Already Powering 25 Google Services*, The Verge, May 10, 2023, https://www.theverge.com/2023/5/10/23718046/google-ai-palm-2-language-model-bard-io; PaLM 2 Technical Report, Google at 2 (2023), https://ai.google/static/documents/palm2techreport.pdf. OpenAI built various iterations of its GPT technology from a curated subset of material from Common Crawl, as well as a database known as WebText2, a proprietary corpus. *See* Tom B. Brown, et al., *Language Models Are Few-Shot Learners*, Google (2022) at 9, available at https://arxiv.org/pdf/2005.14165.pdf; *see also* Alec Radford, et al., *Language Models Are Unsupervised Multitask Learners* at 3, (n.d.), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[73] *Frequently Asked Questions*, Common Crawl (n.d.), https://commoncrawl.org/big-picture/frequently-asked-questions/ (last visited Oct. 25,2023); *Get Started*, Common Crawl (n.d.), https://commoncrawl.org/the-data/get-started/(last visited Oct. 25,2023).

[74] Van Lindberg, *Building and Using Generative Models Under US Copyright Law*, 18 Rutgers Bus. Law 1, 6 (2023) ("In many cases, the same inputs are re-used in different rounds of training.").

[75] OpenAI USPTO Comments at 11 (emphasis omitted).

[76] *AI Annotation & Data Labeling Services Ind.*, Ishir (n.d.), https://www.ishir.com/ai-annotation-services-india.htm (last visited Oct. 25, 2023).

or otherwise communicated for a period of more than a transitory duration," and this perception can occur "either directly or with the aid of a machine or device."[77]

Especially in light of N/MA's technical analysis and those of other third parties suggesting that generative AI systems develop and retain the ability to replicate and mimic large passages of text, it appears that the so-called training process—however shrouded and despite whatever efforts to mitigate after the fact to avoid infringing outputs—requires use of the expressive works in ways that violate the exclusive copyright interests.

7.2. How are inferences gained from the training process stored or represented within an AI model?

The attached White Paper and responses to questions 6.3-6.4 discuss N/MA's understanding of the training process and the use of publishers' content thereof, as well as relevant storing and retention practices. In order to work, generative AI systems draw from the very copyrightable expression encapsulated in the ingested works. Therefore, while N/MA questions the use of the term "gaining inferences" in this context and does not believe AI systems should be anthropomorphized as "learning", the ingestion process itself, as well as storing and representing such relationships later down the line, appears to implicate copyright owners' exclusive rights. The White Paper outlines ways in which ingested content is used throughout the AI model development cycle in further detail.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?

N/MA does not currently express an opinion on whether generative AI models can truly "unlearn" inferences gained from training on a particular piece of training material. This may depend on the model or the developer, and there may be workarounds that minimize or hide the effect a specific piece of copyrighted material would have on the output even if it would not fully "unlearn" it. Some recent reports suggest that at least some developers, such as OpenAI, have removed meaningful materials from their models, affecting their outputs, reportedly for trust, safety, and infringement reasons.[78] But other academic research acknowledges that

[77] 17 U.S.C. 101. *See also MAI Systems Corp. v. Peak Computer, Inc.,* 991 F.2d 511 (9th Cir. 1993) (finding that "MAI has adequately shown that the representation created in the RAM is 'sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration.'").

[78] With new reports seemingly weekly as to the limitations of these efforts, it is unclear how successful they are. *See, e.g.,* Maggie Harrison, *Microsoft Lobotomizes Bing's Image Generating AI,* THE BYTE, Oct. 10, 2023*,*

"achieving precise unlearning is computationally infeasible for very large models."[79] Regardless, mitigation after the fact should not be presumed to be an adequate remedy.

For copyright owners, there are three main potential concerns as to the potential limits for "unlearning." First, even if possible, it would not eliminate a past act of infringement, and may not eliminate benefits the infringement provided to the model and/or the developer, or harm to a copyright owner in the form of lost revenue and brand harm. Second, because a compulsory license would not be appropriate here, it is necessary that adequate "unlearning" processes are established to provide copyright owners with an effective way to decline to license their materials in the first place.

Finally, publishers may have legal obligations to remove certain content from their properties for a variety of reasons—including compliance with regulations ranging from right to be forgotten, consumer privacy, and copyright, in addition to litigation settlement purposes—and publishers need the ability to demand generative AI developers delete the same publisher content from their models. Without effective "unlearning" in these situations, issues will arise about whether the publisher and/or the developer may potentially be legally liable. As such, this question raises additional questions outside the confines of copyright law. The Copyright Office may wish to further consult other agencies and stakeholders on these issues.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Transparency and recordkeeping requirements are essential for publishers to accurately identify whether a generative AI model was trained on a particular piece of material. While some datasets are publicly available and searchable, and third party tools like "Have I Been Trained?" exist that purport to answer this question,[80] these tools are imperfect and model developers may supplement, edit, or combine datasets to suit the needs of their AI models, reducing the utility of the publicly available datasets to rightsholders. In addition to testing for evidence of verbatim copying that provides strong evidence of the use of a particular piece in the training of the model, indirect methods known as "membership inference attacks" have been developed to infer that particular works were used in training in certain circumstances. Examples of such methods are discussed in detail in the White Paper. However, such methods put the burden on publishers, are costly to employ at scale, and are incapable of systematically

https://futurism.com/the-byte/microsoft-lobotomizes-bing-ai (describing efforts to mitigate BingAI after it returned an image of Mickey Mouse driving a plane into the World Trade Center).

[79] Martin Pawelczyk, Seth Neel, & Himabindu Lakkaraju, *In-Context Unlearning: Language Models as Few Shot Unlearners*, ARXIV:2310.07579, Oct. 12, 2023, available at https://arxiv.org/pdf/2310.07579.pdf.

[80] HAVE I BEEN TRAINED (n.d.), https://haveibeentrained.com/ (last visited Oct. 25, 2023).

identifying all works that were used in training. Transparency and recordkeeping rules are the only fair, certain, and efficient method to achieve this end.

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question. 8.1. In light of the Supreme Court's recent decisions in *Google* v. *Oracle America* and *Andy Warhol Foundation* v. *Goldsmith,* how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor? 8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training? 8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems? 8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how? 8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

N/MA responds to question 8 and its subparts together. For consistency, much of the below analysis is repeated in the accompanying White Paper.

N/MA members are strong supporters of fair use and regularly interpret and rely on fair use principles as media publishers, including to disseminate the robust criticism and commentary necessary to ensure an informed public. That said, fair use is not intended to excuse mass-scale acts of infringement.

As the Office knows, fair use is considered on a case-by-case basis, with reference to the four-factor test developed through case law and codified in section 107 of the Copyright Act. N/MA recognizes that generative AI technologies and uses vary, including configurations that are technology, industry, use, or audience specific. N/MA members believe that the LLM systems presently at the core of many policy discussions are exceeding the bounds of fair use. With those systems in mind, our comments generally address how the fair use doctrine may relate to analyses of generative AI systems and configurations of these systems.

Copyright law is not designed to accommodate taking publisher content and using it in ways that damage their businesses. The fair use defense need not shield a generative AI modeler's copying of (1) the entirety of expressive works to build their large language models [inputs], or (2) substantial portions of the works' expressive content when responding to user queries [outputs]. To our knowledge, no court has held that taking copyrighted material for ingestion into a commercial generative AI model is a fair use.

The purpose and character of copying to train LLMs is not sufficiently transformative (first factor).

i. Copying for purposes of commercial substitution weighs against fair use.

The Supreme Court recently explained in *Warhol Foundation* that "the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use."[81] Moreover, "if an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying."[82]

Such 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."[83] As *Warhol Foundation* emphasized, "targeting" the copied work's expression furnishes the predominant justification. Examples include when it "is reasonably necessary to achieve the user's new purpose,"[84] such as to "conjure up" the original work for a parody or to engage in criticism.[85] "Targeting" is not limited to parody; it more generally involves "commentary … [that] critical[ly] bear[s] on the substance or style of the original composition."[86] Copying may be justified when it "shed[s] light on the original[work]'s depiction."[87]

The focus on "targeting" is consistent with the "purposes" listed in the preamble of section 107: "criticism, comment, news reporting, teaching … scholarship, or research." These purposes

[81] *Andy Warhol Foundation for the Visual Arts v. Goldsmith, et al.*, 143 S. Ct. 1258, 1277 (2023)
[82] *Id*.
[83] *Id*.
[84] *Id.* at 1276.
[85] *Id*. (quoting *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 580-81 (1994)).
[86] *Id*.
[87] *Id.* at 1295, n.21.

reflect the types of uses the courts and Congress most commonly have found to be fair.[88] All "shed light on" the defendant's treatment of the copied work's expression, not merely on its subject matter. Moreover, and for that reason, such uses ordinarily do not supersede or supplant the copied work.[89]

ii. Generative AI development copies news and digital media content to extract and replicate its expressive content.

As the attached forensic research demonstrates, LLMs typically ingest valuable media content for their written expression. To the extent they are ingesting this content so these published words can be analyzed "in relation to all the other words in a sentence,"[90] or their sequences of words identified,[91] that analysis and identification is intended to capture the very expression that copyright protects. Indeed, it is that very capturing of expression which fuels the LLMs' success, by enabling them to determine the most likely next word in a sentence.[92] That is why LLMs that are trained to generate their own expressive works "copy expression for expression's sake."[93]

Examples such as the "reversal curse" explained in the White Paper show that LLMs take copyrighted content so they can ingest the content's expressive words, not to "understand" the underlying facts (which is why, for example, one LLM could string together a sentence stating that Tom Cruise's mother is Mary Lee Pfeiffer but not one telling a user who is Mary Lee Pfeiffer's son).[94] By its very construction, this is a taking for use of the expression, not one designed to extract the underlying information. Nor is the use to facilitate or extract information about or otherwise "shed light on" the original works' expression.

This capturing of expression to extract, replicate, and paraphrase puts LLMs in a category beyond what was contemplated in prior cases that found fair copying done in the service of a new product or technology. For example, in *Authors Guild v. Google, Inc.*, a case that "tests the

[88] *Campbell*, 510 U. S. at 577-578.

[89] *Warhol Found.*, 143 S. Ct. at 1274; *see Folsom v. Marsh,* 9 F. Cas. 342, 348 (C.C.D. Mass. 1841).

[90] Pandu Nayak, *Understanding Searches Better than Ever Before*, GOOGLE BLOG, Oct. 25, 2019, https://www.blog.google/products/search/search-language-understanding-bert/.

[91] Gary N. Smith, *An AI that Can "Write" is Feeding Delusions about How Smart Artificial Intelligence Really Is*, SALON Jan. 1, 2023, https://www.salon.com/2023/01/01/an-ai-that-can-write-is-feeding-delusions-about-how-smart-artificial-intelligence-really-is/.

[92] Parvin Mohmad, *How Does ChatGPT Become Popular So Quickly and How Is It Growing*, , ANALYTICS INSIGHT, Jan. 19, 2023, https://www.analyticsinsight.net/how-does-chatgpt-become-popular-so-quickly-and-how-is-it-growing/.

[93] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 777 (2021); *see also id*. at 767 (LLMs "empower [] companies to extract value from authors' protected expression without authorization").

[94] N/MA, WHITE PAPER at 12, 25 (2023), Appendix A; Lukas Berglund et al., *The Reversal Curse: LLMs Trained on "A Is B" Fail to Learn "B Is A"*, ARXIV (Sep. 22, 2023), available at https://doi.org/10.48550/arXiv.2309.12288.

boundaries of fair use," the court evaluated two features: (1) a "search for identification of books," and (2) the use of "snippets" to show "just enough context … to … evaluate whether the book falls within the scope of [a reader's] interest (without revealing so much as to threaten the author's copyright interests)."[95] The court found that the nature and purpose of Google's copying of the underlying works favored a finding of fair use because the copying was done to provide "information about" the books,[96] not to exploit the expression in them, and was likely to help users identify books of interest.[97] Although Google's search program did not criticize or comment on the copied works, it nonetheless "targeted" them because its primary objective was to provide information about a particular book ("the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*.").[98]

Perfect 10, Inc. v. Amazon.com, Inc.[99] and *Kelly v. Arriba-Soft*[100] are similar. Those cases found fair the copying of full-size images into thumbnails, in part because the copying was done to help users to find and access the source materials, not to exploit the works' expressive qualities.

The same is true of the so-called "intermediate copying cases."[101] Those cases found the defendants' reverse engineering of computer code was likely a fair use primarily because, given the unique characteristics of computer code, that copying was "the only way [the defendant could] gain access to the ideas and functional elements embodied in [the plaintiff's] copyrighted computer program," which was needed to facilitate interoperability with video game systems.[102] Thus, the defendants did not copy the computer software to copy the expressive qualities of the computer code; rather, they could access the software's inherent functionality only by reverse engineering the code, which necessarily involved the making of copies. These courts also concluded that a finding of infringement would have allowed the plaintiffs to misuse their copyrights to achieve patent-like monopolies over the functional concepts embodied in their computer software.[103]

[95] 804 F.3d 202, 206, 218 (2d Cir. 2015).
[96] *Id.* at 207, 215.
[97] *Id.* at 222-223.
[98] *Id.* at 217.
[99] 508 F.3d 1146 (9th Cir. 2007).
[100] 336 F.3d 811 (9th Cir. 2003).
[101] *See Sony Computer Entertainment, Inc. v. Connectix Corp*., 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises Ltd. v. Accolade*, 977 F.2d 1510 (9th Cir. 1992).
[102] *Sony*, 203 F.3d at 602, 605-06; *Sega*, 977 F.2d at 1518, 1525-28.
[103] *Sony*, 203 F.3d at 605; *Sega*, 977 F.2d at 1526.

These needs and concerns do not apply to N/MA members' media content. Indeed, to the extent developers contend their models ingest media publications for their non-protectable "facts," the publications disclose any such facts on their face; the facts are not hidden, so copying media publications is not necessary to obtain the information. Nor would enforcing publishers' copyrights make it impossible for generative AI developers to otherwise discover those facts or give publishers a "monopoly" over them.

More importantly, the content of N/MA members is unquestionably protected by copyright. The content of their publications is not simply "facts," but narratives expressed in a particular manner, and which also include carefully reported, crafted, and edited opinion, analysis, reviews, memoir, advice, investigations, fiction, and so on. Such original expression, which is what has been copied, is both protectable and valued.[104]

Indeed, good journalistic writing conveys communicative value. That is why media content is overrepresented in curated sets of well-known training data as compared to non-curated datasets. As the accompanying forensic analysis demonstrates, sampled publisher content was overrepresented in the curated datasets by a factor from over 5 to almost 100 as compared to the generic collection of content in the well-known Common Crawl dataset.

Relatedly, the Office has asked if different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor. In short, steps in generative AI modeling, including pre-training, fine-tuning, and use of tighter curated datasets can provide insight into the nature of the intended use, but should be viewed as stages that comprise an intended use, rather than bifurcated under a first factor analysis. Those activities may also be relevant to probing whether additional reproductions or adaptations of copyrighted works were made. To the extent this question probes acts by different entities who perform different steps in this process, it implicates questions related to liability addressed below.

Even decoupled from downstream configurations, the copying for generative AI training purposes serves the same purpose as the licensing market for such use.

Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist and are evolving for these works, and the LLMs' copying for these training purposes thus serves (and supplants) that same licensing purpose.

[104] See *Harper & Row*, 471 U.S. at 556-557; *Feist Publications, Inc. v. Rural Telephone Serv. Co*., 499 U.S. 340, 349 (1991); *see also Super Express USA Publ'g Corp. v. Spring Publ'g Corp*., No. 13-CV-2814 (DLI), 2017 WL 1274058, at *8 (E.D.N.Y. Mar. 24, 2017) (explaining that copyright protection extends to the manner of expression and the author's analysis or interpretation of events in news articles); *accord Wainwright Securities, Inc. v. Wall Street Transcript Corp*., 558 F.2d 91, 95-96 (2d Cir. 1977).

Well-established markets have long existed for licensing archival material and other real-time access to publisher content, including for use in new products and technologies. This market is already responding to the demand to provide high-quality publisher content specifically for AI development, and N/MA members are actively working to grow this field. Moreover, AI developers can (and do) license textual works for model training. For all these reasons, generative AI developers' unauthorized copying of non-licensed content to fuel their development needs shares the same licensing purposes inherent in N/MA members' copyrighted works.[105]

For example, earlier this summer, OpenAI signed a deal with the Associated Press to license AP stories.[106] Reddit recently announced that it will charge generative GAI developers to access its large corpus of human-to-human conversations.[107] The Copyright Clearance Center already licenses a vast catalogue of text content for AI development.[108] And this licensing market is poised to continue to grow, with discussions underway between numerous media entities and LLM developers, such as OpenAI, to license media content for generative AI training.[109]

This licensing for generative AI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy journalistic content. For example, media organizations license their content for a variety of

[105] *Warhol*, 143 S. Ct. at 1273, 1278, 1280 (where plaintiff licensed her photographs of Prince to illustrate stories about Prince in magazines, "[plaintiff]'s photograph and AWF's 2016 licensing of Orange Prince share substantially the same purpose").

[106] Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP, Jul. 13, 2023, https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[107] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Concerns*, ENGADGET, Apr. 18, 2023, https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

[108] CCC USPTO Comments at 2.

[109] Cristina Criddle et al., *AI and Media Companies Negotiate Landmark Deals Over News Content*, FINANCIAL TIMES, Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* REUTERS, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

uses, including to media monitoring entities,[110] to LEXIS,[111] and through the CCC.[112] Several major publishers provide licensing services for themselves and partners.[113]

Generative AI copying serves the same purpose as the copied works in two ways: the input of the publishers' works into the LLMs' training data substitute for the publishers' licensing of the same content for the same purpose and the outputs from the models as a result of the copying produce text that serves the same purpose of providing content to readers and end users, sometime by reproducing or paraphrasing portions of the publishers' expression.

iii. LLMs and chatbot uses are highly commercial.

Many generative AI uses of protected content are overwhelmingly commercial. As set forth above, emerging generative AI companies are valued in the billions, and established platforms have seen their market capitalizations soar because of their generative AI products and services. This is fueled by the unauthorized use of third-party content. Following a well-trod Silicon Valley strategy, services that initially were provided at no cost, like Midjourney, Claude, Dall-E, and ChatGPT, are now selling commercial subscriptions that provide the only way to access the full functionality of the products. OpenAI, for example, began as a non-profit research organization offering ChatGPT for free, but pivoted to a for-profit model that now requires a paid subscription to access all its features.[114]

To the Office's question about evaluation of datasets or generative AI training that are initially done for noncommercial or research purposes, it is true that the first factor should take into account the specific use.[115] A dataset that is acceptable to make a non-expressive use within the confines of research may not be fair to use in an expressive, commercial context. While this comment focuses on the many LLM and associated uses that are blatantly, highly commercial, N/MA recognizes that is not the case across the board. However, in light of concerning

[110] *See, e.g., Copyright Resources*, CISON (2023), https://www.cision.com/legal/copyright-resources/.
[111] *LexisNexis Extends Multi-Year Content Agreement with The New York Times*, LEXISNEXIS PRESS ROOM, Sep. 20, 2021, https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times.
[112] *Annual Copyright License*, Copyright Clearance Center, (2023) https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf.
[113] *What We Do*, N.Y. TIMES, [n.d.], https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, WASH. POST, [n.d.], https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).
[114] Alex Konrad, *OpenAI Releases First $20 Subscription Version of ChatGPT AI Tool*, FORBES, Feb. 1, 2023, https://www.forbes.com/sites/alexkonrad/2023/02/01/openai-releases-first-subscription-chatgpt/?sh=b4debac7f5f1); *see also* Mark A. Lemley & Brian Casey, *Fair Learning*, 99 TEX. L. REV.743, 746 (2021) ("[ML] systems . . . rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that.").
[115] *Warhol*, 598 U.S. 508; *Chapman v. Nicki Minaj*, 2:18-CV-09088 (C.D. CAL. OCT 22, 2018).

practices of "data laundering" and 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.[116]

The Office has previously addressed similar questions, including in connection with the triennial Section 1201 rulemaking and regulatory implementation of a "noncommercial use" exception under the new protection for pre-1972 sound recordings established by the Music Modernization Act.[117] Even when limited TDM uses for non-consumptive, academic research purposes were contemplated in the Section 1201 triennial rulemaking, the Register of Copyrights noted that the "case law has not established that all copying of works for the purpose of TDM is necessarily a fair use." The Office further noted that the exemption request was based on representations that the ingested text would only be accessible for purposes of verifying research findings (and not to analyze or view the works for any other purpose). It therefore appears that some of the generative AI products on the market now, with their copying of expression for expression's sake and the ability to produce paraphrased and in some cases near-verbatim outputs, go beyond the proposal before the Office at that time. In that rulemaking, the Register also required that academic institutions employ substantial security measures to limit access to the corpus of circumvented works only to other researchers affiliated with qualifying institutions for purposes of collaboration or the replication and verification of research findings, and that the circumvention of technical measures for research purposes only be allowed "on copies of the copyrighted works that were lawfully acquired and that the institution owns or for which it has a non-time-limited license," not including renting or borrowing.

iv. There is no satisfactory independent justification for the copying.

There is no independent reason why generative AI models must ingest valuable copyright-protected expressive works apart from the desire to incorporate that very expression. While GAI developers may prefer to copy such high-quality media unburdened from any licensing obligations, some of the very companies that have infringed the copyrighted content of N/MA members have licensed content from others for similar purposes. For example, Stability AI and Meta have launched text-to-music generators trained solely on licensed musical works and

[116] Relatedly, similar logic, as well as considerations of secondary liability and agency principles, may be relevant to the Office's question with respect to entities that collect and distribute copyrighted material for training but may not themselves engage in the training.

[117] *See* Noncommercial Use of Pre-1972 Sound Recordings That Are Not Being Commercially Exploited, 84 Fed. Red. 14242 (Apr. 9, 2019); Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 86 Fed. Red. 59,627 (Oct. 28, 2021).

sound recordings,[118] and Google is in discussions to develop a similar tool using music licensed from Universal Music Group.[119] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[120] Others seems to be trying to get this right from the start. Adobe Firefly is a text-to-image generator trained solely on Adobe Stock images, openly licensed content, and public domain content.[121] Getty has developed a text-to-image generator trained solely on licensed images.[122]

In an acknowledgment that generative GAI development can continue and flourish without training LLMs on unauthorized copies, Google recently announced a new mechanism, Google-Extended, which will allow website publishers to opt out of having their content used to improve the company's AI models in the future while maintaining access to such content through Google Search.[123] OpenAI has similarly announced that internet sites can now block OpenAI's GPTBot and keep their sites out of ChatGPT.[124] In addition, this "opt-out" approach is antithetical to U.S. copyright law (and does not allow for opt-out of the content already scraped). There is also a wealth of material in the public domain or available under open licenses available for the LLMs to use to build their models.

Notably, NMA members stand ready to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content, something

[118] Daniel Tencer, *Stability AI Launches Text-to-Music Generator Trained on Licensed Content Via a Partnership with Music Library AudioSparx,* Music Business Worldwide, Sep. 14, 2023, https://www.musicbusinessworldwide.com/stability-ai-launches-text-to-music-generator-trained-on-licensed-content-via-a-partnership-with-music-library-audiosparx/; Justinas Vainilavicius, *Meta Releases Music Generator Called MusicGen*, Cybernews, Aug. 3, 2023, https://cybernews.com/tech/meta-releases-music-generator-musicgen/.

[119] Hibaq Farah, *Google and Universal Music Working on Licensing Voices for AI-Generated Songs*, The Guardian, Aug. 9, 2023, https://www.theguardian.com/technology/2023/aug/09/google-and-universal-music-working-on-licensing-voices-for-ai-generated-songs.

[120] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock . . . Including Its Music Libraries*, Music Business Worldwide, Jul. 12, 2023, https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[121] *Firefly FAQ for Adobe Stock Contributors*, Adobe (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

[122] Emilia David, *Getty Made an AI Generator that Only Trained on its Licensed Images*, The Verge, Sep. 25, 2023, https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch.

[123] Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, The Verge, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.

[124] Emilia David, *Now You Can Block OpenAI's Webcrawler*, The Verge, Aug. 7, 2023, https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai.

that will benefit all interested parties and society at large, rather than engage in litigation to protect their rights.

In this setting, the developers' goal to create LLMs or to employ those models to power generative AI products, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression. Sam Altman, the founder of OpenAI, and Brad Smith, President of Microsoft, each acknowledged this point in their recent testimony before Congress, explaining that creators of expressive works deserve to control the rights to, and must benefit from, their creations.[125]

Indeed, courts have long recognized that such generalized fair use justifications should not be used to insulate widespread infringement. *American Geophysical Union v. Texaco, Inc*., for example, found that Texaco's photocopying of scientific journals for purposes of commercial R&D was not a fair use, even where the company had made the copies to enrich their researchers' knowledge, because the company was engaged in a "systematic process of encouraging employee researchers to copy articles so as to multiply available copies while avoiding payment."[126] As the court explained:

> The purposes illustrated by the categories listed in section 107 refer primarily to the work of authorship alleged to be a fair use, not to the activity in which the alleged infringer is engaged. Texaco cannot gain fair use insulation for [its employee]'s archival photocopying of articles (or books) simply because such copying is done by a company doing research. It would be equally extravagant for a newspaper to contend that because its business is "news reporting" it may line the shelves of its reporters with photocopies of books on journalism or that schools engaged in "teaching" may supply its faculty members with personal photocopies of books on educational techniques or substantive fields. Whatever

[125] *Oversight of A.I.: Rules for Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/ (statement of Sam Altman) ("we think that creators deserve control over how their creations are used and what happens sort of beyond the point of, of them releasing it into the world . . . . we think that content creators, content owners, need to benefit from this technology . . . . We're still talking to artists and content owners about what they want. I think there's a lot of ways this can happen, but very clearly, no matter what the law is, the right thing to do is to make sure people get significant upside benefit from this new technology. And we believe that it's really going to deliver that. But that content owners likenesses people totally deserve control over how that's used and to benefit from it."); *Oversight of A.I.: Legislating on Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-us-senate-judiciary-hearing-on-oversight-of-a-i/ (statement of Brad Smith) ("generally I think we should let local journalists and publications make decisions about whether they want their content to be available for training or grounding and the like. And that's a big topic and it's worthy of more discussion. And we should certainly let them, in my view, negotiate collectively because that's the only way local journalism is really going to negotiate effectively.").

[126] 60 F.3d 913, 920 (2d Cir. 1994).

> benefit copying and reading such books might contribute to the process of "teaching" would not for that reason satisfy the test of a "teaching" purpose.[127]

This principle applies in full force to generative AI development. While developers have contended that their unlicensed use of material for LLM training and generative AI development purposes is justifiable because the LLMs ingest the copyrighted content to "learn" from the content, just like a human being, no one is allowed to copy an underlying work just because they have an alleged good reason to read the underlying document but don't want to buy (or otherwise lawfully access) a copy. As one scholar explains:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. 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 . . . . A teacher who copies to broaden his personal understanding is a productive consumer, but he nonetheless must pay for the works he consumes. If the teacher's consumption of copyrighted works inspires him to create new scholarship, so much the better, but his subsequent productivity does not entitle him to a refund for the works that influenced him. In much the same way, machine learning makes consumptive use of copyrighted materials in order to facilitate future productivity. If future productivity is no defense for unauthorized human consumption, it should not excuse robotic consumption, either.[128]

[127] *Id*. at 924; *see also Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1263-64 (11th Cir. 2014) ("[A]llowing some leeway for educational fair use furthers the purpose of copyright by providing students and teachers with a means to lawfully access works . . . . But, as always, care must be taken not to allow too much educational use, lest [the court] undermine the goals of copyright by enervating the incentive for authors to create the works upon which students and teachers depend."); *Princeton Univ. Press v. Mich. Document Servs., Inc*., 99 F.3d 1381 (6th Cir. 1996) (reproduction of significant portions of copyrighted works for use in course packets is not fair use); *Marcus v. Rowley*, 695 F.2d 1171 (9th Cir. 1983) (same for teacher's educational booklet); H.R. Rep. No. 94-1476, at 66-67 (1976), https://www.copyright.gov/history/law/clrev_94-1476.pdf ("[A] specific exemption freeing certain reproductions of copyrighted works for educational and scholarly purposes from copyright control is not justified."); Linda Starr, *Is Fair Use a License to Steal?*, Education World, May 25, 2010, https://www.educationworld.com/a_curr/curr280b.shtml#:~:text=The%20fair%20use%20doctrine%20is,and%20scholarship%2C%20and%20classroom%20instruction.

[128] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J. L. & Arts 45, 73-74 (2017); *id*. at 74 (suggesting "a constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose (quoting *Sony Corp. of Am. v. Universal City Studios, Inc*., 464 U.S. 417, 455 n.40 (1984))).

Of course, LLM machines are not humans. As explained in the White Paper, they do not "learn"—they copy, and they do so on a massive scale that no human could replicate. Because a market exists to provide high quality publisher content for purposes such as AI training, the goal of building LLMs does not justify the unlicensed copying of N/MA members' expressive works.

The Copyright Office asks specifically how recent Supreme Court precedent is relevant to this analysis. N/MA has incorporated the *Warhol* decision throughout its analysis. With respect to *Google v. Oracle,* the Court repeatedly took precautions to limit its reasoning to the specific software code at issue and, as a result, is not directly relevant.[129] Indeed, by beginning the analysis with factor two, and emphasizing the inherent functional nature of computer programs, the opinion is grounded in a very different factual surrounding than generative AI training, which ingests publisher works that have been repeatedly described by the Court as expressive works protected by copyright. [130] The limited applicability of *Google v. Oracle* to this instance is shored by *Warhol,* which, as noted, explains that the degree of transformation depends on the specific use.[131]

v. The unlicensed use of training materials serves a system designed to produce substitutional outputs.

LLMs are designed to produce outputs that can substantially copy from, compete with, and substitute for original text content. Even in the furtherance of new technological development, no court has held fair the copying of content to develop a system whose purpose is to substitute for the original works. Rather, cases holding "fair" the use of copyrighted materials to develop a new technology or further a technological purpose are grounded on findings that the ultimate use *did not* compete with the copyrighted works. The first fair use factor does not require news and media publications to be mined to fuel their replacements.

In *Authors Guild*, for example, the court found that neither of the challenged uses (for "search" and "snippets") could provide a meaningful substitute for the copied books and instead were likely to help users identify books of interest.[132] It concluded that if the snippets were arranged into a coherent aggregate "manner and order" (which the challenged system disallowed) "that would raise a very different question beyond the scope of our inquiry."[133] Similarly, in *Kelly v.*

[129] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021).
[130]*Harper & Row,* 471 U.S. at 556-557; *Feist Publications, Inc. v. Rural Telephone Serv. Co*., 499 U.S. 340, 349 (1991)
[131] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021).
[132] 804 F.3d at 218.
[133] *Id.* at 223.

Arriba Soft Corp., the court found that the search engine "Arriba's use of Kelly's images in its thumbnails does not harm the market for Kelly's images or the value of his images."[134]

In contrast, LLMs can and do generate outputs that replicate or closely paraphrase the original expressive works. Consumer-facing chatbot services built around these models, including those integrated into search engines like Bing or Google, are well poised to directly substitute for publishers and to usurp their valuable relationships with readers of news, magazine, and web content. Marketing for these new features makes clear that they are intended to create substitutional narratives that can substantially copy from, compete with, and substitute for the primary expressive material. Unchained from constraints to serve as no more than an electronic reference or bridge to a primary source, narrative search results can provide users with sufficient content (full key portions and highlights of expressive content), that substitutes for any need to read the original. As a recent New Yorker article explains, the "goal" of "large language models, like OpenAI's ChatGPT and Google's Bard" "is to ingest the Web so comprehensively that it might as well not exist."[135]

These chatbot search uses thus go well beyond the nuanced reasoning and careful guardrails established by cases like *Authors Guild* and *Kelly* and into competitive, consumptive uses that are distinctly unfair to content owners. Indeed, courts routinely dismiss fair use arguments for new digital products that have a similar purpose to, and could supplant, the original work.[136] That reasoning applies here.

[134] 336 F.3d at 821; *see also Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021) ("*Oracle*") (jury's fair use determination barred Oracle from "overcom[ing] evidence that, at a minimum, it would have been difficult for Sun [Oracle's predecessor] to enter the smartphone market" even without Google's alleged infringement, including Sun's former CEO's testimony that Sun's failure to build a smartphone was not attributable to Google's alleged infringement); *cf. Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417, 456 (1984) (noting that plaintiffs "failed to demonstrate that time-shifting would cause any likelihood of nonminimal harm to the potential market for, or the value of, their copyrighted works.").

[135] James Somers, *How Will A.I. Learn Next?*, THE NEW YORKER, Oct. 5, 2023, (reporting that the number of new posts the website Stack Overflow, where computer programmers went to ask and answer programming questions, has decreased by 16% since the debut of ChatGPT).

[136] *See*, *e.g.*, *Fox News Network, LLC v. TV Eyes, Inc*., 883 F.3d 169, 177, 181 (2d Cir. 2018) (media monitoring service, while transformative, was not fair, because it usurped plaintiff's market); *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, *18-25 (S.D.N.Y. Mar. 24, 2023) (Internet Archive's electronic copying and unauthorized lending of 3.6 million books protected by valid copyrights is not a fair use because it competed with plaintiff's licensing market); *Associated Press v. Meltwater U.S. Holdings, Inc*., 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (crawling of various websites for Associated Press's stories and scraping "snippets" of those stories for use in notifying and informing Meltwater's own customers of certain stories directly competed with the Associated Press such that Meltwater's copying would deprive the Associated Press of a stream of income to which it was entitled).

Moreover, recent lawsuits have alleged that certain LLMs knowingly ingested material from notorious pirate sites, and publishers have used their terms of service or using technical measures like robots.txt to prohibit crawling for purposes of generative AI ingestion.[137] If 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, making a fair use defense less likely to succeed. This concept—that to invoke fair use, an individual must possess an authorized copy of a work—was addressed by the Court in *Harper & Row Publishers Inc. v. Nation Enterprises*, which confirmed that "[f]air use presupposes good faith" and found that Nation acted in bad faith because it "knowing exploited a purloined manuscript."[138] The Federal Circuit expanded on the concept in *Atari Games Corp. v. Nintendo of America, Inc.*, finding that because Atari gained access to an unauthorized copy of the Nintendo's source code by submitting false information to the U.S. Copyright Office, "any copying or derivative copying…does not qualify as a fair use."[139]

For these reasons, it is likely that with respect to LLMs, the first factor favors a finding of infringement and not fair use.

The effect of generative AI copying on the market for publisher content is predictable and real (fourth factor).

The fourth fair use factor directs courts to consider "the effect of the use upon the potential market for or value of the copyrighted work."[140] The focus is on whether widespread conduct like the conduct of the alleged infringer "would adversely affect the potential market for the copyrighted work," including market harm to the original work and to derivative works.[141] While the examination of potential markets is not without limit, "traditional, reasonable, or likely to be developed markets" are considered.[142] As the *Texaco* court recognized, "[i]t is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for

[137] *See Complaint, Authors Guild v. OpenAI*, at paras. 97-110 (Sep. 2023) available at https://authorsguild.org/app/uploads/2023/09/Authors-Guild-OpenAI-Class-Action-Complaint-Sep-2023.pdf; *Complaint, Tremblay v. OpenAI*, at paras. 31-34 (Jun. 28, 2023) available at https://torrentfreak.com/images/authors-vs-openai.pdf; *Complaint, Chabon v. Meta Platforms, Inc*. at paras. 26-39 (Sep. 12, 2023) available at https://fingfx.thomsonreuters.com/gfx/legaldocs/lbpgolxxmpq/META%20AI%20COPYRIGHT%20LAWSUIT%20complaint.pdf; Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, THE VERGE, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.
[138] 471 U.S. 539, 547 (1985).
[139] 975 F. 2d 832 at 843(Fed. Cir. 1992).
[140] 17 U.S.C. § 107(4).
[141] *Harper & Row*, 471 U.S. at 566, 568 (emphasis omitted).
[142] *Am. Geophysical Union v. Texaco, Inc*., 60 F.3d 913, 929-30 (2d Cir. 1994).

licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor."[143]

GAI's unauthorized use of copyrighted material harms the market in two ways.

First, with respect to inputs, generative AI developers' unauthorized use of publisher content to build their LLMs deprives publishers of an available licensing market, such that the fourth factor also should favor a finding of infringement when publisher content is used without authorization for training purposes.[144]

While developers complain that it is unworkable to license content for their ingestion needs,[145] there is a long history of publishers licensing their content for a variety of uses and licensing deals, and negotiations are occurring in the open market specifically for GAI uses, as documented *infra*.

As explained above, including in response to questions 6 and 10, and in discussion of the first factor, there is also a long history of media organizations and associations licensing their content for a variety of uses, including to media monitoring entities, to LEXIS, and through the CCC.

Examples also abound, both here and abroad, of collective licensing of copyrighted content, and these models demonstrate the paths that exist for efficient licensing frameworks to meet AI needs. CCC, for example, was formed by authors, publishers, and users to facilitate "centralized licensing of text-based copyrighted materials," and it has grown to represent copyright holders from nearly every country, with access to millions of sources.[146] Outside the

[143] *Id*. at 929 (citation omitted).

[144] *Texaco*, 60 F.3d at 930 (finding fourth factor favored a finding of infringement where the challenged photocopying harmed an existing "workable market for institutional users to obtain licenses for the right to produce their own copies of individual articles via photocopying"); *see also Fox News Network, LLC v. TVEyes, Inc*, 883 F.3d 169, 180 (2d Cir. 2018) (by using content without payment, Fox was deprived of "licensing revenues from TVEyes"); *Davis v. Gap, Inc*, 246 F.3d 152, 175-76 (2d Cir. 2001) (freely taking a copyrighted work allowed defendant to avoid "paying the customary price," that plaintiff "was entitled to charge" for use of work, and that, as a result, plaintiff "suffered market harm through his loss of the royalty revenue to which he was reasonably entitled in the circumstances, as well as through the diminution of his opportunity to license to others").

[145] OpenAI, LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

[146] Comments of Copyright Clearance Center, Inc. 79 Fed. Reg. 2696 (Mar. 3, 2024), https://www.copyright.gov/docs/recordation/comments/79fr2696/CCC.pdf; *Annual Copyright License*, Copyright Clearance Center.

United States, collective management organizations broadly manage news and media licensing, such as NLA Media Access in the U.K.[147]

Second, it is indisputable that generative AI output is intended to, and does, substitute for human-generated content, including publisher content.[148] As explained above, already less than 65% of searches result in clicking through to the underlying source.[149] That percentage is only going to increase with narrative search results. Indeed, marketing experts expect click-through rates for generative search responses to be even lower than already declining rates for organic results.[150] "Particularly for informational searches, Google will aggregate (or flat-out plagiarize) from the search results and give users much of what they're looking for."[151] "Users may find all the information they need directly on the search page, so there's no need to click on the source

[147] Tarja Koskinen-Olsson, *Collective Management of Text and Image-Based Works*, WIPO (Updated 2023) https://www.wipo.int/edocs/pubdocs/en/wipo-pub-924-2023-en-collective-management-of-text-and-image-based-works.pdf; *A Guide to Media Monitoring and Corporate Licensing*, Press Database and Licensing Network, at 14 (Oct. 2017), https://static1.squarespace.com/static/5eca9a7fe349354c54ae6cab/t/5ef2b3025a06263ec1a24a14/1592963847770/pdln_guide+to+corporate+and+mmo+licensing.pdf; *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.
Collective licensing has also flourished in the music industry, further demonstrating the potential to develop efficient, large-scale licensing models for GAI needs. The performing rights organizations (PROs) such as ASCAP, BMI, and SESAC license the right to publicly perform musical compositions on behalf of copyright owners. PROs collectively "cover[] almost all of the millions of songs currently copyright protected," and they operate by offering "blanket authorization to use the music [each organization] represents in exchange for license fees," which are then distributed "as royalties to its affiliated songwriters, composers, and music publishers." *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.

[148] *See also*, *e.g*., Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the USPTO, at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("Writers who were employed to perform formulaic composition might be able to devote their energies to more creative forms of self-expression *once machines supplant them.*" (quoting Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 80 (2017))); Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 767 (2021) (Machine learning "empowers [] companies to extract value from authors' protected expression without authorization" or compensation "and to use that value for commercial purposes that may someday jeopardize the livelihoods of human creators." (quoting Sobel, *Artificial Intelligence's Fair Use Crisis*)); *id.* at 777 (AI systems trained "to generate their own expressive works . . . pose a threat of significant substitutive competition to the work originally copied." (internal quotation marks omitted)).

[149] *See* George Nguyen, *Zero-click Google Searches Rose to Nearly 65% in 2020*, Search Engine Land, Mar. 22, 2021, https://searchengineland.com/zero-click-google-searches-rose-to-nearly-65-in-2020-347115.

[150] *See*, *e.g*., Rebecca Krause, *Google's Search Generative Experience (SGE): A Marketer's Guide*, Seer Interactive, Aug. 10, 2023, https://www.seerinteractive.com/insights/googles-search-generative-experience ("As SGE rolls out to more users, the click-through-rate of the ten organic links (even position 1) may lower.")

[151] Dave Shapiro, *Generative AI in Search*,Neil Patel (2023) https://neilpatel.com/blog/generative-ai-in-search/ ("people will find enough of what they need in the SGE and not click on organic results.").

website."[152] As set forth above, no court has deemed fair the copying of expressive works, even at the development stage, for the purposes of eventually competing with and substituting for the original work. The substitutional use of the generative AI outputs is a further reason why the fourth factor favors a finding of infringement with respect to the unauthorized use of publisher content at the training stage.

The effect of generative AI copying at the output stage is self-evident. Where the outputs replicate or closely paraphrase the original expressive works and thus infringe upon and substitute for them, such that users no longer need to connect with or obtain the original works from their original sources, such uses harm the market for the publishers' works.

Generative AI copying takes substantial portions of expressive works in their entirety (second and third factors).

Under the second factor, courts consider whether a work is creative or functional, "recogn[izing] that some works are closer to the core of intended copyright protection than others."[153] The second factor is typically less important than the first and fourth factors.[154]

Although news, magazine, and digital media content includes underlying facts, the reporting seeks to determine which facts are significant and to recount them in an interesting manner, and they are thus creative in nature.[155] Such content also extends well beyond traditional news reporting and includes pieces devoted to opinion and analysis. Here, where developers copy publisher content so that LLMs can best mimic human speech,[156] the copying is necessarily exploiting the content for its expressive qualities and the second factor favors a finding of infringement for both inputs and outputs.

The third factor evaluates both the quantity and quality of the copying, and "examine[s] the amount and substantiality of the portion used in relation to the copyrighted work as a whole,"

[152] Sam Stemler, *9 Things You Need to Know about Google Search Generative Experience (SGE)*, WEB ASCENDER, Aug. 29, 2023, https://www.webascender.com/blog/9-things-you-need-to-know-about-google-search-generative-experience-sge/.

[153] *Campbell*, 510 U.S. at 586; *Oracle*, 141 S. Ct. at 1202.

[154] *Authors Guild*, 804 F.3d at 220.

[155] *See Harper & Row*, 471 U.S. at 547 ("Creation of a nonfiction work, even a compilation of pure fact, entails originality."); *see also Authors Guild*, 804 F.3d at 220 ("Those who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports."); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 178 (2d Cir. 2018) (rejecting argument that, since facts are not copyrightable, the factual nature of a creative compilation favors a finding of fair use).

[156] *See* N/MA, WHITE PAPER at p. 8, 21-22 (2023), Appendix A; Stephen Wolfram, *What Is ChatGPT Doing ... and Why Does It Work?*, Feb. 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

including whether the “heart” of the work is copied.[157] “[T]he fact that a substantial portion of the infringing work was copied verbatim is evidence of the qualitative value of the copied material, both to the originator and to the plagiarist who seeks to profit from marketing someone else’s copyrighted expression.”[158] The massive scale of copying also favors a finding of infringement.[159]

Here, for inputs, the developers copy all or substantial portions of the publisher content during the course of LLM training and development of generative AI tools, and it is reasonable to conclude that the “heart” of the work is copied. Moreover, copying for generative AI development can be viewed as excessive given the degree to which the copies usurp the available licensing market.[160]

Application of the third factor at the output stage must be evaluated on a case-by-case basis, depending on the portions of the works which the outputs copy. Suffice to say, the third factor will favor a finding of infringement at the output stage whenever the outputs copy sufficient portions or the heart of the copied works.

Question 8.4 asks whether the volume of material used to train an AI model affects the fair use analysis. Because LLMs and other generative AI models ingest a large amount of material, it does not appear necessary to ingest any one particular work. This fact would weigh against fair use, because there are other ways to develop a model beyond taking a particular work. By contrast, courts have found fair use favored when copying was “necessary” to gain access to functional elements of computer software,[161] and the Copyright Office has considered whether a potential licensing market exists when determining whether proposed uses of audiovisual clips for documentary filmmaking is likely to be fair.[162]

And the taking of a copyrighted work is not more likely to be fair because the allegedly infringing act also incorporated other material that was not infringing. As the *Harper & Row* Court explained:

[157] *Harper & Row*, 471 U.S. at 564-65).
[158] *Id*. at 565.
[159] *See*, *e.g.*, *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, at *8 (S.D.N.Y. Mar. 24, 2023) (“Unlike Sony, which only sold the machines, IA scans a massive number of copies of books and makes them available to patrons . . . .”).
[160] *Campbell*, 510 U.S. at 587-88; *see also* N/MA, White Paper at p. 37 (2023), Appendix A.
[161] *Sony Computer Ent. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises, Ltd. v. Accolade, Inc.*, 977 F. 2d 1510 , 1529 (9th. Cir. 1992);
[162] *See, e.g.,* Acting Registrar of Copyrights, Recommendation: Section 1201 Rulemaking: Seventh Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention, at 60- 61(Oct. 2018) available at https://cdn.loc.gov/copyright/1201/2018/2018_Section_1201_Acting_Registers_Recommendation.pdf.

> As the statutory language indicates, a taking may not be excused merely because it is insubstantial with respect to the *infringing* work. As Judge Learned Hand cogently remarked, "no plagiarist can excuse the wrong by showing how much of his work he did not pirate." *Sheldon* v. *Metro-Goldwyn Pictures Corp.,* 81 F. 2d 49, 56 (CA2), cert. denied, 298 U. S. 669 (1936).[163]

The proper question is how much of the copyrighted work has been used by the infringer in creating a secondary work. In this case, not only does it appear that generative AI developers have copied and used entire protected individual works, they have likely copied the entire corpora of our members' newspapers, magazines, and websites. A systemic disregard or carelessness towards copying large volumes of expressive works looks different than the targeted taking of a specific individual work, and should disfavor a finding of fair use.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)? 9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? 9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses? 9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners? 9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

N/MA responds to question 9 and subparts 9.1-9.4 together.

The starting point to answer this question is the Copyright Act, which provides rightsholders with a bundle of rights that may be employed to provide necessary authorization for the use of copyrighted works absent applicable exceptions or defenses. That is, the existing law is "opt in." Consent can be provided by various means, which is also outlined by legal doctrines, but the general principle in copyright is to require affirmative consent absent an applicable exception or limitation. Changing 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.

[163] *Harper & Row Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 547 (1985).

Discussions around opt-out are more relevant in countries and regions that, unlike the United States, may already have a statutory text and data mining (TDM) exception that allows some or all users to engage in TDM for limited AI training purposes. It is rare indeed to have a sweeping exception for TDM that extends to highly commercial uses without the ability to opt out. In addition to raising other potential concerns, including compliance with international agreements, retaining the ability to opt out of such exceptions is important in those countries or regions. The United States, however, has not adopted an exception to our copyright laws for TDM. N/MA opposes the creation of a new or expanded exception to copyright law that would change the status quo to permit AI training without the rightsholder's authorization.

To date, current tools present a Potemkin village of a solution, providing limited benefits to publishers while creating a patina of responsibility to justify positions that copying is legal absent affirmative opt-out. It is inappropriate industrial policy to place the burden on a copyright owner to remedy a potentially infringing act, rather than on a generative AI developer or deployer who already possesses the right and ability to control what material is used for training (whether by selecting, cleaning, or fine tuning a dataset, licensing content, or by paying a low wage to someone overseas to mitigate the worst violations). And the necessary act of choosing what copyrighted works an AI system is trained on distinguishes these developments from the architectures that gave rise to the section 512 safe harbor.

To be sure, there may be limited room for voluntary signals or solutions that may simplify licensing. This is particularly the case as publishers and other rightsholders explore reasonable technical and collective licensing solutions in response to developments in other parts of the world, including the EU's Directive on Copyright in the Digital Single Market.[164] These measures should be industry-led and agreed to by rights holders of particular sectors to prevent very large platforms from imposing methods on publishers and ensure a workable framework for all. The government could play a limited role in facilitating these conversations and in ensuring compliance with the voluntary measures, potentially by imposing penalties for generative AI developers who fail to honor such opt-in or opt-out measures or protocols.

A voluntary opt-in system could be aided by a publisher-led collective licensing entity or a technical measure that allows publishers to signal to generative AI developers that their content is available for AI training purposes, subject to any relevant terms. Such a solution could lower the burden of acquiring licenses for developers—including retroactively for content

[164] 2019, O.J. (Directive (EU) 2019/790) The European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, at Art. 3 & 4, available at https://eur-lex.europa.eu/eli/dir/2019/790/oj.

that has already been scraped for existing applications—while making it easier for publishers to scale up their licensing to multiple licensors, thus facilitating increased choice.

Meanwhile, it is important to distinguish opt-out signals for scraping from copyright licenses, express or implied. For example, some uses of content may constitute fair use (such as certain uses for search purposes) and the owner of the site may wish to signal no more than that scraping is permitted for such purposes. In some cases, the site owner may not be in a position to authorize certain uses—either because they may not have or may not know the full scope of the rights it controls to all of the content on the site for every possible use.[165] This is a common situation for many publishers who make available to the public a wide variety of content, some of which is work-for-hire, some of which was created for the publisher by independent third parties, and some of which is licensed. In other cases, the site owner may wish to authorize access to its sites only for certain limited purposes and only to certain authorized parties for commercial reasons. Any automatic signal, whether opt-in or opt-out, must account for these differences.

One of the most widely advocated exclusion protocols, robots.txt, is currently a blunt tool that does not offer sufficient granular control over the types of uses for which scraping is allowed. As a result, site owners are forced to choose between authorizing ***any*** use and authorizing no uses. Media publishers often depend on search to generate a significant part of their traffic. Consumers also depend on search to locate material online. Blocking ***all*** scraping would eliminate this important source of traffic, but permitting scraping by not including the robots.txt signal certainly doesn't extend permission for developers to make ***all*** potential uses of the content. The availability of the robots.txt signal is insufficient to solve this problem. Robots.txt is a voluntary regime and many scrapers disregard the signal. It would be helpful to both developers and rightsholders for scrapers to honestly and transparently identify the entity that is scraping, and abide by industry standard licenses that can be identified automatically, in a signal similar to robots.txt. N/MA would support the Copyright Office facilitating discussions on voluntary opt-out signals, while ensuring that scrapers have incentives to respect them. Incentives could include potential legal penalties for scrapers who disregard such signals, or fail to provide truthful information regarding their identity and the uses to which the scraped material will be put. Such a standard could be developed by industry, with the backstop of having the conditions—transparency and an obligation to follow the rule—enforceable by law.

[165] This is especially true for publishers who have accumulated content over decades on many iterations of contracts, sometimes in the tens of thousands, which would need to be reviewed for legal compliance with a new and developing use.

Publishers are also concerned that opt-out systems must be efficient at scale. In an opt-out regime, developers may have incentives to make opt-out difficult for publishers (or at least, expend the minimum compliance efforts required), whereas with an opt-in regime, developers are incentivized to seek efficient licensing solutions. For example, DALL-E's opt-out system requires the "owner or rights holder . . . to submit *an individual copy of each image* they'd like removed from DALL-E's training dataset, *along with a description.*"[166] This is obviously impractical for more than a *de minimis* number of images. The Copyright Office need only recall its years-long DMCA study to predict the difficulties with this system.

Moreover, an opt-out regime puts the burden on copyright owners to find out who is using their material. Not only does this incentivize non-disclosure, but developers commonly train their systems on material acquired from sites that have been identified by the U.S. government as notorious markets for piracy,[167] necessitating that copyright owners enforce rights against infringers as a prerequisite—a burden that is impossible to achieve.

With respect to question 9.1, concerning non-commercial and commercial uses, that question may be better evaluated in the context of evaluating infringement or fair use, *see infra*. But with respect to signaling consent, or the lack thereof, by opting in or out of AI training, it is difficult to make blanket exceptions or judgments based on the identity of the user or category of use. For example, content scraped for a seemingly noncommercial use—potentially at the request or with the support of a commercial developer—can and often is passed onto a commercial entity that may create products or services that directly compete with the content creator. At that point, it becomes much more difficult for a rightsholder to "opt in" to or "opt out" of a use that has already occurred. The prevalence of data laundering, as well as the lack of bright lines distinguishing commercial from noncommercial uses with these technologies, makes this question difficult to answer on a black and white basis.

While objecting to any mandatory opt-out requirement, N/MA would support the Copyright Office facilitating discussions on voluntary opt-out signals, while ensuring that developers have incentives to respect them, potentially by imposing penalties for disregarding such signals.

[166] Kali Hays, *OpenAI Offers a Way for Creators to Opt Out of AI Training Data. It's so Onerous that One Artists Called it 'Enraging'*, Google: Insider, Sep. 29, 2023, https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9?r=US&IR=T.

[167] *See* Kevin Schaul *et al., Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart,* The Washington Post (Apr. 19, 2023), *available at* https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ (last visited Oct. 28, 2023); Alex Hern, *Fresh Concerns Raised Over Sources of Training Material for AI Systems,* The Guardian (Apr. 20, 2023), *available at* https://www.theguardian.com/technology/2023/apr/20/fresh-concerns-training-material-ai-systems-facist-pirated-malicious# (last visited Oct. 28 2023).

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

The Copyright Act and the relevant case law sets out clear rules for how to handle works made for hire and assigned copyrights. There is no need for special rules for AI in this respect. While N/MA expresses no opinion as to aspects of this question that implicate other rights, including moral, privacy, and contracts, there is no need to revisit this principle under copyright law.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained? 10.1. Is direct voluntary licensing feasible in some or all creative sectors? 10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses? 10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed? 10.4. Is an extended collective licensing scheme a feasible or desirable approach? 10.5. Should licensing regimes vary based on the type of work at issue?

N/MA responds to the licensing issues posed by question 10 and its subparts together. Our response focuses on the landscapes and dynamics experienced by our members.

As discussed above and for clarity, commercial generative AI companies do need consent to use our members' content under existing law. The Notice asks many questions around how emerging licensing frameworks can respond and adapt to continued innovations in generative AI technologies. A rights-based regime is best suited to answer these questions flexibly, through direct negotiations among the affected parties. While N/MA can't speak for other creative sectors, we certainly believe that voluntary licensing is feasible—and the most desirable—for publishers of newspapers, magazines and digital media content.

The Office Should Encourage Market-Based License Solutions and Reject Calls for Compulsory Licensing

Marketplace licensing, including on a collective basis where appropriate, is the default legal system under U.S. law and should be the default here. Voluntary licensing is especially preferred here, where generative AI technologies are so new, the uses of AI so unpredictable, and the economics so unknown, that it is imperative that publishers and AI developers be given maximum flexibility in structuring (and restructuring) deals as the marketplace evolves.

The nascency of generative AI is already spawning varied companies, products, and services that will have different economic implications for authors, copyright owners, and their businesses.[168] There is unlikely to be a one-size-fits-all solution to licensing copyrighted material for ingestion and other uses by generative AI-dependent entities.

At a moment when marketplace actors are interested in negotiating private arrangements, the Copyright Office should firmly reject calls to establish a compulsory license to permit copyrighted content to be ingested into AI systems under government-set terms.[169]

This view is most consistent with the international copyright legal framework and the longstanding views of the Copyright Office itself. As former Register of Copyrights Marybeth Peters testified to Congress regarding the section 115 license for musical works, compulsory licensing is a "last resort mechanism," typically only seen where there has been a failure of voluntary agreements.[170] As she further explained, "[a] compulsory license limits an author's bargaining power. It deprives the author of determining with whom and on what terms he wishes to do business."[171] For that reason, as the Office explained in connection with a recommendation to sunset the section 119 license for satellite distant signals, "[h]istorically, the Copyright Office has supported statutory licenses only when warranted by special

[168] The Glossary appended to the Office's Notice and variety of definitions offered by policymakers in the EU, U.S., UK, and other markets to categorize obligations by differing types of AI-related actors illustrates this shifting landscape.

[169] It would similarly be premature for the Office to support calls for extended collective licensing (ECL) models. *See* U.S. Copyright Office, Request for Comments on Artificial Intelligence and Copyright, at 2. (Aug. 30, 2023) 88 FR 59942, available at 10.4 https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright.

[170] *See* Section 115 Compulsory License: Hearing before the Subcom. On Courts, the Internet and Intell. Prop. of the H. Comm. on the Judiciary, 108th Cong. (Mar. 11, 2004) (statement of Marybeth Peters, Register of Copyrights) ("[U[se of the compulsory license should only be made as a last resort, and that licensees should be encouraged to obtain voluntary licenses directly from the copyright owners or their agents, who would offer more congenial terms.") available at: https://www.copyright.gov/docs/regstat031104.html. *See also* U.S. Copyright Office, Copyright and the Music Marketplace at 112 (Feb. 2015) available at https://www.copyright.gov/policy/musiclicensingstudy/copyright-and-the-music-marketplace.pdf (compulsory licensing "removes choice and control from all copyright owners that seek to protect and maximize the value of their assets."); European Commission, A Single Market for Patents: New Rules on Compulsory Licensing (April 2023) available at: https://single-market-economy.ec.europa.eu/system/files/2023-04/Patent%20Package_Compulsory%20Licensing_Final.pdf ("Compulsory licensing is a last resort mechanism which allows a government to authorise the use of a patented invention without the consent of the patent holder").

[171] *Id.* Statement of Marybeth Peters, Section 115 Compulsory License: Hearing (2004).

circumstances and only for as long as necessary to achieve a specific goal."[172] And the Office reaffirmed this position again in 2019:

> A statutory license creates an artificial, government-regulated market that operates as an exception to the general rule that copyright owners hold exclusive rights and can negotiate whether and how and at what cost to distribute their copyrighted works; statutory licenses tend to be below the fair market value.[173]

In addition to denying copyright owners the freedom to license as they see fit, a compulsory license would risk ossifying the innovative potential for generative AI technologies. Because a statutory license must clearly set out the scope and terms of the license, it is unlikely to be sufficiently flexible and adaptable to serve the legitimate needs of both publishers and AI developers and keep pace with technological and market developments.

It is important that copyright preserve the core function of market-based incentives for humans to create and disseminate works of authorship as generative AI products and services gain further traction. And this is especially important in the case of newspapers, magazines, media websites, and books, where a compulsory licensing regime could create a risk of political interference from Congress or the Executive Branch.

Conditions Exist for a Strong Licensing Ecosystem to Flourish Between Media Publishers and AI Developers and Other Licensees

Unlike the rare exceptions where government-regulated licensing is necessary, there is no evidence of market failure here to support intervention at this time. Media publishers already operate robust existing licensing arms as part of their established businesses. Well-established markets exist for the licensing of archival material and other real-time access to news content, including for use in new products and technologies. In fact, some of the major developers that have copied and used content without permission are already business partners and licensees of N/MA member publishers in connection with other products.[174]

[172] U.S. COPYRIGHT OFFICE, SATELLITE TELEVISION EXTENSION AND LOCALISM ACT: A REPORT OF THE REGISTER OF COPYRIGHTS at 1 (Aug. 29, 2011) available at https://copyright.gov/reports/section302-report.pdf.

[173] U.S. COPYRIGHT OFFICE ANALYSIS AND RECOMMENDATIONS REGARDING THE SECTION 119 COMPULSORY LICENSE; RESPONSE TO HOUSE COMM. ON THE JUDICIARY, 115TH[CK] CONG. at 5 (Jun. 3, 2019), available at https://copyright.gov/laws/hearings/views-concerning-section-119-compulsory-license.pdf.

[174] Sarah Fischer, *Google to Launch News Showcase Product in U.S.*, AXIOS, Jun. 8, 2023, https://www.axios.com/2023/06/08/google-news-showcase-us (describing Google licensing deals with 150 news publishers across 39 states); Ahiza Garcia, *Facebook Offers Media Outlets Millions to License Content, WSJ Reports,* CNN, Aug. 9, 2019, https://www.cnn.com/2019/08/08/tech/facebook-news-outlets-license-rights-content/index.html (describing Facebook offers to license with news publishers).

For generative AI development specifically, and as explained in response to questions 6 and 8, the market is already responding to the demand to provide high quality news and media content, and N/MA members are actively working to grow this field. This licensing for generative AI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy news content.[175]

Voluntary Collective Licensing Can Play a Role in Licensing Works for Generative AI Uses

For media publishing, the marketplace can support both individual, direct negotiated arrangements (between a publisher and a LLM provider or other generative AI company) as well as voluntary collective licensing arrangements. Such a structure can be more agile in response to technological and business developments than a regulated solution, while supporting a competitive marketplace for the affected sectors.

While 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. Collective licensing would benefit competition among LLM providers. Today, the largest LLM providers crawl and index online content to build their corpus of training data. That process is expensive and difficult, requiring massive scale. It thus forms an entry barrier to nascent competitors. A collective licensing entity could aggregate, standardize, and distribute content from smaller publishers, allowing smaller LLM competitors to at least partially bypass the need to crawl and index web content.

Voluntary collective licensing would also not be unusual. Collective licensing entities already exist that satisfy competition law requirements, including reproduction licensing organizations like CCC. Examples of entities and models engaged in licensing other forms of copyrighted works include a society that issues licenses and distributes licensing fees for over 70,000 fine artists (ARS); a licensing entity that issues blanket licenses for worship music to churches, schools and religious organizations (CCLI); a licensing entity that authorizes non-theatrical uses of motion pictures by organizations in child care, education, communal living facilities, corporations, and others (MPLC); entities that offer subscription licenses to millions of images, videos and music created by millions of contributors (Shutterstock, Getty Images, Unsplash, Storyblocks, iStock, 123RF, Vecteezy, Pixabay, Adobe Stock, JumpStory); performing rights organizations in the music industry (GMR and SESAC, as well as ASCAP and BMI, which are subject to consent decrees); an indie label organization that negotiates model licenses with

[175] See *What We Do*, N.Y. TIMES, [n.d.],https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, WASH. POST [n.d.], https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).

streaming services and other commercial users to which individual labels can opt in (Merlin); and a trade association that has negotiated agreements with major commercial users to which individual publishers can opt in (NMPA); among others. Taken together, these organizations exemplify that rights can be licensed efficiently when markets are allowed to develop. Moreover, often the "back office" technological and other infrastructure needs of a licensing entity may be able to be outsourced to existing organizations that already have the necessary capabilities, including entities like SoundExchange, SESAC, CCC, and others.

Indeed, while competition considerations must be approached carefully, courts have upheld various structures and models used to facilitate copyright collective licensing, including on a blanket license basis.[176]

Collective licensing may be particularly well-suited with respect to media publishing and generative AI development. Collective licensing nuances can vary by market, the nature of the works and uses, and the licensor/licensee parties involved. The licensing of media publishing content can be expected to operate differently than other copyright markets, such as the licensing of musical works for digital streaming services (with which the Office is familiar, given its work with the Music Modernization Act). There are a few key differences. First, LLMs require ingesting a large amount of textual content, but there does not appear to be an expectation that LLMs were trained on a "full catalog" of content, making it easier for a licensee to walk away. This is unlike licensing for music streaming services, where consumer expectations that streaming services offer a "full catalog" may factor into licensing negotiations.[177] Second, media publishing does not have the same fragmentation of the rights to be licensed (since news and media publishers typically control the necessary rights for their mastheads, and there is no need to "match" pieces of a textual work in the same way licenses for musical works and sound recordings must each be separately licensed). These features thus reduce the risk that "must have" or hold-out publishers would be able to extract supracompetitive pricing, but, by the same token, increase the attractiveness of voluntary collective models to facilitate licensing of material by smaller publisher operations.

[176] *See, e.g.*, *Buffalo Broad. Co., Inc., v. ASCAP,* 744 F.2d 917, 920 (2d Cir. 1984); *Broadcast Music, Inc. v. Columbia Broadcasting System, Inc.*, 441 U.S. 1, 23 (1979) ("Joint ventures and other cooperative arrangements are . . . not usually unlawful, at least not as price-fixing schemes, where the agreement on price is necessary to market the product at all."). *See also Texaco, Inc. v. Dagher,* 547 U.S. 1 (2006) (holding that internal pricing decisions of a legitimate joint venture are not per se unlawful).

[177] U.K. COMPETITION AND MARKETS AUTHORITY, MUSIC AND STREAMING: FINAL REPORT at 14, 73-74, 76, (2022) *available at* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1120610/Music_and_streaming_final_report.pdf.

The Office Should Encourage the Development of Voluntary Marketplaces

In light of the strong opportunities for voluntary individual and collective licensed solutions to be structured in ways that are pro-competitive, N/MA does not believe the Copyright Office needs to recommend intervention at this stage. That said, it is possible that legislation, such as antitrust exceptions, to augment existing abilities to negotiate collectively could be helpful.

Indeed, in other contexts, antitrust exceptions are strongly needed to correct market imbalances that are harming what the Office has called "the press's essential role in our system of government."[178] The Office has previously noted the potential for changes to competition policy to serve as an effective means to improve the position of press publishers in dealing with news aggregators.[179] N/MA supports the Journalism Competition and Preservation Act ("JCPA"). While the Office previously declined to offer a recommendation with respect to competition policy,[180] in light of the Office's current interest in the interrelation between copyright and competition interests, N/MA urges the Office to follow the logical conclusion of its press publisher study and support competition-based policy changes like JCPA to improve protections for sustaining journalism.

However, with respect to licensing of media content for generative AI uses, it is not clear that such legislation is actually necessary given that many collective licensing entities (some described above) currently operate in accordance with antitrust laws without the need for legislative exceptions.

Given the explosion of commercial LLM products, mostly without obtaining the permission necessary to make use of the content they have taken, licensing for current and future LLM models should be put in place swiftly. For N/MA's part, its members are willing to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content (including for past takings), something that will benefit all interested parties and society at large, and avoid protracted litigation.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model,

[178] U.S. COPYRIGHT OFFICE, COPYRIGHT PROTECTIONS FOR PRESS PUBLISHERS at 4 (2020) ("Press Publishers Study"), available at https://copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf.

[179] *Id.*

[180] *Id.* at 24.

and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

N/MA refers to question 10 above. A rights-based framework exists and is best suited to address these questions flexibility, through market negotiations among the affected parties. This will also allow for different transactions to emerge amongst the user-side licensees, including curators, developers, and deployers of generative AI models.

Developers should be prohibited from ingesting materials for training purposes from sources known to contain pirated content. Such sites should be blocked and prohibited for use in training.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

With respect to training, it is evident that particular works may be more or less valuable for training than other works, and this can be reflected in license pricing and terms. There may also be other relevant terms to be negotiated based on what is technically feasible and valued between licensing partners (e.g., territoriality, output similarity, attribution, etc.).

As explained in the White Paper, and in response to questions 6, 7.1, and 8, media content accounts for a substantial volume of the known sources for LLM training, suggesting that this high quality expressive material is especially desirable by developers. Forensic analysis shows:

- Developers have copied and used news, magazine and digital media content to train LLMs.
- Popular curated datasets underlying LLMs significantly overweight publisher content by a factor ranging from over 5 to almost 100 as compared to the generic collection of content that the well-known entity Common Crawl has scraped from the web.
- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's generative AI-powered products like Bard. Half of the top ten sites represented in the data set are news outlets.
- LLMs also copy and use publisher content in their outputs. LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

N/MA notes that showing that a particular work contributes to a particular output from a generative AI system is not required to establish prima facie infringement for purposes of training/ingestion. But outputs can offer clear evidence that a particular work has been copied. In too many cases, N/MA members have documented instances where it appears that an output results from copying one particular work. In those cases, similar outputs (including identical and substantially similar outputs) can make more clear that fair use does not apply.

As noted in response to questions 22-27, outputs can also separately be evaluated for additional infringement claims.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

Generative AI development is unlikely to succeed without a robust ecosystem that facilitates licensed use of valuable, authentic news media material. The failure to license publishing content may negatively impact the valuation of AI companies themselves, creating a cloud on the technology precisely because it is unlicensed.[181] Companies that might otherwise want to license from and deploy generative AI products and services may hang back as long as the IP issues are unresolved.

And a market that facilitates licensed exchanges of human-created content is needed for continued innovation. Researchers have found "that use of model-generated content in training causes irreversible defects in the resulting models," an effect they term "model collapse."[182] Even short of a complete model collapse under a deluge of synthetic content, there is an increased risk that generative AI chatbots could become an unattractive swamp of hallucinations without the ability to use human-created content that reflects thoughtful editorial judgment and creative expression.

The flourishing of AI technologies requires development that incorporates design principles that underscore public safety, security, and trust—as demonstrated by the recent voluntary commitments from leading AI companies to the Biden-Harris Administration and the

[181] Indeed, some AI developers have taken the unusual step of pledging to defend users of their products, in perhaps an implicit recognition of such a cloud. *See* Blake Brittain, *Google to Defend Generative AI Users from Copyright Claims*, REUTERS, Oct. 12, 2023, https://www.reuters.com/technology/google-defend-generative-ai-users-copyright-claims-2023-10-12/.

[182] Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV, at 13 (May 27, 2023), available at https://arxiv.org/abs/2305.17493. *See also* Sina Alemohammad, et al., *Self-Consuming Generative Models Go MAD*, ARXIV, (Jul. 4, 2023), available at https://arxiv.org/abs/2307.01850.

Administration's Executive Order.[183] Companies that adequately account for intellectual property responsibilities in their business models at the outset will be better poised to enjoy the tremendous potential economic benefits promised by AI innovation.

The Notice's question is oddly phrased, suggesting a "licensing requirement" rather than the need to adhere to established law. As explained further in response to question 6.3, in addition to obtaining permission to use third party material, entities may make use of material in the public domain, material they have created themselves, or material that may fall under a relevant exemption of limitation of copyright. But AI developers should not get a pass to create models that usurp licensing markets and compete with publisher content just because. What FTC Chair Lina Khan recently observed in another context holds true for copyright as well: "there is no AI exemption to the laws on the books."[184]

If the law was ignored, the economic impact of generative AI technologies on publishers and the entire information ecosystem, including authors and publishers of copyrighted works, could be catastrophic. The Office should encourage market development that supports the protection and licensed use of expressive content for ingestion into LLMs and other AI models.

In any event, foundational model developers are operating licensing companies themselves, offering access to LLM models in commercial arrangements with a panoply of downstream entities.[185] Some creators of datasets are also licensing the datasets (including on a royalty-free basis). The potential for a robust LLM licensing has fueled significant investments and increased

[183] Statements and Releases, The White House, Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (Jul. 21, 2023) available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-; The White House, Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (October 30, 2023), available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[184] Press Releases, Fed. Trade Comm., FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI (Apr. 25, 2023) available at https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

[185]*See, e.g.*, Katherine Lee, A. Feder Cooper & James Grimmelmann, Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain, at 4-5 (Draft Sep. 15, 2023), available at https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation.pdf (outlining supply chain); Alex Barinka, *Meta to Charge Cloud Providers for AI Tech That It Said Was Free*, Bloomberg, Jul. 26, 2023, https://www.bloomberg.com/news/articles/2023-07-26/meta-to-charge-cloud-providers-for-ai-tech-that-it-said-was-free?embedded-checkout=true; OpenAI, *Pricing* [n.d.], https://openai.com/pricing (last visited Oct. 26, 2023); Amazon & Anthropic, *Expanding access to safer AI with Amazon*, Anthropic, Sep. 25 2023, https://www.anthropic.com/index/anthropic-amazon.

valuation for these developers.[186] The better question for the Office to ask is whether it is sound intellectual property and industrial policy to begin a licensing supply chain at the foundational model provider, rather than further up towards the source, with the authors and publishers who create the content that is a key input for those providers. The economic impacts on publishers should not be considered mere externalities to the hopes for AI innovation.

For these reasons, we do not believe that fair licensing will hinder generative AI development—to the contrary, it is likely to improve the quality and accuracy of generative AI. Indeed, one copyright veteran observed that similar fears were raised in connection with the growth of photocopying in the 1960s.[187] At that time, some entities argued it would be impossible to secure all needed permissions to facilitate scientific progress, and regulation would put the U.S. at a competitive disadvantage. However, judicial recognition that not all photocopying was fair use did not impede innovation but led to a regime of voluntary collective licensing that has facilitated copying, enhanced access, and supported creative incentives by providing compensation to authors and rightsholders.[188]

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

To the extent that material ingested for AI was obtained in a manner contrary to publishers' terms of service, or otherwise exceeding the bounds of access granted on the public internet or otherwise, N/MA notes that this may give rise to additional liability risk. For example, the FTC has opened an investigation into whether OpenAI has engaged in unfair or deceptive privacy or data security practices in scraping public data.[189] Further, the manner in which material was scraped and obtained may be considered when evaluating questions of copyright liability.[190] Whether copyrighted works were scraped from illegal sources, as alleged in currently pending lawsuits, or contrary to terms of service or technical measures, can be relevant to a fair use analysis. It also may be relevant in considering potential damages.

[186] *See, e.g.,* Mary Azevedo, *OpenAI Could See Its Secondary-Market Valuation Soar to $90B,* TECHCRUNCH, Sep. 26, 2023, https://techcrunch.com/2023/09/26/openai-is-reportedly-raising-funds-at-a-valuation-of-80-billion-to-90-billion/?guccounter=1.

[187] Jon Baumgarten, *Former Copyright Office GC Warns Against Blanket Assertions That AI Ingestion of Copyrighted Works 'Is Fair Use'*, COPYRIGHT ALLIANCE, May 23, 2023, https://copyrightalliance.org/warns-assertions-ai-ingestion-is-fair-use/.

[188] *Id.* The CCC is such a model.

[189] Cat Zakrzewski, *FTC Investigates OpenAI over Data Leak and ChatGPT's Inaccuracy*, THE WASH. POST, Jul. 13, 2023, https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

[190] *See*, *e.g.*, *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 547 (1985), 17 U.S.C. 1201.

Transparency and Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation? 15.1. What level of specificity should be required?

N/MA supports the development and adoption of strong transparency requirements for generative AI developers because the status quo is insufficient. Indeed, Stanford University's Institute for Human-Centered AI recently published an index rating the transparency of 10 foundational model companies, finding each of them "lacking."[191] Having actual, verifiable, and accurate information regarding the uses of protected publisher content is vital for effective copyright enforcement. Such transparency requirements will likely also benefit other policy objectives outside copyright, such as safety audits, bias mitigation, risk assessments, and combating deepfakes, mis- and disinformation, hate speech, and other online harms. While this may be a multi-agency effort, N/MA believes that the Copyright Office, FTC, and USPTO can, and should, play a key role in these discussions from an IP perspective, due to their significant importance for rightsholders' ability to protect their copyrights online.

The United States is not alone in tackling issues related to AI transparency requirements.[192] Other international bodies, countries, and regions are also actively considering similar measures, including the European Union and the G7 through the Hiroshima Process.[193] While the G7 countries aim to develop global AI standards that can serve as the baseline for domestic AI laws and regulations, the European Union is already actively considering a proposal related to copyright and transparency in the AI Act. The EU institutions are currently engaged in trilogue negotiations, where the negotiators are weighing the Parliament's proposal to require

[191] Katharine Miller, Introducing The Foundation Model Transparency Index, STANFORD UNIVERSITY, Oct. 18 2023, available at https://hai.stanford.edu/news/introducing-foundation-model-transparency-index.

[192] Nor is transparency reporting a new concept for many of the very platforms that are engaging in LLM development. Many have experience preparing reports in the context of compliance obligations under online safety, privacy, or existing copyright-related duties, particularly outside the U.S.

[193] See, e.g., G7 HIROSHIMA AI PROCESS, G7 DIGITAL & TECH MINISTERS' STATEMENT (2023) available at https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf; Supantha Mukherjee, Foo Yun Chee & Martin Coulter, *EU Proposes New Copyright Rules For Generative AI*, REUTERS, Apr. 28, 2023, https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/.

generative AI developers to provide disclosures about the inclusion of copyrighted content in their training data.[194]

While the Parliament's original proposed amendment is somewhat ambiguous and should be adjusted to ensure it provides tangible, enforceable benefits to rightsholders, it is a positive step forward that can be emulated, and improved upon, outside the EU. International harmonization is particularly important as divergence in standards and enforcement may facilitate circumvention due to the borderless nature of the online ecosystem. Further, international standards or policy alignment would also lower compliance and litigation costs, and increase legal certainty and predictability to generative AI developers and rightsholders alike. Relatedly, N/MA has encouraged the Administration to take a leading role in the global discussions and to remain active in international fora finding solutions to these issues.[195]

To be meaningful, transparency standards should require generative AI and dataset developers to keep records about the protected works included in the training data and associated metadata, perhaps alongside an explanation of the legal basis on which their scraping, access, or inclusion is based. Such information should be categorized and provided in a manner that is manageable and easily searchable.

The minimum floor should be set at a level that allows rightsholders to easily and unambiguously identify when their content is being used or has been used for AI training purposes in order to enable rightsholders to effectively choose how to exercise their rights. Applicable disclosures may include not only information identifying the content used, but also the type of use, the time and method of collection and scraping, applicable retention practices, provenance, any alterations made to the content, and any third-parties who have access to the database or have already purchased it.

The goal should be to be able to construct a full chain of use. The creators of large data sets are presumably best placed to collect, retain, and disclose records regarding the information, materials, and sources included in the datasets they have built. Any downstream users, including developers, could then build on that information, and account for curation, editing, and other modification of material.

[194] Jeremy Fleming-Jones, *EU AI Act nearing agreement despite three key roadblocks – co-rapporteur*, EURONEWS, Oct. 23, 2023, https://www.euronews.com/next/2023/10/23/eu-ai-act-nearing-agreement-despite-three-key-roadblocks-co-rapporteur.

[195] DIGITAL CONTENT NEXT, EUROPEAN PUBLISHERS COUNCIL & NEWS/MEDIA ALLIANCE, JOINT LETTER ON AI (Oct. 19, 2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/10/DCN-EPC-NMA-Joint-Letter-on-AI_US.pdf.

15.2. To whom should disclosures be made?

N/MA supports an open discussion about the most efficient solutions for disclosures concerning training data. In principle, at least for publicly facing datasets, a publicly accessible disclosure depository or clearinghouse could arguably minimize the costs and burden on dataset creators, developers, and copyright owners. A centralized solution may also be preferable as generative AI applications and datasets develop and proliferate. As an alternative or supplementary measure, an industry-led technical standard that would allow rightsholders to read disclosures automatically in addition to establishing a standardized way of organizing and finding information may be worth exploring. Disclosure obligations could consider appropriate differential treatment or exceptions for legitimate proprietary, trade secret, business confidential, or directly licensed material.

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

Developers who incorporate models from third parties into their systems or applications should be subject to the same transparency requirements as other developers.[196] Further policy formulation could consider permitting compliance to be made by disclosing the underlying models and datasets used with adequate links to the disclosures made by dataset creators regarding the use of copyrighted content in their datasets. The developers should also be responsible for disclosing material changes or additions they may have made to the third-party models or datasets that are relevant to such copyright-related aspects. The opposite result, creating different obligations based on an artificial hierarchy of AI developers, may facilitate circumvention and data laundering, undermining the purpose and efficacy of potential transparency and recordkeeping requirements.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

The development and adoption of generative AI transparency and recordkeeping requirements—and the scope and subjects of such requirements—must be policy and public interest-based. It is critical that AI technologies evolve with proper guardrails and safety protocols, including strong and enforceable recordkeeping obligations. Simply put, transparency and recordkeeping requirements, as well as applicability of a rights based licensing framework, should be a cost of doing business. The government should not effectively

[196] As noted in responses to questions 22-27, they may also risk direct or secondary legal liability for infringing uses of content to train those models.

subsidize generative AI developers at the expense of authors and publishers by not adopting transparency and recordkeeping rules necessary to enforce existing copyrights out of a desire to protect developers from compliance costs. This is particularly the case considering the scale of profits anticipated by large generative AI developers whose models are especially likely to compete with creative content.[197]

Further, recordkeeping requirements may carry additional benefits, including reducing legal uncertainty and risk for AI developers, and providing rights holders with a more efficient ability to protect their content and investments against unauthorized uses, and reach negotiated agreement. In the absence of adequate recordkeeping systems, enforcement and negotiations may be considerably more cumbersome, expensive, and time-consuming, rendering such actions out of reach for far too many publishers.

In addition, such measures would facilitate greater public trust in generative AI applications and their outputs—an increasingly important benefit as AI applications and systems proliferate and become more intertwined with people's lives. Transparency and recordkeeping requirements could facilitate efforts to analyze and combat biases in AI, increase national security by helping identify harmful data sources that drive or affect generative AI outputs, and serve a variety of other interests, ranging from consumer protection to financial regulation and consumer privacy.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

The basic principle of copyright law—discussed in response to Question 9—is that where permission is required, it should be obtained before the use is made.

However, in cases where developers of generative AI have not acquired a license before the training took place, the developer should have a duty to notify applicable rightsholders as soon as practicable. The burden should not be on copyright owners to undertake the expense to reverse engineer AI training datasets and conduct forensic analysis to learn whether and how their property was used.

[197] See, e.g., Richard Waters and Camilla Hodgson, *Microsoft's Edge in AI Pays Off While Google Tries to Catch Up in the Cloud*, FINANCIAL TIMES, Oct. 25, 2023, https://www.ft.com/content/b20f9491-34b5-409c-b084-68169be6638c; Arthur Sants, *AI Helps Microsoft Pull Ahead of Google*, INVESTORS' CHRONICLE, Oct. 25, 2023, https://www.investorschronicle. co.uk/news/2023/10/25/ai-helps-microsoft-pull-ahead-of-google/; Deepa Seetharaman & Berber Jin, *OpenAI Seeks New Valuation of Up to $90 Billion in Share Sale*, WALL ST. J. (Sep. 26, 2023) https://www.msn.com/en-us/money/companies/openai-seeks-new-valuation-of-up-to-90-billion-in-share-sale/ar-AA1hiJ9W.

As discussed in answers to Questions 15-15.4., transparency and recordkeeping systems can support potential notification obligations imposed on AI developers. In addition, publicly identifying training datasets or licenses, such as OpenAI's announcement about a licensing deal with Shutterstock,[198] and the creation of searchable databases of URLs and works used in training could increase general transparency around AI training.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

Other bodies of law may already impose record keeping or disclosure obligations on developers of AI models (including privacy, consumer protection, document retention, and antidiscrimination laws, such as fairness in lending obligations), and there is ongoing interest among lawmakers in whether it would be appropriate to amend those laws. For the purposes of this Notice of Inquiry, N/MA focuses solely on copyright-related issues in these comments.

Generative AI Outputs

Infringement

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

Yes, AI-generated outputs can infringe copyrighted works, including by violating the right of reproduction and the derivative work right. Existing legal doctrines relevant to copyright infringement can be used to analyze AI-generated outputs the same as other potentially infringing materials.

For example, well-settled legal principles establish that AI-generated outputs infringe the reproduction right in media articles and other literary works where outputs are comprised of verbatim content, and also may infringe where they contain close paraphrasing or closely detailed summaries and/or substantially similar structure and expression to the original work.[199] Our response to question 23 expands on this further.

[198] *See, e.g.,* Shutterstock, *Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data*, Jul. 11, 2023, https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year.

[199] *See, e.g., Wainwright Sec. Inc. v. Wall St. Transcript Corp.,* 558 F.2d 91, 95-96 (2d Cir. 1977) (affirming finding of infringement where summaries of Wall Street Journal articles appropriated "the manner of expression, the author's analysis or interpretation of events, the way he structures his material and marshals facts, his choice of words, and the emphasis he gives to particular developments"); *Associated Press v. Meltwater US Holdings, Inc.*,

For decades, modern media, publishing, distribution, licensing, and software business models and related transactions have been developed upon this shared understanding of these metes and bounds of copyright law. The advent and use of AI-generated outputs can and must be integrated into this shared legal framework to incentivize continued creativity and innovation.

AI-generated outputs can be examined whether they infringe the derivative work right even in cases where the output itself would not otherwise qualify for copyright protection because it is not the product of human authorship. As the Office has recently correctly noted, "the test for copyrightability and the test for infringement of the derivative-works right are distinct," and "the derivative-works right is framed in terms of 'preparation,' indicating that non-human actions may be sufficient to infringe the right."[200]

In addition to potentially giving rise to infringement claims, unauthorized use of copyrighted works can also preclude protection for AI-generated outputs, even assuming that there is sufficient human authorship attached to that work. Section 103 provides that "protection for a work employing preexisting material in which copyright subsists does not extend to any part of the work in which such material has been used unlawfully." Moreover, the copyright in a "derivative work extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work, and does not imply any exclusive right in the preexisting material."

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

Substantial similarity is the dominant test applied to determine whether there has been an infringement of a copyrighted work. This test can be applied to address whether outputs from a generative AI system are infringing, including outputs of paraphrasing tools such as Quillbot AI or AI-chatbot regurgitations of protected news media content, such as examples shown in the attached Technical Annex.

931 F. Supp. 2d 537 (SDNY 2013) (finding excerpting of AP news articles to be infringing and not fair use); *Warner Bros. Ent. Inc. v. RDR Books, 575 F.Supp.2d 513 (SDNY 2008)* (finding "Lexicon" of facts, summaries, and supplemental material drawn from the Harry Potter series was infringing and not fair use). In addition, the use of copyrighted works to create other, supplemental works infringes the copyright owner's exclusive right to prepare derivative works*. Castle Rock Entertainment v. Carol Publishing Group*, 150 F. 3d 132 (2d. Cir. 1998) (affirming finding that "Seinfeld Aptitude Test" was an infringing derivative work that did not constitute fair use).

[200] Suzy Wilson & Rob Kasunic, Letter to ALI re Preliminary Draft No. (Sep. 26, 2023) available at https://www.copyright.gov/rulings-filings/restatement/comments/2023-09-26-Preliminary-Draft-No-9.pdf*.*

Additional judicial precedents have developed to help courts analyze questions of substantial similarity.[201] N/MA expects judicial doctrine to continue to evolve, including to provide any clarification necessary with respect to outputs from a generative AI system.

In the context of journalistic works and other writings published by N/MA members, including opinion, analysis, reviews, advice, investigations, and fictive works, judicial precedent is well-suited to address claims of infringement based on outputs from a generative AI system. It is black letter law that news reporting may be infringed by quoting too much of its content: the Supreme Court addressed this squarely in *Harper & Row,* holding that quoting 300-400 words verbatim from a 450-page biography was infringement, not fair use.[202]

With journalistic content, the line between copying copyrighted expression versus unprotectable facts has been frequently analyzed, and the right of news publishers to protect their copyrighted expression against overzealous borrowers repeatedly upheld. While a free Press itself depends upon facts remaining in the public domain,[203] U.S. copyright law has always aimed to incentivize the original expression of facts; the originating Copyright Act of 1790 was limited in scope to protect three types of works: books, maps, and charts.[204]

Courts navigate the facts/expression distinction by analyzing how expressive the copied material is. One illustrative case is *Salinger v. Random House,* where the Second Circuit reversed a finding by then-district Judge Leval that a biography of writer J.D. Salinger made fair use by paraphrasing letters from the famous author. The Second Circuit sharply disagreed with Judge Leval's weighing of the third fair use factor, the amount and substantiality of the portion used, noting that "protected expression has been 'used' whether it has been quoted verbatim or only paraphrased."[205] The appellate court updated the fair use analysis by considering both paraphrases and finding that the lower decision erroneously rejected claims of infringement because they employed "a cliche or a word-combination that is so ordinary that it does not qualify for the copyright law's protection." It explained:

> The "ordinary" phrase may enjoy no protection as such, but its use in a sequence of expressive words does not cause the entire passage to lose protection. And though the

[201] *See., e.g., Cavalier v. Random House, Inc.*, 297 F.3d 815, 822 (9th Cir. 2002); *Rentmeester v. Nike, Inc.*, 883 F.3d 1111, 1118 (9th Cir. 2018).

[202] *Harper & Row,* 471 U.S. at 569.

[203] 17 U.S.C. 102(b). *See, e.g. Narell v. Freeman*, 872 F.2d 907, 911 (9th Cir. 1989).

[204] U.S., Copyright Act of 1790 (1970); *See* U.S. Copyright Office, *The 18th Century* [n.d.], https://copyright.gov/timeline/timeline_18th_century.html (last visited Oct. 27, 2023).

[205] *Salinger v. Random House, Inc.,* 811 F.2d 90, 97-98 (2d Cir. 1987). The opinion also addresses that the Salinger letters were unpublished under the second factor, but status of publication was not relevant to the third factor analysis.

> "ordinary" phrase may be quoted without fear of infringement, a copier may not quote or paraphrase the sequence of creative expression that includes such a phrase. [The question is whether] the passage as a whole displays a sufficient degree of creativity as to sequence of thoughts, choice of words, emphasis, and arrangement to satisfy the minimal threshold of required creativity.[206]

Other cases draw similar conclusions. In *Wainwright,* the Second Circuit noted that although facts are not protectable, one may not take "the manner of expression, the author's analysis or interpretation of events, the way he structures his material and marshals facts, his choice of words, and the emphasis he gives to particular developments."[207] In *Robinson v. Random House, Inc.*, "approximately 25-30 percent of the words and phrases" were "used verbatim or through close paraphrasing" in an infringing book."[208] The court pointed to a side-by-side analysis to underscore how the defendant "went far beyond the use of mere facts contained in the [original book]—the appropriation included [the author's] expression " by taking "organization, writing style, even punctuation."[209] Similarly, when determining that a "Lexicon" of facts, summaries, and supplemental material drawn from the *Harry Potter* series was infringement and not a fair use, the court considered direct quotations, close paraphrases, and scene summaries, noting, "the law in this Circuit is clear that 'the concept of similarity embraces not only global similarities in structure and sequence, but localized similarity in language.'"[210]

The same analysis would apply in the generative AI context: a model's output need not replicate full passages to establish infringement, but a court may consider lengthy summaries, close paraphrases, verbatim excerpts, and whether the structure of the original work was lifted to determine substantial similarity.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

[206] *Id.*
[207] *Wainwright Sec. Inc. v. Wall St. Transcript Corp.,* 558 F.2d 91, 95-96 (2d Cir. 1977) (affirming finding of infringement based on abstract summaries of Wall Street Journal articles). *See also Associated Press v. Meltwater US Holdings, Inc.*, 931 F. Supp. 2d 537 (SDNY 2013) (excerpts of AP news articles was infringing and not fair use).
[208] *Robinson v. Random House, Inc.*, 877 F.Supp. 830, 835 (S.D.N.Y. 1995) (finding use was infringing and not fair).
[209] *Id.* at 837-838.
[210] *Warner Bros. Ent. Inc. v. RDR Books*, 575 F.Supp.2d 513 (SDNY 2008).

To the extent that generative AI developers and deployers do not maintain adequate recordkeeping or retention practices or disclose them, existing discovery practices may not be sufficient or well-tailored to address these questions. Moreover, strong public policy considerations counsel against litigation as the place of first resort. In addition to conserving judicial economy, the discovery process can be time consuming, inefficient, and imperfect. N/MA refers to its responses to questions 15-17 concerning the need for adequate transparency and recordkeeping practices.

That said, existing legal rules are currently applicable, including the obligation to preserve evidence when a party should know that the evidence may be relevant to future litigation.[211] Given the multitude of copyright infringement and other lawsuits already commenced against generative AI companies, N/MA members believe similar developers are already under an obligation to preserve and eventually disclose records of what copyrighted materials they used in "training" their systems, how the training works, and what materials are retained.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties? 25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs?

Question 25, like other infringement-related questions, will have fact-dependent answers depending on the specific circumstances of infringement. Copyright liability is joint and several, and there may be more than one direct infringer, involved in different stages of the development, deployment, or use of a generative AI model. In addition, principles of secondary liability would also apply. *See, e.g., Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.,* 545 U.S. 913 (2005). We are aware that many companies have announced intentions to indemnify certain end users against claims of copyright infringement related to the outputs generated by their generative AI models.[212]

[211] *See, e.g.,* Fed. R. Civ. Pro. 37(e) (providing for sanctions where a party failed to take reasonable steps to preserve electronically stored information in anticipation of litigation); *Fujitsu Ltd. v. Federal Exp. Corp*., 247 F. 3d 423 (2d Cir. 2001).

[212] *See* Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT ON THE ISSUES, Sep. 7, 2023, https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/; Stephen Nellis, *Adobe Pushes Firefly AI into Big Business, with Financial Cover*, REUTERS, Jun. 8, 2023, https://www.reuters.com/technology/adobe-pushes-firefly-ai-into-big-business-with-financial-cover-2023-06-08/; Neal Suggs & Phil Venables, *Shared Fate: Protecting Customers with Generative AI Indemnification*, AI & MACHINE LEARNING, Oct. 13, 2023, https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification.

That said, we believe that, at a minimum, developers of generative AI models and the interfaces incorporating them are directly liable for their own infringing output. With respect to open-source practices, the reliance on open-source AI models or sources should not obviate the need to adhere to transparency or licensing obligations. Indeed, in other contexts, open-source licensing has been a valuable and flexible tool to facilitate the permissive use of a wide range of copyrighted content--working within, as opposed to against, the overall legal framework of copyright. To the extent some users of open source material may be confused, and think that open source material is not subject to copyright protections (including publisher content incorporated therein), the Copyright Office should educate to clarify this folk misconception.

N/MA would be particularly concerned by attempts to otherwise skirt responsibility by designing conditions for "divided infringement" to escape liability for acts that would otherwise be infringing. To be sure, open-source AI models like LLAMA2 appear to have a direct financial interest in the use of its models by downstream commercial actors, as well as the right and ability to supervise its licensees.

As the marketplace and legal landscape continue to develop, the Copyright Office can consider whether guidance or recommendations may be needed to avoid incentives that shift responsibility away from the developers of generative AI models who are typically best placed to bear those compliance obligations and make it more difficult for copyright holders to effectively enforce their rights.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

Section 1202(b) is intended to prevent the deliberate stripping of copyright management information (CMI) to facilitate infringement. In enacting section 1202, Congress noted that CMI is essential to "establishing an efficient Internet marketplace" by facilitating the tracking and monitoring of copyright uses as well as licensing agreements.[213] But as the Copyright Office previously noted in its study on Moral Rights, the precise dual scienter standard was strongly debated in international fora when the related WIPO Treaties were negotiated, and decades later, many contend this standard has impeded the practical usefulness of section 1202 to protect an author's attribution's rights.[214] The Office therefore recommended a legislative

[213] THE REGISTER OF COPYRIGHTS, THE COPYRIGHT OFFICE, REPORT: AUTHORS, ATTRIBUTION, AND INTEGRITY: EXAMINING MORAL RIGHTS IN THE UNITED STATES (Apr. 2019) available at https://www.copyright.gov/policy/moralrights/full-report.pdf.
[214] *Id*. at 93-98.

amendment to this standard, which N/MA believes would be a good step.[215] The Office has also expressed concern over interpretations, like the Ninth Circuit's *Core Logic* opinion, that would raise this knowledge bar even higher.[216]

In the context of generative AI, removal of CMI can hinder the determination whether a copyrighted work has been ingested in the training process and inhibit complete and accurate recordkeeping activities. And many recent litigations around generative AI products and services have involved claims under section 1202, including *Anderson v. Stability AI*, *Doe v GitHub*, *Tremblay v. OpenAI, Inc., Silverman v Open AI, Inc.* and *Getty Images v. Stability AI*. For example, one currently active docket, *Doe v. GitHub,* involves the use of automated removal of metadata from open-source computer code used to train generative AI tools offered by Microsoft and OpenAI, where such tools "were not programmed to treat attribution, copyright notices, and license terms as legally essential."[217]

The Office should build upon its previous analyses of section 1202 and encourage legal interpretations and, if necessary, legislative reforms that allow for a balanced law regarding removal of CMI. It should discourage reckless practices like automated metadata stripping for purposes of ingesting copyright-protected works into generative AI models.

<u>Labeling or Identification</u>

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

This is a complicated question that has wider implications beyond copyright law, including potential First Amendment considerations, and the Copyright Office should exercise caution if it decides to address this issue. If any labeling requirements are adopted, appropriate agencies, including the FTC and USPTO should be consulted, and they must not be one-size-fits-all but rather should recognize the variety of AI-generated uses and be appropriately narrowly tailored. As a starting point, labeling disclosures should not apply to instances where a human person reviews and edits content that was assisted by generative AI, and remains legally liable and editorially responsible for the content. The level and format of any labeling disclosures should also be carefully considered as labeling that works for a certain type of creative work

[215] *Id.* at 98.

[216] *Id.* at 96, *citing Stevens v. Corelogic, Inc.,* 899 F.3d 666 (9th Cir. 2018), cert denied, 586 U.S. __ (U.S. Feb. 19, 2019) (No. 18-878).

[217]*Doe V. Github Inc,* No. 22-Cv-06823-Jst, 2023 U.S. Dist. (N.D. Cal. May 11, 2023) available at https://caselaw.findlaw.com/court/us-dis-crt-n-d-cal/2200493.html.

may not work for another—for example, while AI-generated photographs could be watermarked, repeated pop-ups identifying AI-generated scenes or components may seriously disrupt an audiovisual experience.

The Office could facilitate stakeholder dialogues within and between industries to facilitate the development of marketplace standards, and consider whether consultation with additional agencies on matters adjacent to copyright, such as USPTO or FTC, would be beneficial.

Additional Questions About Issues Related to Copyright

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works "in the style of" a specific artist)? Who should be eligible for such protection? What form should it take?

U.S. copyright law does not protect the "style" of a specific creator *per se*, although in some instances characters and other motifs can be protected when they are significantly distinctive and unique. As exemplified by *Steinberg v. Columbia Pictures Industries, Inc.*, the line between "style" and expression is not always clear and easy to draw.[218] Finding and preserving the appropriate balance is important for creative expression to flourish and to provide sufficient legal certainty to both original and secondary creators alike.

Related to, but separate from the specific questions posed by the Office, news, magazine, and digital media publishers are concerned about the potential of generative AI models and applications to misrepresent the source of information or the sources of other goods and services in violation of interests of trademark owners. N/MA is also concerned by the ability of generative AI to create outputs in the style of a media outlet or a high-profile journalist or other contributor or content creator while misattributing the content to said media or individual. Such misrepresentations may implicate—and potentially require changes to—other areas of law, including the Lanham Act, right of publicity, or other related laws. Absent effective ways to combat these misattributions, publishers of all types and sizes risk reputational, brand, and financial harms caused by mis- or disinformation they have not published nor generated.

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

[218] *Steinberg v. Columbia Pictures Industries, Inc*. 663 F. Supp. 706 (S.D.N.Y. 1987).

N/MA recommends the Copyright Office consider three policy recommendations and initiatives not explicitly raised by its Notice, namely publishers' ability to register online web content by submitting identifying material, the Journalism Competition and Preservation Act, and voluntary guidance and the facilitation of industry-led solutions.

First, most importantly--and urgently--the Copyright Office should adopt regulations to enable publishers to group register online web content in an efficient, economical, and simple manner. Currently, publishers are effectively unable to register their online-only content as there is no group registration option allowing for the registration of groups of frequently updated website content. We understand the constraints of the legacy eCO registration system impede the Office's ability to nimbly update the registration options it offers the public. But for news publishers, registering each individual online article under existing registration options would be burdensome, economically punitive, and contrary to the general goals of the registration system. As AI developers exclusively use online content to train their models and applications, publishers' inability to adequately register their copyrights has wide-reaching consequences to their ability to enforce their rights, monetize their content, and continue investing in the production of high-quality original content.

N/MA urges the Office to swiftly adopt regulations to enable publishers to group register news website content in an efficient manner. We are encouraged by recent suggestions that the Office has identified a solution that the eCO system may accommodate and recommend immediate adoption of this solution on at least an interim, pilot basis, and then examination to see if subsequent updates are required (including when a modernized registration system comes to fruition). We support an option to facilitate the registration of publisher owned copyrightable content on a website at a designated period of time, subject to verification. Considering the substantial market harms that systemic, unauthorized scraping for AI purposes may cause, N/MA believes that the registration option should be construed to allow publishers to seek statutory damages for the infringement of each article or other work copied. Regardless, we welcome creative thinking from the Office to introduce an updated option within the eCO system. We thank the Office for its attention to this matter and our members are ready to provide any business or technical information that would be helpful.

Second, the Office could recommend that Congress consider the passage of the Journalism Competition and Preservation Act (JCPA). The Office previously highlighted JCPA as a potential competition law-based solution to the issue of systemic unauthorized use of publisher content by the dominant online platforms, including in connection with generative AI, examined in more detail in the Office's Study on Ancillary Copyright Protections for Publishers.[219] While our comment here focuses on copyright concerns, attention to competition issues should also be given to ensure market conditions facilitate adequate compensation for use of publishers' valuable expressive material.

In its Study, the Office acknowledged that "economic trends in the news industry all point to a sea change in the press publishing ecosystem, with especially damaging consequences for local newspapers."[220] AI poses a similar existential challenge to publishers of all types and sizes, requiring an array of policy, technical, and regulatory solutions, while publishers meanwhile remain challenged by the existing practices of dominant platforms. N/MA understands that antitrust solutions are outside the scope of the Office's purview but would encourage the Office to mention such options as potential non-copyright tools in your Study.

Third, the N/MA recommends that the Copyright Office consider facilitating stakeholder dialogues in order to develop voluntary guidance documents, policy recommendations, and toolkits—similar to the NTIA's work as part of the Biden-Harris Administration's Task Force on Kids Online Health & Safety.[221] Relatedly, the Office may wish to establish a standing consultative group to ensure it can keep pace with generative AI developments as its study processes. Convening such dialogues would encourage market-led solutions that could form a significant part of a sustainable approach to AI development that protects and values publishers' copyrights and contributions to the economy and establishes a healthy growth environment for continued generative AI development.

Respectfully submitted,

Regan Smith

Danielle Coffey
President & CEO
News/Media Alliance

Regan Smith
Senior Vice President & General Counsel
News/Media Alliance

[219] "Should Congress wish to explore non-copyright measures for supporting journalism, the comments on this Study offered several proposals, including the JCPA, a levy on digital advertising revenue, increased public funding, or tax breaks for journalism. All of these proposals, however, lie beyond the expertise of the Copyright Office, and we make no findings on their merits." Press Publishers Study at 59.
[220] *Id.*
[221] NTIA, *Press Release, NTIA Seeks Comment on Protecting Kids Online*, United States Department of Commerce (Sep. 28, 2023), https://www.ntia.gov/press-release/2023/ntia-seeks-comment-protecting-kids-online.