News Corp

October 30, 2023

Comments of News Corporation
to the Copyright Office

Re: Artificial Intelligence & Copyright

News Corporation ("News Corp") appreciates the opportunity to submit these comments in response to the Copyright Office's Notice of Inquiry and Request for Comments regarding Artificial Intelligence and Copyright.

News Corp is a global media and information services company comprising businesses across news and information services, digital real estate services, book publishing, and subscription video services. Our news publishing businesses in the United States include the leading daily newspaper, *The Wall Street Journal*, and the oldest continuously published title in the country, the *New York Post*. Like all publishers of news, News Corp depends on intellectual property laws to support its endeavors.

The novel challenges that artificial intelligence—and in particular the emerging field of generative artificial intelligence—present for copyright law and policy are complex, and we commend the Copyright Office for undertaking an appropriately broad inquiry. As part of that inquiry, we urge the Copyright Office to focus on the most critical challenge that generative artificial intelligence presents for publishers. The large language models that underpin generative artificial intelligence were built using, without permission or compensation, massive amounts of journalistic works generated at a substantial cost by organizations like News Corp. And applications generated from large language models deliver to users information about, among other things, the same topics covered in News Corp publications, presenting the risk, if not the likelihood, that these products will supplant the original publications as the place users go for the information they need. Copying original content to produce a competing product that supplants the original is at the heart of what copyright law is designed to prevent.

If copyright law and policy do not prevent this massive freeriding, which siphons off reader attention and monetization opportunities for publishers, the market forces encouraging the production of professional news content will be eroded. Thus, while the Copyright Office wades into the complexities presented by generative artificial intelligence, we encourage it to not lose sight of this simple truth: protecting content creators is one of copyright law's core missions, and doing so is necessary to allow publishers to produce the kinds of news and information that News Corp employees generate every day. The implications for publishers, their readers, and democratic values could not be more profound.

The potential of generative artificial intelligence to enhance the way people interact, communicate, think, learn, and innovate cannot be overstated. At News Corp, we also appreciate the thoughtful leadership that many of the leading proponents of generative artificial intelligence have exhibited as they endeavor to create a sustainable market for their services, including their recognition that sensible regulation is appropriate and that the economics underlying professional news generation must be preserved. Our goal is to

contribute to this important conversation in a way that promotes and enhances the production and distribution of both professional news content and generative artificial intelligence products and services.

A. The use and value of news works to train artificial intelligence models and applications

Modern large language models ("LLMs") and their applications rely heavily on journalistic works for their creation. LLMs are constructed from three primary components: content, model design, and computing power. But of these, content—used as training data for LLMs—is the most pivotal influence on a model's underlying performance.[1] Critically, however, not all content has the same value as an LLM input. Adhering to a "garbage in, garbage out" dynamic, models trained on high-quality content yield superior output.[2] Widely used corpora for LLM training, including C4 and WebText, contain news articles published by leading national and international newspapers in their ten most common sources of content.[3]

Journalistic works are considered high quality for artificial intelligence ("AI") applications for several reasons. The rich news archives provided by publications like *The Wall Street Journal* provide LLMs with deep vertical subject-matter expertise. An LLM trained on *The Wall Street Journal*'s historical corpus of Pulitzer Prize-winning coverage of business, finance, and politics is empowered to synthesize nuanced insight into each of these critical topics.[4]

---

[1] For GPT-3, increasing the volume of training data by 1%, or by 3 billion tokens, should improve accuracy by approximately 3.5%, whereas increasing the number of model parameters by 1% or 1.75 billion parameters leads only to a 1.0% model accuracy gain. These statistics imply that increasing training data size leads to larger gains in model performance, and AI researchers working on the models should prioritize training dataset size over model size. Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, ARXIV (2022), https://arxiv.org/abs/2203.15556.

[2] Lora Aroyo et al., *Data Excellence for AI: Why Should You Care?*, 29 INTERACTIONS 66 (2022) ("Real-world datasets are often 'dirty', with various data quality problems and present the risk of 'garbage in = garbage out' in terms of the downstream AI systems we train and test on such data.").

[3] News articles published by leading newspapers are some of the most common sources of content in the C4 and WebText datasets because each dataset was filtered to remove low-quality data. Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, ARXIV (2021), https://arxiv.org/abs/2104.08758 (documenting characteristics of the C4 dataset and further noting that "[t]wo well-represented domains of text [in this dataset] are Wikipedia and news (NYTimes, LATimes, Al Jazeera, etc.)"). Alan D. Thompson, *What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher*, (2022), https://LifeArchitect.ai/whats-in-my-ai.

[4] *See* Pablo Villalobos et al., *Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning*, ARXIV (2022), https://arxiv.org/abs/2211.04325 ("A common property of these sources is that they contain data that has passed usefulness or quality filters. For example, in the case of news, scientific articles, or open-source code projects, the usefulness filter is imposed by professional standards (like peer review).").

Professional news articles also help LLMs to produce reliable answers and differentiate fact from fiction. News publishers adhere to long-established legal and ethical frameworks, such as constraints against libel and invasions of privacy, which provide a foundation of meaningful and effective governance. Comprehensive systems proactively ensure the integrity of news publisher content and actively respond to oversights post-publication. This is in stark contrast to internet companies that can avail themselves of Section 230 of the Communications Decency Act to amplify false and damaging content with impunity. When LLMs are trained on meticulously curated news content, they are better equipped to produce reliable, factual, and accurate answers to user queries.

Journalistic works are also exceptionally well-written: clear sentences with precise word choices; well-structured paragraphs presenting information cogently and in order of importance; and thoughtfully conceived articles providing readers with the optimal level of detail needed to understand a topic without being overburdened by extraneous information. Each of these attributes is highly valuable to LLMs, which rely on these qualities to learn language and the ability to generate longer, coherent, structurally sound outputs that users consider excellent.[5] Consider, for instance, an LLM that has been trained on *The Wall Street Journal*'s prize-winning journalism in the category of explanatory reporting for lucid writing and clear presentation. Such a model could potentially be capable of articulating analogous explanations of other complex topics.

B. Artificial intelligence models obtain journalistic works in two ways: with authorization and without

Developers of foundation models employ two main strategies to procure journalistic works for model training. First, they can obtain authorized access to journalistic works.[6] OpenAI, for example, struck a licensing deal with the Associated Press to use and train on its news articles dating back to 1985.[7] As OpenAI's chief operating officer Brad Lightcap shared, the Associated Press's "high-quality, factual text archive . . . will help to improve the

---

[5] *See* Alycia Lee et al., *Beyond Scale: The Diversity Coefficient as a Data Quality Metric Demonstrates LLMs Are Pre-trained on Formally Diverse Data*, ARXIV (2023), https://arxiv.org/abs/2306.13840; Lukas Budach, *The Effects of Data Quality on Machine Learning Performance*, ARXIV (2022), https://arxiv.org/abs/2207.14529 (setting out six dimensions for assessing data quality such as "consistent representation" (that content refers to entities or concepts by only one representation e.g., "New York" is not also represented by "NYC" or "NY"), "completeness" (that sentences are complete and do not contain missing values such as "unknown" or "NaN"), and "target accuracy" (that words are used correctly)); Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, ARXIV (2019), https://arxiv.org/abs/1810.04805; Ramesh Nallapati et al., *Abstractive Text Summarization Using Sequence-To-Sequence RNNs and Beyond*, ARXIV (2016), https://arxiv.org/abs/1602.06023; Jack W. Rae et al., *Scaling Language Models: Methods, Analysis & Insights from Training* Gopher, ARXIV (2022), https://arxiv.org/abs/2112.11446.

[6] OpenAI, *GPT-4 Technical Report*, ARXIV (2023), https://arxiv.org/abs/2303.08774.

[7] Sara Fischer, AP Strikes News-Sharing and Tech Deal With OpenAI, AXIOS (July 13, 2023), https://www.axios.com/2023/07/13/ap-openai-news-sharing-tech-deal.

capabilities and usefulness of OpenAI's systems."[8] Also relying on authorized access, Bloomberg trained its BloombergGPT model on news articles written by Bloomberg journalists and Bloomberg Television news transcripts spanning two decades, including content only available to subscribers of the Bloomberg Terminal, which can cost more than $25,000 per person per year.[9]

Whereas news organizations like the Associated Press and Bloomberg made their content available for LLM training by choice and through commercial arrangement, thousands of other news websites—including some of the country's most trusted brands such as *The Wall Street Journal*—had their valuable content taken and used without their authorization and without any commercial benefit to them. Many news publications had made their news articles available to the general public online. Despite terms of service prohibiting the unauthorized copying and use of those works, an entity registered as a non-profit named Common Crawl replicated millions of their journalistic pieces. In one month in 2018 alone, Common Crawl copied more than 180,000 works belonging to the *Chicago Tribune*, 180,000 works belonging to the *Washington Post*, and 230,000 works belonging to *The Wall Street Journal*.[10] Common Crawl then pooled news companies' works together and marketed the collective dataset to third parties under the guise of "for scientific research."[11] In reality, however, Common Crawl provided the collective dataset to both for- and non-profit entities developing LLMs, making the Common Crawl dataset a primary training corpus for nearly all LLMs.[12]

---

[8] *AP, Open AI agree to share select news content and technology in new collaboration*, ASSOCIATED PRESS (July 13, 2023), https://ap.org/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration.

[9] Shijie Wu et al., *BloomberGPT: A Large Language Model for Finance*, ARXIV (2023), https://arxiv.org/abs/2303.17564.

[10] Based on analysis of the publicly available Common Crawl June 2018 scraped dataset. Common Crawl, June 2018 Crawl Archive Now Available (2018), https://commoncrawl.org/blog/june-2018-crawl-archive-now-available. Scraped datasets contained far fewer Bloomberg material. *See* 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/ (reporting the C4 dataset contained 100M tokens of nytimes.com materials but only 22K tokens of bloomberg.com material).

[11] *In a Nutshell, Here's What We Do*, COMMON CRAWL, https://web.archive.org/web/20230626095340/https://commoncrawl.org/big-picture/ ("The web is the largest and most diverse collection of information in human history. Web crawl data can provide an immensely rich corpus for scientific research . . . ."). Common Crawl's website was updated in August 2023, deemphasizing the value of its dataset for scientific research, and adding, for the first time, discussion about its value for large language models.

[12] Alexandra (Sasha) Luccioni & Joseph D. Viviano, *What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus*, 2021 PROC. OF THE 59TH ANNUAL MEETING OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS 182, 182 ("The Common Crawl has been used to train many of the recent neural language models in recent years, including the GPT model series, BERT and FastText and, given its size, often represents the majority of data used to train these architectures.").

In addition to depriving thousands of newsrooms of the opportunity to commercially license their content, the unauthorized use of journalistic content for LLM training cannibalizes newsrooms' subscription and advertising revenue. AI models ingest copyrighted works and "synthesize" their content to produce answers, insights, and other outputs, which can substantially or identically mirror the underlying content. To illustrate, when prompted with the first few lines of a news article or book chapter, models recite verbatim the next sections or entirety of the article or book chapter; approximately 50 percent of the answers produced by the Bing AI application reproduce creators' copyrighted works word-for-word.[13]

The production of substantially similar and identical works puts LLMs and their applications in direct competition with newsrooms. A Bing executive highlighted this point candidly at a recent global conference, explaining how Bing's capability to synthesize creative works and address user queries eliminates the need to visit publishers' websites, remarking: "I'm now able to stop wasting my time reading those sites itself."[14] This potentially massive redirection of user traffic away from newspapers to generative AI ("GAI") products will inevitably lead to diminished subscription sales and a downturn in advertising revenues, amplifying the difficulties of already struggling news business models.

GAI represents the next, and potentially final, phase of modern technology's decimation of the business of news. Before the internet, news organizations recouped their significant investment in original works from a period of exclusivity dictated by the logistical demands of news dissemination. It took time for a second newspaper to rewrite, print, and distribute a competing publication's breaking piece of investigative journalism. This natural period of exclusivity was pivotal given the economics of news creation; while news content is expensive to originate, once created, the marginal cost of reproducing news reports is dramatically lower. However, the internet and search engines reduced that period of exclusivity to almost zero. When one publication invests significant capital in breaking an investigative piece of journalism, thousands of other publications can freeride off the original publisher's efforts and republish stories almost instantly, obviating the need for users to visit and reward the original publication.

Use of journalistic content for LLMs without authorization furthers this pernicious trend; the illicit appropriation of news content for model training, as well as models' capacity to synthesize and, at times, engage in the verbatim regurgitation of news content, further undermine the incentives for users to visit news sites. During the era in which news publishers benefitted from a natural period of exclusivity, copyright law needed to do less to protect and foster the generation of news; given the disappearance of that period of exclusivity, copyright law now needs to do more.

---

[13] Based on an internal study of the Bing AI application.

[14] Mike Schecter, head of the Bing product, at the Microsoft Start conference in March 2023. Even before the problem of AI, 64.82% of internet searches ended in no clicks through to web properties. AI will exacerbate this by synthesizing even more information upfront compared to traditional search. Rand Fishkin, *In 2020, Two Thirds of Google Searches Ended Without a Click*, Sparktoro (Mar. 22, 2021), https://sparktoro.com/blog/in-2020-two-thirds-of-google-searches-ended-without-a-click/.

Major AI firms have justified their appropriation of journalistic works on the basis that these works were "publicly available" on the internet for the public to read.[15] But giving the biggest commercial entities in the world carte blanche to make copies of publicly available journalistic works and use them for commercial applications is not equitable or consistent with copyright principles, particularly when considering the direct financial and labor investments embodied in these works. Accurate, reliable news content requires rigorous and multilayered editing and fact-checking from pre-publication to post-publication updates subject to vigorous internal standards of quality and reliability. News publishers employ journalists, subject-matter experts, photographers, videographers, fact-checkers, editors, and administrative staff. They incur the cost of maintaining equipment and offices, including far-flung global networks of correspondents and bureaus, to provide comprehensive coverage of events around the world. Often, a single groundbreaking investigative report may be the result of months or more of time-intensive work.

Journalists bear costs that extend far beyond the financial realm. Fearless journalists navigate life-threatening situations daily in their commitment to nurturing an informed populace. Yet tragically, journalists often pay the ultimate price for bringing news to the world: the Committee to Protect Journalists reports that at least 67 journalists and media workers were killed in the line of duty during 2022, an increase of nearly 50 percent from the prior year.[16] The risk to journalists in wartime is even greater: in the two weeks since Hamas' October 7 attacks on Israel, 29 journalists have died, both in the attacks and in covering

---

[15] Michael Schade, *How ChatGPT and Our Language Models Are Developed*, OPENAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed (last visited Oct. 22, 2023) ("OpenAI's large language models, including the models that power ChatGPT, are developed using three primary sources of information: (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or our human trainers provide."); Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, ARXIV (2023), https://arxiv.org/abs/2302.13971 ("[Meta] introduce[s] LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets."); Hugo Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ARXIV (2023), https://arxiv.org/abs/2307.09288 ("Llama 2, an updated version of Llama 1, trained on a new mix of publicly available data."); Jess Weatherbed, *Google Confirms It's Training Bard on Scraped Web Data, Too*, VERGE (July 5, 2023), https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping ("'Our privacy policy has long been transparent that Google uses publicly available information from the open web to train language models for services like Google Translate,' said Google spokesperson Christa Muldoon to The Verge. 'This latest update simply clarifies that newer services like Bard are also included. We incorporate privacy principles and safeguards into the development of our AI technologies, in line with our AI Principles.'").

[16] Jennifer Dunham, *Deadly Year for Journalists as Killings Rose Sharply in 2022*, COMM. TO PROTECT JOURNALISTS (Jan. 24, 2023), https://cpj.org/reports/2023/01/deadly-year-for-journalists-as-killings-rose-sharply-in-2022/.

Israel's response.[17] Even more journalists face the risk of unjust arrest, incarceration, and prosecution in totalitarian regimes like Russia.

Without swift and certain protection from this country's copyright law, newsrooms are responding by hiding their news works behind curtains and restricting access to paying subscribers.[18] In recent months, media companies have deployed paywalls,[19] imposed per-day article reading limitations,[20] and increased prices on content licenses. Reddit recently opted to terminate free programmatic access to its platform, which users relied on to enhance the site's functionality, in an effort to prevent AI firms from harnessing this access to acquire training data for competing commercial applications. CEO Steve Huffman lamented, "We don't need to give all of that value to some of the largest companies in the world for free."[21]

But only a select group of publishers, primarily the largest brands with premium content, established audiences, and robust infrastructure, have the privilege to capitalize on their works through paywalls and licensing. Those outside this echelon will curtail their production of news, and many will simply go out of business. When the creative value of professional writers—who take years to hone their skills—is stolen and swiftly rehashed by GAI models, it eradicates the incentive for human authors to produce new content.

Reinvigorated copyright protection for news is not just necessary to protect and foster news generation; it is necessary to foster the growth of the emerging GAI market because reducing authors' incentives to make high-quality content in turn adds fuel to an impending bottleneck in written materials available for LLM training.[22] Firms could have chosen to train models exclusively on materials in the public domain or, alternatively, they could have obtained permission for copyrighted works. Instead, by denying content creators control over the usage of their works, GAI firms have exacerbated their own data-shortage crisis. Stated differently, as set forth above, news content is a critical input for training LLMs and many of their applications; it is a critical raw material that the GAI market will rely on for its growth. Without legal protections that support the continued production of these goods, the GAI

---

[17] *Journalist Casualties in the Israel-Gaza Conflict*, COMM. TO PROTECT JOURNALISTS (Oct. 28, 2023, 8:46 AM EDT), https://cpj.org/2023/10/journalist-casualties-in-the-israel-gaza-conflict/.

[18] Although *The Wall Street Journal* has paywalls, it has historically provided access to many of its articles to users for free. This access is why over 700,000 of its articles nonetheless ended up in Common Crawl datasets. Allowing AI firms to infringe on publishers' copyright will impact market dynamics, leading to a reduction in freely available articles for the public.

[19] Newsrooms like *The New York Times*, *NBC News*, and others are exploring how to erect guardrails around journalists' written works. Alex Sherman & Lillian Rizzo, *A.I. Poses New Threats to Newsrooms, and They're Taking Action*, CNBC (June 6, 2023, 8:49 AM EDT), https://www.cnbc.com/2023/06/06/news-organizations-ai-disinformation.html.

[20] Twitter in July 2023 began limiting the number of posts that users could read per day in response to data scraping. @elonmusk, X (July 1, 2023, 1:01 PM), https://twitter.com/elonmusk/status/1675187969420828672.

[21] Mike Isaac, *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems*, N.Y. TIMES (Apr. 18, 2023), https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html.

[22] Researchers are already warning that the supply of high-quality data online is running out. By 2026, they predict that the expansion of training datasets will be capped, as high-quality data requires professionally written materials. Pablo Villalobos et al., *Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning*, ARXIV (2022), https://arxiv.org/abs/2211.04325.

market will atrophy. This situation is far from unprecedented; the inception of copyright protections was precisely to incentivize writers, researchers, and other creatives to commit the essential time, resources, and energy towards producing original content.

C. Considerations for the Copyright Office and the United States

Whereas several jurisdictions outside of the U.S. have adopted copyright exceptions for text and data mining, such exceptions only stifle citizens' access to news sources through the pernicious market failures described above. As journalists in these areas lose revenue from website traffic to GAI models that output similar content without attribution or compensation, they will be incentivized to hide their content behind paywalls to prevent further data scraping. As that occurs, citizens lose access to copyrighted material and news currently available to the public. The U.S. should avoid the same mistake.

A mandatory licensing regime would also harm AI markets by suppressing competition between firms on data sourcing. GAI firms differentiate themselves through heterogeneous decisions regarding the data included and prioritized within the training datasets for their models. Developers meticulously curate and refine their datasets to concretely align with the intended functionality and performance of their models. For instance, the data that Bloomberg's finance LLM BloombergGPT relied on diverges considerably from Google's Gemini training data, which included YouTube data like user comments, transcriptions, and video descriptions. Firms also make different decisions as to what data to "clean" from sources and what data to license from third parties. The subjective value of each parcel of data is not equal to all models. Each unit of data carries a disparate subjective value across different models, consequently establishing the foundation for competitive specialization amongst AI entities.

A government-imposed licensing regime would further have to account for the complex commercial terms currently used in the free market—who gets to use a licensor's content; the number of times a licensor's content is queried or used; the types of use and use cases (e.g., whether a licensor's content is used for research or commercial purposes, and the specific types of commercial purposes proposed); the volume of content that a licensee has access to, measured in units like words or articles; the length of time that a licensee has access to content; the ability of the licensee to query fresh and updated data over time; a licensor's ability to audit destruction and non-use of their data; how the licensee uses a licensor's content in their end product; requirements that the licensee protect the licensor's quality, reputation and risk (e.g., consumer protection risk). Each of these terms protects a copyright holder's ability to control how their content is used and to be fairly compensated for its use. A simple mandatory licensing regime would erase content creators' power to insist on contractual terms that enforce their copyright ownership rights.

Regulators should also be wary of consent-based frameworks in the face of dominant tech platforms that can leverage their market power in ancillary markets to coerce third parties into agreement. Leading internet companies have access to creators' content via several channels outside of scraped datasets, such as through crawlers and browsers, and can condition continued access to their market-dominant platforms on creators agreeing to the use

of their content to train LLMs. Moreover, consent-based frameworks will not rectify the damages publishers have and will continue to suffer before the availability of an opt-out option.

Instead, AI companies should be compelled to stop using journalistic works without authorization. This would pave the way for the free market to develop organic solutions for connecting licensors and licensees. This approach has been proven before, as evidenced by the world of digital advertising. There, where publishers are tasked with serving curated ads to individual users in a fraction of a second, the market birthed high-performance ad exchanges that transact hundreds of millions of ads each day.

D. Judicial interpretation of the scope of Section 301 preemption has not kept pace with modern artificial intelligence technology

The advent of the internet has led to new forms of infringement of news content by third parties. During the pre-internet era, infringements commonly involved a single victim and distribution by the infringer of copies or products derived from that creator's original work. The internet era, by contrast, has seen the proliferation of so-called "aggregators" that gather content from effectively *all* news publishers to fabricate and offer to consumers an interface for obtaining news, one that lacks depth but has massive breadth. This scheme provides aggregators with a profound advantage: because they aggregate all sites, they are always co-first to publish, along with the publisher breaking the story. This advantage is magnified by the fact that many aggregators are tied to products that have enormous consumer penetration, allowing the aggregators the ability to funnel their captured audience into their news interfaces. In short, while any individual publisher may object under copyright law to an aggregator's copying and repurposing of their own content, the crux of the problem lies in the fact that aggregators scrape and repurpose *all* publishers' content to compete against those publishers for user attention.

GAI presents the latest evolution in news aggregation. While the consumer interface and product may be different, the underlying scheme is the same: scraping and repurposing effectively all news content to provide a substitute product that competes with publishers for user attention. This aggregative freeriding presents a kind of industry-wide infringement that mirrors claims of unfair competition, misappropriation, and unjust enrichment that historically were actionable under state law.

Despite these clear state-law protections, some courts have found such claims preempted by the Copyright Act, which future courts could mistakenly believe preclude applications to the facts and circumstances presented by the emergence of GAI. This could be solved legislatively, but it also could be solved in the courts. Given that prior decisions ignored the firm roots on which such state protections were founded and were made by courts based on facts and technologies that were presented to the jurists at the time, courts presented with GAI-related facts could reach different conclusions today. Indeed, facts have changed, and AI has created new ways for users to essentially steal the work product of authors. The areas of law not addressed by the Copyright Act must now fill in this hole rather than simply be deemed to have been preempted by copyright law.

In the past, before the advent of the internet and GAI, it was also much easier to analyze texts to determine if there was a substantial similarity between them. It was possible to roughly determine the percentage and importance of copied text to analyze whether a copyright infringement had occurred. The advent of electronic copying, especially by GAI companies that ingest libraries worth of text and then subject the libraries to a "training" process over which they assert trade secret protection, makes the nature and amount of copying harder, if not impossible, to assess. GAI has the ability to combine multiple news stories and take a relatively small percentage of text from each. GAI companies exploit this theft-by-multiple-source-obfuscation thereby attempting to skirt the substantial similarity that would trigger clear-cut copyright infringement liability, while still serving as a substitute work to the consumer, who will in turn be disincentivized to read (and pay for) the underlying original news reporting.

The result is that U.S. copyright law also risks undervaluing the creative elements of news reporting, which include not only the composition of words used to express the facts reported, but also the editorial functions of selecting what to report and not to report, which facts are pertinent, which sources of information are authoritative and reliable, what kind of balance to strike between the presentation of conflicting viewpoints, which news items take precedence over others in importance, how deeply to go into a story, whether to credit information provided by sources, whether to name those sources, and all the other myriad editorial judgments that shape the final written composition.

The Copyright Office study asks in Questions 30-34 about additional rights that should be considered for enactment in this new AI-driven environment. The preemption issue highlights that either the Copyright Act should be amended, or the preemption doctrine be re-evaluated, to make clear that many of the types of state law claims enumerated above, when brought by news publishers and others, are not summarily deemed preempted and dismissed.