NEWS MEDIA ALLIANCE

Statement of Danielle Coffey
President & CEO, News/Media Alliance

U.S. Senate AI Insight Forum: November 29, 2023
Transparency, Explainability, Intellectual Property, & Copyright

Leader Schumer, Sen. Rounds, Sen. Heinrich, Sen. Young, and members of the Senate, thank you for inviting me to participate in today's critical discussion around artificial intelligence (AI), transparency, explainability, intellectual property and copyright. My name is Danielle Coffey, and I am President and CEO of News/Media Alliance (N/MA), representing over 2,200 news, magazine, and digital media publishers, ranging from the largest news and magazine publishers to hyperlocal newspapers, and from digital-only outlets to papers who have printed news since before the Constitutional Convention.

Today's issues are top of mind for N/MA's members, who play a vital role in their communities and in fostering an informed public. Publishers invest considerable time and resources to produce journalism and creative content that encourages democratic engagement, strengthens community ties, lowers municipal borrowing costs, safeguards consumers, keeps decision makers accountable, enriches individual lifestyles, and supports the free flow of ideas and information. Our members invest heavily in covering current events and matters of public interest, while adhering to principles and processes that support verification, accuracy, and fidelity to facts. This standard of care, and assumption of responsibility for such content, ensures quality and reliability inherent in news publishers' brands that are exposed and pressure-tested on a constant basis.

The last few years have witnessed the rise of generative AI (GAI) systems that have the potential to greatly reshape the digital marketplace and alter many features of public life. Media publishers recognize the exciting potential in responsibly deployed AI innovations, and their newsrooms and creative content teams are at the world's forefront in covering and reporting on the technologies, while also actively evaluating it for appropriate internal use. But unregulated, GAI also poses a significant threat to the pillars of a healthy and informed democracy.

GAI developers have developed large language model (LLM) systems by copying massive amounts of the creative content of media publishers without consent, credit, or compensation. These systems do not actually "learn" facts; rather, they produce sequences of words that mimic human speech. And they can be deployed for harmful uses in a variety of ways: they may make up answers, paraphrase, or alter photographs to deceive. They may create or promote false narratives that damage reputations, expose highly personal information, or present professional journalism in conjunction with unrelated vulgar and repelling content. The pervasive, unauthorized use of publisher content to produce outputs that include inaccuracies and other harmful attributes then devalues publisher brands by muddying the source of the original content and misattributing information or false information to unrelated publishers or journalists.

As we enter an upcoming election year when readers will depend on outlets that report on news from both sides of the aisle, publications will be held accountable for what is reported. By contrast, GAI systems are already working to disclaim any responsibility for their content. Low quality, often erroneous, synthetic information will make it difficult to maintain trust in our democratic system.

The importance of journalism and creativity is embedded in the constitution. Article I, Section 8 of the Constitution enables Congress to make copyright law to "promote the progress of Science and the Useful Arts", while the First Amendment establishes the Freedom of the Press. Showing the wisdom of our Founders, the press in recent years has been the first to expose major financial crimes, investigate corruption at home and abroad, and inform the public about everything from national politics to issues of highly local interests. Along the way, the Committee to Project Journalists reports 152 journalists and media workers have been killed for doing their jobs to root out the truth since 1993—with 2023 on track to break records in this terrible statistic. As valuable as AI is, it is not capable of engaging in this important work—indeed it will perniciously act to wipe it out.

Equally concerning is the lawfulness and sustainability of these GAI systems that do not shoulder the cost of reporting the news or producing creative content, but who capitalize on the results of that valuable work and often act as direct competitors to publishers. In short, journalists, writers, publishers, and other creators make the investments and take the risks while generative AI developers reap the rewards of traffic, data, brand creation, subscription fees, and advertising dollars. This is freeriding, and it is antithetical to established copyright law and the public interest that it serves.

GAI companies may argue that the press is not an important part of the way they train their LLMs. This too, is inaccurate. In a White Paper published last month, N/MA described its forensic analysis which demonstrated that developers disproportionally use online news, magazine, and digital media content to train their GAI models—sucking up America's journalistic corpus for its value and expressive nature. Our analysis shows:

- 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, used to develop 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.

This rampant copying infringes on the exclusive rights protected by copyright and far exceeds the bounds of fair use. Two key points separate LLM technology from other copying that has been found to be a fair use. First, LLMs ingest valuable media content to copy "expression for expression's sake,"[1] targeting the very aspects protected by copyright. To the extent they are ingesting content so that published words can be analyzed "in relation to all the other words in a sentence,"[2] or their sequences of words identified, that analysis and identification is intended to copy and retain the very expression that copyright protects. It is inaccurate and dangerous to anthropomorphize GAI models as "learning" unprotectable facts—these technologies are not "learning" as humans do, but memorizing to regurgitate and mimic copyright-protected expression without ever absorbing any underlying concepts.

Second, the outputs of GAI models directly compete with the protected content that was copied and used to train them. The rise of chatbots that provide detailed narrative answers to prompts goes far beyond prior judicial holdings that the carefully articulated purpose of helping users navigate to original sources could justify the wholesale copying of online content. In fact, leading developers boast that users no longer need to access or review original sources. Worse yet, an increasing number of GAI products are designed to fetch fresh news stories to "ground" generative AI output and better summarize publisher content, through a process known as "retrieval augmented generation" or "RAG". Both archived content and breaking news will be combined to compete with the delivery of news in every form. This will become unsustainable if the source of original content can no longer be funded, and the models have nothing of quality left to feed on.

Some suggest that all of the benefits of GAI are impossible if developers are required to secure permission to scrape behind the paywalls of media companies, and that it is all fair use. They may claim that free speech permits them to siphon from human creativity for their own tremendous economic benefit. These arguments are dangerous misdirection.

N/MA members are strong supporters of fair use, which the Supreme Court has called a "built-in free speech safeguard," and regularly rely on it to publish the commentary necessary to ensure an informed public. But the untested and overly expansive view of fair use proposed by GAI developers would swallow copyright in its entirety. Copyright should serve the public interest, and the humans who do the difficult work of innovating and reporting, not companies that hoover up this work to regurgitate it for immense corporate profit.

Instead, companies that adequately account for intellectual property responsibilities in their business models at the outset—just like any other business cost—will be best poised to enjoy the tremendous economic benefits promised by AI innovation. Marketplace licensing, including on a voluntary, collective basis, is the default rule under U.S. law and should be the default

[1] 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").

[2] Pandu Nayak, *Understanding Searches Better than Ever Before*, Google The Keyword (Oct. 25, 2019), (authored by Google Fellow and Vice President, Search).

here. Especially with nascent technology and emerging markets, privately negotiated arrangements will be flexible, efficient, and able to address the nuances in product markets.

As stated, while AI technologies present many potential benefits, unregulated, GAI risks driving existing publishers out of business and disincentivizing investments in new, original content. This result would undermine the purpose of the Copyright Clause of the Constitution, diminish the essential role of the Press, and jeopardize the healthy development of GAI models themselves, who depend on original human-created content to work.

To mitigate these risks, it is essential that GAI training datasets, systems, and applications be based on reliable, trustworthy content with adequate safeguards to deter the creation of false information based on that content. To do so lawfully—in a manner that protects the public interest, including professional journalism—generative AI developers should license content based on fair negotiations.

We offer the following suggestions for Congress to ensure these innovations advance in a sustainable manner:

Transparency, explainability, and traceability: *Congress should support legislation that requires the recordkeeping and disclosure of unauthorized training uses of material that is protected by copyright, by technical protection measures, or governed by contractual terms prohibiting scraping to disclose the use and weighting of specific usage of third-party content***.** Substantial transparency measures must develop around the use of copyrighted materials in GAI technologies. The public should be able to know what AI models were trained on, and make the evaluations needed to select more ethically sourced or reliable models if they choose. And publishers have a right to know who copied their content and what they are using it for. The incentives to avoid disclosure are too strong to bet on a self-regulatory solution. Obligations should be tailored to scrapers, developers, and those who configure foundational models into customized applications.

Guidelines should be issued on when and how entities should provide explanations around outputs, as well as “directionality” obligations, so that outputs include links to materials cited in summaries. While there is value in international harmonization, and addressing other data-related concerns together, any outcome should 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.

Copyright infringement and other harmful usage: *The unauthorized copying of publisher content to train and fuel commercial systems that produce substitutional output must be recognized as infringing.* Policymakers should push industry to acknowledge that the rampant copying of expressive media content to train LLMs that then compete with that content violates publishers’ exclusive rights and unfairly usurps their markets. N/MA believes that existing law establishes that this systematic and competitive conduct is infringing. And we are heartened by emerging signs of licensed GAI uses for other creative works, like music and images. But wider recognition that media content is not free for the taking is critical to foster meaningful negotiations between GAI developers and publishers.

Congress should also consider legislation clarifying that GAI developers and deployers are responsible for the design of their technology, and do not qualify for safe harbors set up decades ago to encourage a nascent tech industry to host or actively monitor and remove harmful content created wholly by third party persons. GAI systems and those who develop them should be held responsible and accountable, just like any other business.

Licensing and competition: *Policymakers should encourage market-based licensing solutions, and honor established law and policy that discourages government regulation of licensing markets as a first resort.* For some markets, this can include voluntary collective licensing as already permitted under law. Government should prevent developers from conditioning or modifying the provision of other services, such as advertising or search ranking, on (a) a content owner or site operator making available content for training or (b) a content owner or site operator permitting use of such content or site in search or other services or imposing reasonable terms and conditions.

Responsible design and accountability: *Congress should ensure generative AI development is designed to be responsible, rather than mitigating after the fact. It would be useful to make explicit that the immunities provided by Section 230 of the Communications Decency Act do not apply to AI-generated content.* Publisher experience in other contexts is that lowered standards for liability reduces the incentives for platforms to negotiate for uses of quality content, increases harm to the public, and places media publishers at a competitive disadvantage. Early GAI trends reveal so-called "overfitting" or "unintentional" harms to be all-too common. Developers should be incentivized to incorporate safety by design principles and maintain programs to prevent dangerous outcomes of AI-generated content, such as illegal content and other serious online harms.

Enforcement and anti-piracy: *Congress should ensure adequate tools to prevent unwanted scraping and "laundering" of copyrighted content.* Web scrapers must respect and follow terms of use and abide by automated flags that signal that online content be limited to specified uses. For example, tools could flag users' desire to block crawling for training while permitting beneficial uses, or for training to be limited to particular uses or users. Known pirate sites should be off-limits for AI training purposes, even if those site owners would allow data scraping, and enforcement efforts of DOJ, the IPEC,[3] and other parts of government should be empowered and funded.

Thank you for the opportunity to participate in this discussion. We look forward to working with Congress, the Administration, the States, and our counterparts around the world as this important discussion moves forward.

[3] We support swift confirmation of the IPEC, the head of the office empowered to advise the Administration on IP enforcement issues, which has been vacant since 2021.