Expert Brief

What is the Value of Journalism to AI?

A Framework for Establishing Journalism's Value in Artificial Intelligence Systems

By Courtney C. Radsch, PhD

Director, Center for Journalism and Liberty

Overview:

Problem: Big Tech is building its latest technology on the intellectual property and uncompensated use of expression, content, and data collected online and in databases. Journalistic content, which is far more than just a collection of facts and is often gathered at great costs to the journalists who report the news, is indispensable to these new AI technologies. The journalism sector needs a more sophisticated framework for how to determine the value of their content and what fair compensation would look like throughout various parts of the AI value chain. The legal regulatory system has lagged recent rapid-fire developments in AI. By failing to enforce intellectual property rights, regulators have allowed a handful of companies to further entrench their dominance and develop technologies and business models that undermine the viability of entire sectors of the economy, including journalism.

Solution: News publishers, along with the creative industries more broadly, must actively define the worth of their content and data by understanding how and why value is created throughout the generative AI process, from developing foundation models to powering real-time search, if they want to obtain fair compensation. Journalism cannot be expected to adapt its business models to the AI era without interventions by policymakers to correct market imbalances, enforce intellectual property rights, and require data access and transparency of AI systems.

Background:

After decades of giving away their content for free and being held hostage to the power of social media and search platforms, news publishers are realizing that they need to be more proactive in the era of artificial intelligence. As AI companies rely on news content to train their large language models and make AI applications more relevant, publishers already contending with a precipitous decline in referral traffic and the continued monopolization of digital advertising by Big Tech are being exploited even further.

The journalism industry shed nearly 3,000 jobs in the U.S. alone and scores of publications closed over the past year, exposing the unviability of the business models that had propped up news providers well into the 21st century. Publishers have seen referral traffic, already in decline since Facebook de-prioritized news, plummet even as they are trying to figure out how to navigate the

Courtney C. Radsch, PhD, is the Director of the Center for Journalism and Liberty at the Open Markets Institute where she produces and oversees cutting-edge research into news media market structures, technology policy, and AI governance and helps design smart policy solutions to protect and bolster journalism's financial and editorial independence in the digital age.

OPEN MARKETS

CENTER FOR JOURNALISM & LIBERTY

demise of cookies and the implications of AI for the future of their business. Meanwhile, the tech companies propelling AI have enjoyed revenue growth and valuations that have turned them into the most valuable companies in the world with market capitalizations of more than a trillion dollars each.

This disconnect can be traced back to the damage tech corporations have wrought on news publishers by cannibalizing their original content and data, displaying them in their search results or social media feeds, and then diverting advertisers, readers, and potential subscribers away from the news sites themselves. This reduces revenues earned from subscription, advertising, licensing, and affiliates, undermining not just the ability to produce quality journalism but also the industry's underlying business model.

To adapt their business models for the AI era, news publishers need to demand their rights and work collectively to put a figure on the value of journalism to artificial intelligence systems and assess the threat posed to future revenue and business models. But journalism cannot be expected to adapt its business models to the AI era without interventions by policymakers to correct market imbalances, enforce intellectual property rights, and require data access and transparency. Industry action must go hand in hand with legislative and regulatory action.

From RAGs to Riches: Leveraging Three Stages of Value Creation in AI

There are three primary stages of value creation in AI that publishers can leverage: model inputs and development, training and improving models, and applications. Journalism content can serve as rich, diverse data that improve accuracy and reliability of AI models while helping them better understand and interact with the world, particularly as synthetic media becomes more prominent online. But too narrowly focusing on the use of their content just to develop and train large language models means publishers are bypassing several other opportunities to translate value into revenue. Journalism provides ongoing value because of its quality, timeliness, and empirical grounding, and it could become even more valuable as the amount of AI-generated content increases.

Access to human-created, high-quality content that is a relatively accurate and timely portrayal of reality, like journalism, is an important input for machine learning models. Journalism is a primary provider of high-quality, relevant, and current information underpinning generative AI search, summarization, and content generation. News outlets must therefore consider how to optimize revenue streams and assert their pricing autonomy throughout the AI value chain. They will need to figure out how to unlock the value of journalism by adopting sophisticated and dynamic compensation frameworks and pricing strategies for news content in various parts of AI systems and applications, which are laid out in the next section. They will need access to data, including data sets and foundational model weights, and they need regulations that enable them to do so, regardless of whether they decide to litigate or license.

Whether opting for fixed rates or dynamic pricing based on use cases or consumption metrics, aligning pricing with the intrinsic value of journalistic content in AI is crucial if publishers are to successfully navigate the AI landscape. The following section outlines a three-pronged model for assessing value.

- **Foundation Models**: Data and content used to build foundation models, including large language models (LLMs), multimodal models (MMMs are text and images), and computer vision models (CVMs).
- **Improving models:** Training, updating, and improving models through fine-tuning, alignment, scaling, and other processes.
- **Outputs & Applications:** Retrieval Augmented Generation and real-time news: Generative search, summarization, content creation, and other applications that make use of journalism to provide more accurate, timely, and relevant results, for example, through retrieval augmented generation or grounding.

Foundational models	Data inputs, neural networks, training	Historical news data/Archives; metadata; translations
Improving Models	Model and transformer training, fine-tuning, alignment, scaling, reinforcement learning	Historical news data/Archives; prompt engineering and use of AI in the newsroom
RAGs & Applications	Retrieval Augmented Generation, grounding	Real-time news; Historical news data/Archives; search, chatbots, summarization, content creation, enterprise specific uses

FOUNDATIONAL MODELS: ARCHIVES AND HISTORICAL NEWS DATA

Journalism is an essential part of many of the foundational data sets used to develop and train generative artificial intelligence systems. News comprises *half* of the top 10 sites in the training data of a Google dataset that is used to train some of the most popular LLMs, including models from both Google and Meta. News makes up a significant part of the Common Crawl dataset, one of the oldest unstructured datasets used in many LLMs that spans 16 years of unfiltered, unlabeled content culled from across the internet and social media. Common Crawl, a nonprofit that makes its datasets available for free, began offering a regularly updated news dataset in 2016. The News and Media category is the third most prevalent source of data and makes up 13 percent of the dataset. News accounts for nearly half of the top 25 most represented sites in the Colossal Clean Crawled Corpus, a snapshot of the open-source Common Crawl dataset filtered to retain high-quality English sources and discard low-quality and problematic content like profanity and hate speech. Even content that was put behind paywalls and intended to be restricted to paid users is present in LLMs and recycled in generated responses. Last year, ChatGPT and Bing had to stop a new product partnership because users were able to bypass publisher paywalls. The *Los Angeles Times*, which relies on subscriptions and employs a paywall, is among the top sites in that dataset. In January, the newspaper laid off 20 percent of its workforce after losing tens of millions of dollars a year. Another publisher, the *New York Times*, was the fifth most used source to train ChatGPT, according to the newspaper's copyright infringement lawsuit against Open AI and Microsoft. News organizations have increasingly relied on paywalls and subscriptions amid declines in digital advertising revenue as the Google and Meta duopoly became inescapable intermediaries. Big Tech freely used publishers' content to improve the value of their search and social media platforms while controlling the underlying adtech and cloud infrastructure that publishers and advertisers rely on. Now they are once again freely using journalism to fuel their AI models, products, and services, and continuing to undermine the news industry's business model.

AI's need for data is insatiable and experts envisage that creating and training new LLMs will become increasingly difficult as AI-generated content becomes more prevalent online and in the data sets used to create and train LLMs. OpenAI reportedly transcribed upwards of one million hours of YouTube videos while Meta explored buying publishing house Simon & Schuster for this purpose. Google also transcribed YouTube videos and granted itself the right to use online content from public Google Docs, reviews on Google Maps, and a host of other applications where its users generate ostensibly public-facing content to fuel its AI products, according to the *New York Times*.

Access to human created, high-quality content that is a relatively accurate portrayal of reality is therefore an important input for the models that fuel machine learning and generative AI applications that require veracity or information retrieval; without it the models malfunction, degrade and potentially even collapse, putting the entire system at risk. And this is not a theoretical risk — Europol estimated that more than 90 percent of internet content will be AI-generated by 2026. Which makes human generated data more important, and thus more valuable. As GenAI is integrated into content production and labor markets that support human content production become more precarious -- think journalism, entertainment, and writing – AI companies will need to figure out how to maintain access to a steady supply of quality data.

IMPROVING MODELS: TRAINING, FINE-TUNING, REINFORCEMENT LEARNING

News content enriches pretrained models, reinforcement learning, and fine-tuning that help AI models to excel at specific tasks, such as summarization or text-to-image generation. Natural language processing (NLP) and fine-tuning AI models involves training them with specific types of content or human feedback.

Bigger is better, according to many in the AI field, and it was the scale of these models that launched the current wave of generative AI developments. But as the science and math advances, researchers are also learning that smaller amounts of high-quality data are more important than vast troves of lower quality data. Journalism provides a regular supply of relatively high-quality data that includes metadata and multimedia, and many news publishers are sitting on archives – what AI startup founder Lucky Gunasekara calls the "fat head" of value — that AI companies would love to get their hands on. Curated content like journalism is considered high quality and particularly useful for training and fine-tuning.

Scaling solutions that grow the value created by a core trained model and its repurposing for offshoot models, will become increasingly important as the volume of data increases, companies compete on processing speeds, and the environmental and carbon impact of AI technologies comes under greater scrutiny.

OUTPUTS & APPLICATIONS: RETRIEVAL AUGMENTED GENERATION (RAG) AND REAL-TIME NEWS

Journalism can be a particularly valuable source for grounding, which involves connecting outputs with a given data source, and retrieval augmented generation (RAG), which improves static LLM results by retrieving and connecting the model with relevant external or proprietary data. RAG is a cost-effective way to update static LLMs with more timely, relevant, or domain-specific information, which improve accuracy and predictability and reduces the likelihood and prevalence of hallucinations. The new generation of generative search engines and answer machines are powered by RAGs, which also enable chatbots and generative search and provide real-time and external context. In enterprise applications, RAGs can reduce and potentially eliminate hallucinations. RAG inputs can come from any source of content, but the bulk of real use-case for queries appears to be everyday journalism and specific genres, like finance or reviews. Perplexity AI, for example, prioritizes what founder Aravind Srinivas calls "peer-reviewed domains" such as leading journalism outlets because it is high-quality, has been through an editing process and includes background research and source verification.

Perplexity AI, a so-called AI unicorn, has attracted investors like Jeff Bezos and chip manufacturer Nvidia and earned the company a nearly $3 billion valuation, despite the fact that the AI startup pays publishers nothing and privacy concerns have led several companies, including notably Microsoft, to ban employees from the chatbot at work. The company's CEO admitted that the economic value of quality journalism is "very high" but seemed to think that visibility of content was incentive enough for news publishers. Srinivas admitted that using news inputs "doesn't actually lead to direct monetization," which presents a problem that he acknowledged "companies relying on the quality of the output for their own services should help them" with. Perplexity is rumored to be getting into the advertising business, but there is no indication that any of that revenue would be returned to the sources upon which it depends for accurate answers.

> *Access to human created, high-quality content that is a relatively accurate portrayal of reality is therefore an important input for the models that fuel machine learning and generative AI applications that require veracity or information retrieval; without it the models malfunction, degrade and potentially even collapse, putting the entire system at risk.*

AI-powered search and conversational "answer engines" are gaining in popularity and predicted to replace traditional search. Search queries are one of the most important sources of referral traffic for publishers, who are deeply concerned about how AI will further exacerbate the trend toward zero-click searches, which have been on a steady upward trend since 2019. A 2022 study found that half of all Google searches were zero-click, meaning that Google displayed or summarized the user's queried information such that the user did not click through to the original content, and just a tiny fraction of Facebook users click through on the content in their newsfeeds.

The Faustian bargain publishers made with tech platforms to exchange access to content for access to audiences via referrals seems unlikely to pay off in the new generation of AI chatbots, which do very poorly on news retrieval, according to a recent study by the Reuters institute for the Study of Journalism. It found that chatbots did a poor job of basic headline retrieval, failing to retrieve headline news accurately or consistently despite very specific prompts. And even when results provided a link to the publisher's website, it was rarely to the specific article referenced.

Silicon Valley startup Miso.ai found clickthrough rates of only 10 to 15 percent on its Answers platform, an alternative to chatbots that provides bulleted briefings with citations in contrast to narrative answers that typically contain few, if any, references. Miso.ai's low clickthrough rates indicate the new generation of search engines are unlikely to drive sufficient traffic to news publishers, indicating the need for alternative revenue models not just based on referrals.

Yet generative search and answer machines are where journalism, particularly local journalism, could be particularly valuable and thus must be able to monetize. Searching for information about local businesses, community issues, or government is going to be lot less useful if there is no local journalism informing the results. Similarly, journalism that focuses on niche topics, breaking news, or specific domains are also likely to be especially valuable to applications that want to provide up-to-date, relevant, and timely information to their users while fighting the scourge of misinformation and low-quality content online.

"Yesterday's news is actually super important," said Gunasekara. Miso's data shows that 80 percent of Answers rely on data that is more than 30 days old in order to provide context and background. "Archives, in our opinion, are extremely valuable," he said. Establishing the value of journalism throughout the search pipeline would entail understanding how it is used by, or creates value for, the crawler, index, query processor, and ad engine.

Currently this value can only be extrapolated from inadequate data, although the EU AI Act may hold out some hope for improving transparency. Furthermore, different types of journalism and news publisher content — such as archives, paywall-protected or premium content, fact-checks, and human- as opposed to AI-generated data, photos, videos, and audio — may offer different value to various parts of the AI tech stack. Publishers need to consider tailoring rates to content type and use case.

Publishers must also be realistic about the value created by generative applications such as search. According to one estimate of inference costs, the cost each ChatGPT query is .36 cents, meaning that "a search query with an LLM has to be significantly less than <0.5 cents per query, or the search business would become tremendously unprofitable for Google."

But journalism must not get locked into the current version of how search, content production, dissemination, and digital advertising work. Publishers need to remain flexible enough to update and revise agreements as technology develops and the political economy of the information ecosystem evolves. Equally important is the need for public policies, including enforcing intellectual property and contract rights that limit unfettered scraping of publisher and other creator websites while giving publishers the right to collectively negotiate and create a viable markets solution.

Strategic Rate Setting: Leveraging Uniqueness

Content isn't one-size-fits-all. Breaking news, investigative journalism, foreign coverage, local reporting, and other types of premium content possess distinct value propositions. Breaking news, thematic verticals and reviews, or local journalism can make real-time searches for information more relevant and accurate. News organizations should strategically set rates that reflect this and consider tiered pricing models that are tailored to the types of content needed for specific use cases. For example, foundational training licenses for commercial firms may command higher rates than API or on-demand access, which could be priced dynamically.

In defining their value proposition, news outlets could take a page from the Big Tech companies squeezing them. The ad tech system fueling digital advertising is based on real-time bidding that allows advertisers and publishers to connect using automated systems that (theoretically at least) optimize the cost of advertising on a specific site. Dynamic scalable licensing or royalty schemes that allow AI companies to bid for access to specific aspects of a publisher's content for various purposes could play a similar role in expanding and streamlining the remuneration process without extensive legal or business development efforts. Publishers already have licensing and royalty systems that cover different types of uses, for example, individuals versus commercial companies, and establish fees accordingly. The music industry's mechanical and performance license frameworks, along with licensing for interactive versus noninteractive platforms, are another way to account for different uses. The same could be done with AI.

Publishers could also deploy tiered licensing based on the type of content that caters to different types of AI needs. For example, a generative search engine could bid on licensing breaking news or local news focused on certain geographies. Publishers could adopt a different fee structure for AI companies that want to access basic news articles, reviews, or historical archives for model development or fine-tuning or those that want to use it for content generation applications. News summaries, translations, multimedia, and metadata all have particular relevance for AI training and improvement.

Creating a digital marketplace where AI companies can bid on access to news content that adjust based on demand, use-case, relevance, or other factors would empower news organizations to reclaim the value of their content and ensure that they maintain some level of control over how their intellectual property is used by AI systems. However, given that Google and Meta dominate the current ad tech ecosystem (with Amazon gaining market share) and significant parts of the AI ecosystem, preventing them from extracting monopoly profits by controlling the entire system will be essential. Any new licensing bidding system would need to be transparent and structurally separate from the powerful entities that control the ad tech system.

Rather than negotiating individual contracts, which is largely undoable for all but the largest publishers, news organizations should be able to set prices strategically and dynamically and will need to create collectives that can lead negotiations with AI companies.

Efforts to create a marketplace for publishers and AI companies are nascent but promising. Venture capital-backed TollBit, for example, has raised several million dollars. The startup promises to create a frictionless way for AI companies and publishers to transact, but these voluntary efforts will still need to be shaped by public policy that ensures there is sufficient information available to determine fee structures. Also, policymakers should allow small publishers to collectively bargain given the inefficiency, difficulty, and improbability of each outlet trying to get a deal on their own. Smaller publishers are not prioritized by AI companies, noted Srinivas, CEO of Perplexity AI, and is reflected in the fact that only the biggest or most prominent publishers have secured AI deals with tech firms.

This is also where news media bargaining codes could be especially consequential. More than a dozen jurisdictions around the world have passed or are considering passing laws that require dominant platforms to negotiate with publishers for the right to use their content, although the laws as currently envisioned cover just search and social media and not AI. They could expand to require that AI companies of a certain size come to the table while empowering smaller publishers to pool their resources and collectively negotiate (as CJL recommended to the South African Competition Authority in its Media and Digital Platforms Market Inquiry).

To Litigate or License? Copyright, Contracts, and the Market

Publishers around the world are considering whether to litigate or license. Publishers, musicians and record labels, photographers and photo agencies, authors, entertainers, and artists have filed lawsuits against the AI companies at the forefront of what one plaintiff characterized as "systematic theft on a mass scale." A top executive at open-source Stability AI quit in protest over the theft of copyrighted works by wealthy AI companies.

AI companies have freely admitted that requiring licensing would stall "progress" and potentially make some tools impossible but have also inked deals with dozens of media organizations. "If licenses were required to train LLMs on copyrighted content, today's general-purpose AI tools simply could not exist," according to Anthropic, the Amazon and Google-backed generative AI firm. And limiting model training to content in the public domain would not meet the needs of their models, according to OpenAI.

More than half of 1,159 publishers surveyed have requested AI web crawlers stop scanning their sites in hopes of forestalling the theft and monetization of their content by AI companies, but compliance is voluntary and can be ignored with impunity. Others have filed lawsuits against AI companies for copyright infringement, including under the Digital Millennium Copyright Act.

The New York Times filed suit in late 2023 against OpenAI and Microsoft for copyright violations after it could not reach a licensing agreement for the use of its content. In February, several leading independent news outlets including RawStory, Alternet, and The Intercept sued OpenAI for violating the Digital Millennium Copyright Act, seeking statutory damages. In April eight leading U.S. news publications owned by hedge fund Alden Capital also sued OpenAI, demanding that publishers be compensated for use of the content rather than seeking monetary damages. The Times lawsuit appears to have come after they could not reach a voluntary licensing agreement, which suggests that the companies were too far apart on their value estimates and one reason why the final arbitration offer model of Australia's News Media Bargaining Code is attractive. The 2021 law and a similar effort in Canada required designated platforms to negotiate with news publishers for the use of their content and ensured that negotiations took place in good faith and did not drag on indefinitely by giving the arbitrator the right to pick one side or the other if a mutually agreed upon figure could not be reached.

Legal regulatory efforts to enforce copyright and impose mandatory negotiating frameworks on search and social media companies have gained popularity around the world recently, with at least a dozen jurisdictions and EU member states considering or passing such legislation. Allowing collective bargaining by publishers, requiring access to data held by designated tech platforms, and imposing transparency requirements will bolster a regulatory framework that not only increases the power of local and smaller news outlets but could be applied to AI companies that crawl and scrape publisher websites.

But to sue or sign is not an either-or proposition, and to some extent ensuring that a market for licensing publisher data exists could help boost copyright claims by mitigating fair use arguments. This happens because in many jurisdictions, market replacement is a key factor in determining whether the unlicensed use of journalistic content is protected by copyright. The leading AI companies, with their Big Tech partnerships, are projected to reap billion-dollar revenues with valuations approaching a trillion dollars. By contrast, news publishers — whose content is integral to AI models — are either shutting down or struggling to remain viable.

WHAT WE CAN LEARN FROM EXISTING DEALS

Reddit inked a $60 million partnership with Google in February, effectively planting a flag with a number on it in the ground. Reddit will provide training data and more efficient ways to train models by allowing Google to access its Data API, while Google locks in Reddit's use of its VertexAI cloud and gains access to a real-time fresh structured data source. Given Reddit's 50 million daily active users, that translates to a value of about 83 cents per user per year. Given the prevalence of misinformation, hate speech, extremism, and "norm-violating influencers" on Reddit, journalism could be valued far higher for its accurate and higher quality data.

Reddit had already started featuring more prominently in Google search results prior to the deal announcement, which came just as the company filed for its initial public offering IPO, which revealed that the company's licensing agreements with a number of outside parties amount to $203 million over the next two to three years. We do not know whether this includes deals with OpenAI, whose CEO is a major Reddit shareholder, or Tencent Holdings, which owns 11 percent of outstanding shares and is one of the China's leading AI companies. Reddit also signed a deal in May with OpenAI, whose CEO Sam Altman is the social media platforms' third largest shareholder, though the amount and terms have not been disclosed.

We know much less about the deals that AI companies have already made with news publishers, though OpenAI has been the most aggressive in pursuing voluntary licensing deals. OpenAI, in which Microsoft has a major ownership stake, has made licensing agreements with some of the largest journalism organizations in the world, including the Associated Press, Axel Springer, Le Monde, Spanish media conglomerate Prisa, and DotDash Meredith, the largest print and digital publisher in the U.S., while several more are reportedly in discussion with Apple and Google. Although the terms are largely unknown, analysis of publicly available announcements and news reports indicate that many of the deals cover licensing content, including archives and contemporary content, for a defined period (two years seems to be the norm) as well as access to AI tools in the newsrooms. Reports indicate that OpenAI offered between $1 and $5 million annually while Apple appears to be offering more money for a wider array of uses to a handful of large publishers including Condé Nast and NBC News. The AP deal with OpenAI provides partial access to its archive going back to 1985 and is likely to set a benchmark for other deals going forward, though industry insiders think the AP undervalued its worth and should have leveraged its power to get deals for the news organizations that work with the cooperative.

Microsoft and Google have not announced any specific AI licensing deals, though they have announced bespoke "collaborations" and "partnerships" to assist newsrooms in adapting and adopting AI in the newsrooms. Microsoft did not respond to specific questions about whether it compensates publishers when their content shows up in generative searches or chats, pointing instead to a page outlining how it is ensuring newsrooms can "innovate" with its products. Google did not respond to a request for comment. Publishers will be creating value for these companies using their products, though it is unlikely that any of them have negotiated with these AI companies for the value created through their use of these tools, such as prompt engineering or fine-tuning.

Onboarding newsrooms to their AI infrastructure and training them in how to integrate AI into the journalism process is redolent of the way that Facebook, now Meta, and Google "helped" newsrooms make better use of their tools and platforms over the past decade, which served to entrench the dependence of publishers on these platforms, even as they pivoted away from news and tried to torpedo regulatory efforts aimed at making them compensate publishers. While newsrooms need to build their capacity to leverage AI, relying on Big Tech to drive these efforts reinforces and deepens platformization and undermines their editorial and economic independence.

In the meantime, some news providers are forging ahead with voluntary agreements in the absence of legal regulatory clarity. But this leaves out smaller and local publishers and could undermine efforts to develop sustainable business model alternatives for journalism. Bespoke, secretive deals with the largest or most influential news outlets are not a replacement for public policy and will not rescue local news from the precarity created by corporations who skirt the law and enjoy dominant market power. Furthermore, regardless of whether news outlets are engaged in individual or collective discussions, developing a robust understanding of the value proposition is critical for ensuring they do not leave money on the table. The framework for licensing or developing a royalty model would include different aspects of their content and data for various stages of the AI model, as well as use cases.

Bespoke, secretive deals with the largest or most influential news outlets are not a replacement for public policy and will not rescue local news from the precarity created by corporations who skirt the law and enjoy dominant market power.

WHY VOLUNTARY BLOCKING AND WALLING OFF CONTENT IS NOT THE RIGHT SOLUTION

More than two-thirds of leading U.S. and EU newspapers, and more than 75 percent of U.S. news outlets, are behind a paywall. More than half of 1,159 publishers surveyed this year have requested at least one AI web crawler to stop scanning their sites in hopes of stalling the theft and monetization of their content by AI companies. But compliance is voluntary and can be ignored with impunity, especially given the existing incentives and a lack of legally enforceable restrictions. And many publishers feel that simply asking companies not to crawl though robots.txt – the voluntary protocol that websites use to provide instructions to automated crawlers and scrapers – is insufficient. Nonetheless, including restrictions on crawling, scraping and commercial use in a site's terms of service and via robots.txt could strengthen a publisher's case if pursuing litigation.

To fend off reproach by publishers and content creators, Google started to allow publishers to de-index their sites or pages from its AI crawlers without also withdrawing from its search crawler in late 2023. But de-indexing news undermines public interest goals by reducing the supply side of quality information while further entrenching the dominance of Big Tech companies that have already built LLMs using news.

Furthermore, walling off news content and preventing it from being used would remove quality information from RAGs and the chatbots, generative search products, and foundational models that underpin generative AI services, resulting in a range of negative impacts. Withdrawing news will exacerbate mis- and disinformation, reinforce harms like "hallucinations" (essentially incorrect answers made up by generated by AI), and undermine a host of downstream applications. The *Washington Post*, for example, found that training sets already include several media outlets that rank low on NewsGuard's independent scale for trustworthiness, or which are backed by a foreign country, including Russia's propaganda arm RT. Removing countervailing quality content would just give junk news, propaganda, disinformation and synthetic media greater prominence. Continued access to quality human data is becoming increasingly important and thus more valuable. Google's data deal with Reddit, for example, means that it can leverage the site's human-generated data to better train Gemini and its other AI models, chatbots, and generative tools to detect bias, misinformation or other malign content, even though Reddit does not offer the level of curated high-quality content that journalism does. The news industry, too, must figure out its value proposition, from original reporting to verification and fact-checking to analysis, reviews, and opinion.

Making AI Safer through Licensing and Compensation

Policymakers around the world are concerned about the ability to document and scrutinize the data used by foundation models. Imposing requirements that publishers, authors, photographers, and other creators receive compensation for their work will help ensure that systems are put in place and technology developed that will allow them to do just that. Regardless of a handful of voluntary agreements, policymakers should explore statutory licensing and taxing generative AI firms to create a compensation fund that rights holders could apply for.

Many are also concerned about the rapid pace of development and deployment of general and generative AI, particularly given the lack of safeguards, regulations, and legislation in place to govern its use. Indeed, many AI luminaries and tech leaders signed a letter last year calling for a temporary halt on AI development, which went nowhere because no one wanted to be left behind in the scramble for AI dominance.

Making Business Models Viable Requires Public Policy

AI companies claim that it would be impossible to license data used in foundation models and compensate rights holders, as if that should absolve them of the responsibility to do so. But acquiescing to this stance means that we are prioritizing one business model over another. We are favoring a business model based on the pervasive theft of intellectual property by the wealthiest companies in the world over the business model of journalism. Journalism cannot be expected to adapt its business models to the AI era without interventions by policymakers to correct market imbalances, enforce intellectual property rights, and require data access and transparency.

How we decide to allocate intellectual property rights and what we decide about how fair use does or does not apply to developing and training artificial intelligence systems will have profound ramifications for business models in a variety of sectors and the further concentration of power in a handful of technology corporations. Over the past nearly two decades, as tech companies like Apple, Amazon, Google, Meta, and Microsoft grew to become some of the most valuable companies in the world, the United States lost a third of its newspaper and two-thirds of its newspaper journalists. They cannot be replaced with AI.

Gone are the days of passive acceptance that enabled social media and search platforms to siphon off value from publishers and journalists without compensation. We know that journalism is essential to democracy. Given AI's well-established harms like the spread of misinformation during elections, we cannot say the same of generative AI.

ACKNOWLEDGMENTS

Thank you to John Villasenor (UCLA), Subramaniam Vincent (Santa Clara University), and Matt Canute (Simon Frazier University) for providing feedback and peer review.

CONTACT:

Center for Journalism & Liberty
at Open Markets Institute

CJL@openmarketsinstitute.org
journalismliberty.org