

## COMMENTARY

# Connecting the dots: AI is eating the web that enabled it

Tom Wheeler

June 24, 2024

- 
- The large language models that power generative AI tools were built using data scraped from countless websites, but they now seek to eliminate the need for users to go to those same sites.
- 
- Already, a quarter of all web pages developed between 2013 and 2023 no longer exist, and traffic from search engines to the web is predicted to fall by 25% in the next two years.
- 
- While some publishers are suing AI developers for using their data to train AI tools, many are entering into partnerships with companies like OpenAI, who promise monetary compensation and the promotion of the publisher's websites within AI-generated content.
- 

Connecting the dots of recent research suggests a new future for traditional websites:

- Artificial Intelligence (AI)-powered search can provide a full answer to a user's query 75% of the time without the need for the user to go to a website, according to [research](#) by The Atlantic.
- A worldwide [survey](#) from the University of Toronto revealed that 22% of ChatGPT users "use it as an alternative to Google."
- Research firm Gartner [forecasts](#) that traffic to the web from search engines will fall 25% by 2026.
- Pew Research [found](#) that a quarter of all web pages developed between 2013 and 2023 no longer exist.

The large language models (LLMs) of generative AI that scraped their training data from websites are now using that data to eliminate the need to go to many of those

# From decentralized information to centralized conclusions

[Created by Sir Tim Berners-Lee](#) in 1989, the World Wide Web redefined the nature of the internet into a user-friendly linkage of diverse information repositories. “The first decade of the web...was decentralized with a long-tail of content and options,” Berners-Lee [wrote](#) this year on the occasion of its 35th anniversary. Over the intervening decades, that vision of distributed sources of information has faced multiple challenges. The dilution of decentralization began with powerful centralized hubs such as Facebook and Google that directed user traffic. Now comes the ultimate disintegration of Berners-Lee’s vision as generative AI reduces traffic to websites by recasting their information.

The web’s open access to the world’s information trained the large language models (LLMs) of generative AI. Now, those generative AI models are coming for their progenitor.

The web allowed users to discover diverse sources of information from which to draw conclusions. AI cuts out the intellectual middleman to go directly to conclusions from a centralized source.

The AI paradigm of cutting out the middleman appears to have been further advanced in Apple’s recent [announcement](#) that it will incorporate OpenAI to enable its Siri app to provide ChatGPT-like answers. With this new deal, Apple becomes an AI-based disintermediator, not only eliminating the need to go to websites, but also potentially disintermediating the need for the Google search engine for which Apple has been paying \$20 billion annually.

The Atlantic, University of Toronto, and Gartner studies suggest the Pew research on website mortality could be just the beginning. Generative AI’s ability to deliver conclusions cannibalizes traffic to individual websites threatening the *raison d’être* of all websites, especially those that are commercially supported.

information providers. “The rise of digital media and technology has transformed the way we access our news and entertainment,” the U.S. Census Bureau [reported](#) in 2022, “It’s also had a devastating impact on print publishing industries.” Thanks to the web, total estimated weekday circulation of U.S. daily newspapers fell from 55.8 million in 2000 to 24.2 million by 2020, [according](#) to the Pew Research Center.

The World Wide Web also pulled the rug out from under the economic foundation of traditional media, forcing an exodus to proprietary websites. At the same time, it spawned a new generation of upstart media and business sites that took advantage of its low-cost distribution and high-impact reach. Both large and small websites now [feel the impact](#) of generative AI.

Barry Diller, CEO of media owner IAC, harkened back to that history when he [warned](#) a year ago, “We are not going to let what happened out of free internet happen to post-AI internet if we can help it.” Ominously, Diller observed, “If all the world’s information is able to be sucked up in this maw, and then essentially repackaged in declarative sentence in what’s called chat but isn’t chat...there will be no publishing; it is not possible.”

The New York Times [filed a lawsuit](#) against OpenAI and Microsoft alleging copyright infringement from the use of Times data to train LLMs. “Defendants seek to free-ride on The Times’s massive investment in its journalism,” the suit [asserts](#), “to create products that substitute for The Times and steal audiences away from it.”<sup>1</sup>

Subsequently, eight daily newspapers owned by Alden Global Capital, the nation’s second largest newspaper publisher, filed a [similar suit](#). “We’ve spent billions of dollars gathering information and reporting news at our publications, and we can’t allow OpenAI and Microsoft to expand the Big Tech playbook of stealing our work to build their own businesses at our expense,” a spokesman [explained](#).

The legal challenges are pending. In a colorful description of the suits’ allegations, journalist Hamilton Nolan described [AI’s threat](#) as an “Automated Death Star.”

In July 2023, the [Associated Press](#) became the first major content provider to license its archive to OpenAI. Recently, however, the deal-making floodgates have opened. Rupert Murdoch's [News Corp](#), home of The Wall Street Journal, New York Post, and multiple other publications in Australia and the United Kingdom, German publishing giant [Axel Springer](#), owner of Politico in the U.S. and Bild and Welt in Germany, venerable media company [The Atlantic](#), along with new media company [Vox Media](#), the [Financial Times](#), Paris' [Le Monde](#), and Spain's [Prisa Media](#) have all contracted with OpenAI for use of their product.

Even Barry Diller's publishing unit, [Dotdash Meredith](#), agreed to license to OpenAI, approximately a year after his apocalyptic [warning](#).

News Corp CEO Robert Thomson [described](#) his company's rationale this way in an employee memo: "The digital age has been characterized by the dominance of distributors, often at the expense of creators, and many media companies have been swept away by a remorseless technological tide. The onus is now on us to make the most of this providential opportunity."

"There is a premium for premium journalism," Thomson observed. That premium, for News Corp, is reportedly \$250 million over five years from OpenAI. Axel Springer's three-year deal is reportedly worth \$25 to \$30 million. The Financial Times terms were reportedly in the annual range of \$5 to \$10 million.

## AI companies' different approaches

While publishers debate whether AI is "providential opportunity" or "stealing our work," a similar debate is ongoing among AI companies. Different generative AI companies have different opinions whether to pay for content, and if so, which kind of content.

Google, whose core business is revenue generated by recommending websites, has not sought licenses to use the content on those websites. “The internet giant has long resisted calls to compensate media companies for their content, arguing that such payments would undermine the nature of the open web,” the New York Times [explained](#) <sup>7</sup>. Google has, however, licensed the user-generated content on social media platform [Reddit](#) <sup>8</sup>, and together with Meta has [pursued Hollywood rights](#) <sup>9</sup>.

OpenAI has followed a different path. Reportedly, the company has been pitching a “Preferred Publisher Program” to select content companies. Industry publication AdWeek [reported](#) <sup>10</sup> on a leaked presentation deck describing the program. The publication said OpenAI “disputed the accuracy of the information” but claimed to have confirmed it with four industry executives. Significantly, the OpenAI pitch reportedly offered not only cash remuneration, but also other benefits to cooperating publishers.

As of early June 2024, other large generative AI companies have not entered into website licensing agreements with publishers.

## Content companies surfing an AI tsunami

On the content creation side of the equation, major publishers are attempting to avoid a repeat of their disastrous experience in the early days of the web while [smaller websites](#) <sup>11</sup> are fearful the impact on them could be even greater.

As the web began to take business from traditional publishers, their leadership scrambled to find a new economic model. Ultimately, that model came to rely on websites, even though website advertising offered them [pennies](#) <sup>12</sup> on their traditional ad dollars. Now, even those assets are under attack by the AI juggernaut. The content companies are in a new race to develop an alternative economic model before their reliance on web search is cannibalized.

The first step in the program is direct compensation. To Barry Diller, for instance, the fact his publications will get “direct compensation for our content” means there is “[no connection ↗](#)” between his apocalyptic warning 14 months ago and his new deal with OpenAI.

Reportedly, the cash compensation OpenAI is offering has two components: “guaranteed value” and “variable value.” Guaranteed value is compensation for access to the publisher’s information archive. Variable value is payment based on usage of the site’s information.

Presumably, those signing with OpenAI see it as only the first such agreement. “It is in my interest to find agreements with everyone,” Le Monde CEO Louis Dreyfus [explained ↗](#).

But the issue of AI search is greater than simply cash. Atlantic CEO Nicolas Thompson [described ↗](#) the challenge: “We believe that people searching with AI models will be one of the fundamental ways that people navigate to the web in the future.” Thus, the second component in OpenAI’s proposal to publishers appears to be promotion of publisher websites within the AI-generated content. Reportedly, when certain publisher content is utilized, there will be hyperlinks and hover links to the websites themselves, in addition to clickable buttons to the publisher.

Finally, the proposal reportedly offers publishers the opportunity to reshape their business using generative AI technology. Such tools include access to OpenAI content for the publishers’ use, as well as the use of OpenAI for writing stories and creating new publishing content.

## Back to the future?

Whether other generative AI and traditional content companies embrace this kind of cooperation model remains to be seen. Without a doubt, however, the initiative by both parties will have its effects.

We have seen new technology-generated copyright issues resolved in this way before.<sup>2</sup> Finding a credible solution that works for both sides is imperative. The promise of AI is an almost boundless expansion of information and the knowledge it creates. At the same time, AI cannot be a continued degradation of the free flow of ideas and journalism that is essential for democracy to function.

## Newton's Law in the AI age

In 1686 Sir Isaac Newton posited his [three laws of motion](#)<sup>3</sup>. The third of these holds that for every action there is an equal and opposite reaction. Newton described the consequence of physical activity; generative AI is raising the same consequential response for informational activity.

The threat of generative AI has pushed into the provision of information and the economics of information companies. We know the precipitating force, the consequential effects on the creation of content and free flow of information remain a work in progress. [\(#\\_ftnref1\)](#)

### AUTHOR



**Tom Wheeler** Visiting Fellow - Governance Studies, Center for Technology Innovation  @tewheels

---

## Acknowledgements and disclosures

## Footnotes

1. To be clear, the lawsuit pertains to the use of the Times information. The company is also exploring the [use of generative AI ↗](#) with a team of dedicated editors and engineers.
2. In the [late 1970s ↗](#), after having won two Supreme Court [decisions ↗](#) holding they had no copyright liability, the nascent cable television industry agreed to a schedule of payments in return for a “[compulsory license ↗](#)” for the retransmission of broadcast television programs.

---

Copyright 2024 The Brookings Institution