**KULA**
Knowledge Creation, Dissemination, and Preservation Studies

# Large Language Publishing
## The Scholarly Publishing Oligopoly's Bet on AI

Jefferson Pooley

Volume 7, Number 1, 2024

URI: https://id.erudit.org/iderudit/1114527ar
DOI: https://doi.org/10.18357/kula.291

See table of contents

Publisher(s)

University of Victoria Libraries

ISSN

2398-4112 (digital)

Explore this journal

Cite this document

Pooley, J. (2024). Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI. *KULA*, *7*(1), 1–11. https://doi.org/10.18357/kula.291

Article abstract

The AI hype cycle has come for scholarly publishing. This essay argues that the industry's feverish—if mostly aspirational—embrace of artificial intelligence should be read as the latest installment of an ongoing campaign. Led by Elsevier, commercial publishers have, for about a decade, layered a second business on top of their legacy publishing operations. That business is to mine and process scholars' works and behavior into prediction products, sold back to universities and research agencies. This article focuses on an offshoot of the big firms' surveillance-publishing businesses: the post-ChatGPT imperative to profit from troves of proprietary "training data," to make new AI products and—the essay predicts—to license academic papers and scholars' tracked behavior to big technology companies. The article points to the potential knowledge effects of AI models in academia: Products and models are poised to serve as knowledge arbitrators, by picking winners and losers according to what they makevisible. I also cite potential knock-on effects, including incentives for publishers to roll back open access (OA) and new restrictions on researchers' access to the open web. The article concludes with a call for a coordinated campaign of advocacy and consciousness-raising, paired with high-quality, in-depth studies of publisher data harvesting—built on the premise that another scholarly-publishing world is possible. There are many good reasons to restore custody to the academy, the essay argues. The latest is to stop our work from fueling the publishers' AI profits.

KULA KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Pooley, Jefferson. 2024. Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7(1). https://doi.org/10.18357/kula.291

COMMENTARY

# Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI

Jefferson Pooley

University of Pennsylvania

The AI hype cycle has come for scholarly publishing. This essay argues that the industry's feverish—if mostly aspirational—embrace of artificial intelligence should be read as the latest installment of an ongoing campaign. Led by Elsevier, commercial publishers have, for about a decade, layered a second business on top of their legacy publishing operations. That business is to mine and process scholars' works and behavior into prediction products, sold back to universities and research agencies. This article focuses on an offshoot of the big firms' surveillance-publishing businesses: the post-ChatGPT imperative to profit from troves of proprietary "training data," to make new AI products and—the essay predicts—to license academic papers and scholars' tracked behavior to big technology companies. The article points to the potential knowledge effects of AI models in academia: Products and models are poised to serve as knowledge arbitrators, by picking winners and losers according to what they make visible. I also cite potential knock-on effects, including incentives for publishers to roll back open access (OA) and new restrictions on researchers' access to the open web. The article concludes with a call for a coordinated campaign of advocacy and consciousness-raising, paired with high-quality, in-depth studies of publisher data harvesting—built on the premise that another scholarly-publishing world is possible. There are many good reasons to restore custody to the academy, the essay argues. The latest is to stop our work from fueling the publishers' AI profits.

**Keywords:** academic publishing; artificial intelligence; data mining; prediction products; surveillance publishing

*The New York Time*s ushered in the new year with a lawsuit against OpenAI and Microsoft. The paper covered the suit, fittingly, as a major business story (Grynbaum and Mac 2023). OpenAI and its Microsoft patron had, according to the filing, stolen "*millions* of The Times' copyrighted news articles, in-depth investigations, opinion pieces, reviews, how-to guides, and more" (New York Times Company v. Microsoft Corp. et al. 2023, 2, italics in original)—all to train OpenAI's large language models (LLMs). The *Times* sued to stop the tech companies' "free-ride" on the newspaper's "uniquely valuable" journalism (New York Times Company v. Microsoft Corp. et al. 2023, 4).

OpenAI and Microsoft have, of course, cited fair use to justify their permissionless borrowing. Across seventy bitter pages, the *Times*' lawyers drove a bulldozer through all four factors that US judges weigh for fair use. The brief also points to reputational harm—from made-up responses that ChatGPT or Bing Chat attribute to the *Times*: "In plain English, it's misinformation" (New York Times Company v. Microsoft Corp. et al. 2023, 52).

There is no question that attorneys for Elsevier and the other scholarly publishing giants read the *Times* filing carefully. They will have noticed a leitmotif: The newspaper's expensively reported stories produce *trusted* knowledge, in otherwise short supply. In a "damaged information ecosystem . . . awash in unreliable content," the *Times*' journalism is an "exceptionally valuable body of data" for AI training, states the filing (New York Times Company v. Microsoft Corp. et al. 2023, 2, 56). Other news organizations have the same view; some have signed licensing deals, while others are negotiating with OpenAI and its peers. No more free rides.

The big scholarly publishers are very likely to agree. And they are sitting on the *other* corpus of vetted knowledge, science and scholarship. A few licensing deals have already been announced, with talks for others almost certainly underway. No doubt threats and lawsuits have also been prepped. At the same time, the commercial publishers are building their own AI products. In the two years since ChatGPT's splashy entrance, at least three of the big five scholarly publishers, plus Clarivate, have announced tools and features powered by LLMs. They are joined by dozens of VC-backed startups—acquisition targets, one and all—promising an AI boost across the scholarly workflow, from literature search to abstract writing to manuscript editing.

Thus the two main sources of trustworthy knowledge, science and journalism, are poised to extract protection money—to otherwise exploit their vast pools of vetted text as "training data." But there's a key difference between the news and science: Journalists' salaries, and the cost of reporting, are covered by the companies. Not so for scholarly publishing: Academics, of course, write and review for free, and much of our research is funded by taxpayers. The *Times* suit is marinated in complaints about the costly business of journalism. The likes of Taylor & Francis and Springer Nature will not have that argument to make. It is hard to call out free-riding when it is your own business model.

## Surveillance Publishing, LLM Edition

The AI hype cycle has come for scholarly publishing. The industry's feverish—if mostly aspirational—embrace of AI should be read as the latest installment of an ongoing campaign.[1] Led by Elsevier, commercial publishers have, for about a decade, layered another business on top of their legacy publishing operations. That business is to mine and process scholars' works and behavior into prediction products, sold back to universities and research agencies. Elsevier, for example, peddles a dashboard software, Pure, to university assessment offices—one that assigns each of the school's researchers a Fingerprint® of weighted keywords. The underlying data comes from Elsevier's Scopus, the firm's propriety database of abstracts and citations. Thus the scholar *is* the product: Her articles and references feed Scopus and Pure, which are then sold back to her university employer. That same university, of course, already shells out usurious subscription and APC dollars to Elsevier—which, in a painful irony, have financed the very acquisition binge that transformed the firm into a full-stack publisher.

Elsevier and the other big firms are, to borrow Sarah Lamdan's (2022) phrase, data cartels. I have called this drive to extract profit from researchers' behavior *surveillance publishing* (Pooley 2022)—by analogy to Shoshana Zuboff's (2019) notion of surveillance capitalism, in which firms like Google and Meta package user data to sell to advertisers. The core business strategy is the same for Silicon Valley and Elsevier: Extract data from behavior to feed predictive models that, in turn, get refined and sold to customers. In one case it is Facebook posts and in the other abstracts and citations, but either way the point is to mint money from the by-products of (consumer or scholarly) behavior. One big difference between the big tech firms and the publishers is that Google et al. entice users with free services like Gmail: If you are not paying for it, the adage goes, then you are the product. In the Elsevier case we are the product *and* we are paying (a lot) for it.

Elsevier and some of the other big publishers already harness their troves of scholarly data to, for example, assign subject keywords to scholars and works. They have, indeed, been using so-called AI for years now, including variations on the ML techniques ascendant in the last fifteen years or so. What is different about the publishers' imminent licensing windfall and wave of announced tools is, in a word, ChatGPT. It is true that successive versions of enormous "large language" models from OpenAI, Google, and others have been kicking around in commercial and academic circles for years. But the November 2022 public release of ChatGPT changed the game. Among other things, and almost overnight, the value of *content* took on a different coloration. Each of the giant "foundation" models, including OpenAI's GPT series, is fed on prodigious helpings of text. The appetite for such training data is not sated, even as the legality of the ongoing ingestion is an open and litigated question.

The big publishers think they are sitting on a gold mine. It is not just their paywalled, full-text scholarship, but also the reams of other data they hoover up from academics across their platforms and products. In theory at least, their proprietary content is—unlike the clown show of the open web—vetted and linked. On those grounds, observers (e.g., Zhavoronkov 2023) have declared that publishers may be the "biggest

[1] The term itself is misleading, though now unavoidable. By AI (artificial intelligence), I am mostly referring to the bundle of techniques now routinely grouped under the *machine learning* (ML) label. There is an irony in this linguistic capture. For decades after its coinage in the mid-1950s, *artificial intelligence* was used to designate a rival approach, grounded in rules and symbols. What most everyone now calls AI was, until about thirty years ago, excluded from the club. The story of how neural networks and other ML techniques won admission has not yet found its chronicler. What is clear is that a steep funding drop-off in the 1980s (the so-called "AI winter") made the once-excluded machine-learning rival—its predictive successes on display over subsequent decades—a very attractive aide to winning back the grant money.

winners" in the generative AI revolution. Maybe. But either way, expect Springer Nature, Taylor & Francis, Elsevier, Wiley, and SAGE to test the theory.

## Early Deals

Sure enough, the publishing giants have begun to quietly strike deals with the tech firms. These agreements follow on the heels of high-profile deals made by news outlets, music companies, and social sites like Reddit (Heikkilä 2024; Brown 2024). The logic is the same: Scholarly publishers are selling the right to train LLMs on their copyrighted books and articles. One such deal surfaced in May, with Taylor & Francis set to receive an initial $10 million, and then "recurring payment[s]" through 2027, from Microsoft (Informa 2024a, 1, 3). In a July investor update, the publisher's parent company, Informa, cited "further momentum" in licensing its "unique specialist content," with 2024 revenues topping $75 million (2024b, 2). Wiley (2024), for its part, has struck content rights agreements with a pair of unnamed "large tech compan[ies]," worth a reported $45 million (Milliot 2024). According to reports, neither firm notified authors about the licensing of their works (Wood 2024); when news of the Taylor & Francis agreement surfaced in the summer, researchers decried the deal in social media posts and opinion pieces (Palmer 2024). As one scholar told *Inside Higher Ed*:

> I've come to terms with the fact, as an author who has published, that at some point my work is going to go into AI, whether that's through an illegal copy published somewhere on the internet or some other means. I just didn't expect it to be my publisher. (quoted in Palmer 2024)

Additional deals are likely to be announced in the coming months in the big publishers' annual reports and other financial disclosures—though two of the oligopolist firms (Springer Nature and SAGE) are privately held and may withhold details. Meanwhile, "reproduction rights" groups like the Copyright Clearance Center (2024) are busy building so-called "collective licensing" platforms, with the aim to streamline the collection of licensing fees from tech companies. On this model, publishers get their training data revenues via third-party platforms—in place of, or in addition to, one-off deals (Cader 2024; on datasets, see Paul 2024).[2]

So the campaign to extract another layer of profit from authors' work— what Lance Eaton (2024) aptly calls "academic fracking"—is well underway. As a recent headline in *Nature* (Gibney 2024, 715) put it, "Has your paper been used to train an AI model? Almost certainly." And the big scholarly publishers are, as expected, pitching their "datasets" as uniquely trustworthy.

## Hallucinating Parrots

Truly large language models, like those driving ChatGPT and Bard, are notorious fabulists. They routinely, and confidently, return what the industry euphemism terms "hallucinations." Some observers expect the problem to keep getting worse as LLM-generated material floods the internet. The big models, on this fear, will feed on their own falsehood-ridden prose in subsequent training rounds—a kind of large-language cannibalism that, over time, could crowd out whatever share of the pre-LLM web that was more or less truthful (see Shumailov et al. 2024).

One solution to the problem, with gathering VC and hype-cycle momentum, is a turn to so-called "small" language models (Criddle and Murgia 2024). The idea is to apply the same pattern-recognition techniques, but on curated, domain-specific datasets. One advantage of the smaller models, according to proponents, is their ability to restrict training data to the known and verifiable. The premise is that, with less garbage in, there will be less garbage out.

So it is no surprise that the published scientific record has emerged, in the industry chatter, as an especially promising hallucination slayer (e.g., Matei 2023). Here is a body of vetted knowledge, the thinking goes, cordoned off from the internet's Babelist free-for-all. What makes the research corpus different is, well, peer review and editorial gatekeeping, together with citation conventions and scholars' putative commitment to a culture of self-correcting criticism. Thus the published record is—among bodies of minable text—uniquely trustworthy. Or that is what small-language evangelists are claiming.

Enter Elsevier and its oligopolistic peers (Lawton 2023). They guard (with paywalled vigilance) a large share of published scholarship, much of which is unscrapable. A growing proportion of their total output is, it is true, open access, but a large share of that material carries a non-commercial license. Standard OA agreements tend to grant publishers blanket rights, so they have a claim—albeit one contested on fair-use grounds by OpenAI and the like—to exclusive exploitation (Kaufman 2023a). Even the balance of OA works that

[2] In a recent column, scholarly publishing veteran Richard Charkin (2024) could hardly contain his excitement at the prospect of collective-licensing fees. Calling reproduction rights groups like the Copyright Clearance Center "our unsung heroes," he wrote that the approach is not only "desirable in its own right," but also "contains the seeds of future profit opportunity."

permit commercial reuse are corralled with the rest on propriety platforms like Elsevier's ScienceDirect. Those platforms also track researcher behavior, like downloads and citations, that can be used to tune their models' outputs (Yoose and Shockey 2023). Such models could, in theory, be fed by proprietary bibliographic platforms, such as Clarivate's Web of Science, Elsevier's Scopus, and Digital Science's Dimensions (owned by Springer Nature's parent company).

## "The World's Largest Collection"

One area where a number of big publishers are already jumping in is search-based summary (Van Noorden 2023). Elsevier (2024), for example, recently released ScopusAI. Researchers type in natural-language questions, and they get a summary spit out, with some suggested follow-up questions and references—those open a ScienceDirect view in the sidebar. ScopusAI results also include a "Concept Map"—an expandable topic-based tree, presumably powered by the firm's Fingerprint keywords.

The tool is combing its Scopus titles and abstracts—from 2018 on—and then feeding the top ten or so results into an OpenAI GPT model for summarizing. Elsevier is not shy about its data trove advantage: Scopus AI is "built on the world's largest collection of trusted peer-reviewed academic literature," proclaims a splashy promotional video (Elsevier 2023).

Springer Nature and Clarivate are also in on the search-summary game. Dimensions, the Scopus competitor from Springer Nature's corporate sibling Digital Science, has a Dimensions AI Assistant in trials (Dimensions 2024). Like Scopus AI, the Dimensions tool is retrieving a small number of abstracts based on conversational search, turning to models from OpenAI and Google for the summaries. More recently, Digital Science has released a Dimensions Research GPT chatbot (Linacre 2024)

Meanwhile, Clarivate—which owns Web of Science and ProQuest—unveiled Clarivate Academic AI Platform, which it describes as the "technology background" for a series of products including Web of Science Research Assistant and Alethea Academic Coach. The whole suite of "solutions" is, as the company boasts, "grounded in our extensive collection of curated scholarly content" (Ben-Porat 2024). Clarivate has also struck a deal with AI21 Labs, an Israeli LLM startup (tagline: "When Machines Become Thought Partners") (Clarivate 2023). Using Clarivate's "trusted content as the foundation," AI21 promises to use its models to generate "high-quality, contextual-based answers and services," with what it frankly calls "Clarivate's troves of content and data" (quoted in Clarivate 2023).

The big firms will be competing with a stable of VC-backed startups, including Ought ("Scale up good reasoning"), Iris.ai ("The Researcher Workspace"), SciSummary ("Use AI to summarize scientific articles in seconds"), Petal ("Chat with your documents"), Jenni ("Supercharge Your Next Research Paper"), Scholarcy ("The AI-powered article summarizer"), Imagetwin ("Increase the Quality in Science"), Keenious ("Find research relevant to any document!"), and Consensus ("AI Search Engine for Research").

An open question is if the startups can compete with the big publishers; many are using the open-access database Semantic Scholar, which excludes the full text of paywalled articles (Lo et al. 2020). They have won plenty of venture-capital backing, but—if the wider AI industry is any guide—the startups will face an uphill climb to stay independent. Commercial AI, after all, is dominated by a handful of giant US and Chinese corporations, nearly all big tech incumbents (Kak et al. 2023; Widder et al. 2023). The industry has ferocious economies of scale, largely because model-building takes enormous financial and human resources.

The big publishers may very well find themselves in a similar pole position. The firms' stores of proprietary full-text papers and other privately held data are a built-in advantage. Their astronomical margins on legacy subscription- and APC-publishing businesses means that they have the capital at hand to invest and acquire. Elsevier's decade-long acquisition binge was, in that same way, financed by its lucrative earnings. There is every reason to expect that the company will fund its costly LLM investments from the same surplus; Elsevier's peers are likely to follow suit. Thus universities and taxpayers are serving, in effect, as a capital fund for AI products that, in turn, will be sold back to us. The independent startups may well be acquired along the way. The giant publishers *themselves* may be acquisition targets to the even larger Silicon Valley firms hungry for training data—as Avi Staiman (2023) observed in *The Scholarly Kitchen*.

The acquisition binge has already begun. Last fall, Springer Nature acquired the Science division of Slimmer AI, a Dutch "AI venture studio" that the publisher has worked with since 2015 on peer-review and plagiarism-detection tools (Springer Nature 2023a). Digital Science, meanwhile, bought Writefull ("Digital Science" 2023), which makes an academic writing assistant. Digital Science pitched the acquisition as a small-language-model play: "While the broader focus is currently on LLMs," said a company executive in the press release, "Writefull's small, specialized models offer more flexibility, at lower cost, with auditable metrics" ("Digital Science" 2023). Research Solutions, a Nevada company that sells access to the big commercial

publishers' paywalled content to corporations, recently bought scite, a startup whose novel offering—citations *contexts* (Nicholson et al. 2021)—has been repackaged as "ChatGPT for research" (Research Solutions 2023).

## Fair Use?

As the *Times* lawsuit suggests, there is a big legal question mark hovering over the big publishers' AI prospects. The key issue, winding its way through the courts, is fair use: Can the likes of OpenAI scrape up copyrighted content into their models without permission or compensation? The Silicon Valley tech companies think so; they are fresh converts to fair-use maximalism, as revealed by their public comments filed with the US Copyright Office (Brewster 2023). The companies' "overall message," reported *The Verge* in a roundup, is that "they don't think they should have to pay to train AI models on copyrighted work" (Davis 2023). Artists and other content creators have begged to differ, filing a handful of high-profile lawsuits.

The publishers have not filed their own suits yet, but they are certainly watching the cases carefully. Wiley, for one, told *Nature* that it was "closely monitoring industry reports and litigation claiming that generative AI models are harvesting protected material for training purposes while disregarding any existing restrictions on that information" (quoted in Conroy 2023). The firm has called for audits and regulatory oversight of AI models to address the "potential for unauthorised use of restricted content as an input for model training" (Williams 2023). Elsevier, for its part, has banned the use of "our content and data" for training (Williams 2023); its sister company, LexisNexis, likewise, recently emailed customers to "remind" them that feeding content to "large language models and generative AI" is forbidden (quoted in Powers 2023). The Copyright Clearance Center (CCC), in its own comments to the US Copyright Office, took a predictably muscular stance on the question:

> There is certainly enough copyrightable material available under license to build reliable, workable, and trustworthy AI. Just because a developer wants to use "everything" does not mean it needs to do so, is entitled to do so, or has the right to do so. Nor should governments and courts twist or modify the law to accommodate them. (Kaufman 2023b)

The for-profit CCC is the publishing industry's main licensing and permission enforcer. Big tech and the commercial publishing giants are already maneuvering for position (Brewster 2023). As Joseph Esposito (2023), a keen observer of scholarly publishing, put the point: "Scientific publishers in particular, may have a special, remunerative role to play here."

One near-term consequence may be a shift in the big publishers' approach to open access. The companies are already updating their licenses and terms to forbid commercial AI training—for anyone but them, of course. The companies could also pull back from OA altogether, to keep a larger share of exclusive content to mine. Esposito (2023) made the argument explicit in a recent *Scholarly Kitchen* post: "The unfortunate fact of the matter is that the OA movement and the people and organizations that support it have been co-opted by the tech world as it builds content-trained AI." Publishers need "more copyright protection, not less," he added (Esposito 2023). Esposito's consulting firm, in its newsletter, called the liberal Creative Commons Attribution (CC BY) license a "mechanism to transfer value from scientific and scholarly publishers to the world's wealthiest tech companies" (Clarke & Esposito 2023). Perhaps, though I would preface the point: *Commercial scholarly publishing* is a mechanism to transfer value from scholars, taxpayers, and universities to the world's most profitable companies.

Another likely knock-on effect is that scholars' access to the open web and other large-text data sources—to conduct text mining and other research—may be increasingly restricted. The reason is that content owners like news outlets and social media sites are updating their terms of service and robots.txt instructions to prevent big AI firms from crawling their pages. Web publishers, as a recent MIT study reported (Longpre et al. 2024), are rapidly locking down their content. University-based researchers are getting caught up in the dragnet, with the result that they may be legally barred from studying large swaths of the web (Koebler 2024).

## The Matthew Effect in AI

There are a hundred and one reasons to worry about Elsevier mining our scholarship to maximize its profits. I want to linger on what is, arguably, the most important: the potential effects on knowledge itself. At the core of these tools—including a predictable avalanche of as-yet-unannounced products—is a series of verbs: to surface, to rank, to summarize, and to recommend. The object of each verb is *us*—our scholarship and our behavior. What is at stake is the kind of knowledge that the models surface, and *whose* knowledge.

AI models are poised to serve as knowledge arbitrators (Gendron et al. 2022) by picking winners and losers according to what they make visible. There are two big and interlocking problems with this role: The models are trained on the past, and their filtering logic is inscrutable. As a result, they may smuggle in the many biases that mark the history of scholarship, around gender, geography, and other lines of difference. In this context it is useful to revive an old concept in the sociology of science. According to the Matthew Effect—named by Robert Merton (1968) decades ago—prominent and well-cited scholars tend to receive still more prominence and citations. The flip side is that less-cited scholars tend to slip into further obscurity over time. ("For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away"—Matthew 25:29.) These dynamics of cumulative advantage have, in practice, served to amplify the knowledge system's patterned inequalities—for example, in the case of gender and twentieth-century scholarship, aptly labeled the Matilda Effect by Margaret Rossiter (1993).

The deployment of AI models in science, especially proprietary ones, may produce a Matthew Effect on the scale of Scopus, and with no paper trail. The problem is analogous to the well-documented bias-smuggling with existing generative models; image tools trained on, for example, mostly white and male photos reproduce the skew in their prompt-generated outputs. With our bias-laden scholarship as training data, the academic models may spit out results that, in effect, double down on inequality. What is worse is that we will not really know, due to the models' black-box character. Thus the tools may act as laundering machines—context-erasing abstractions that disguise their probabilistic "reasoning." Existing biases, like male academics' propensity for self-citation, may win a fresh coat of algorithmic legitimacy. Or consider center-periphery dynamics along North-South and native-English-speaking lines: Gaps traceable to geopolitical history, including the legacy of European colonialism, may be buried still deeper. The models, in short, could serve as privilege multipliers.

The AI models are not going away, but we should demand that—to whatever extent possible—the tools and models are subject to scrutiny and study (Hardinges et al. 2023). This means ruling out propriety products unless they can be pried open by law or regulation. We should, meanwhile, bring the models in house, within the academic fold, using mission-aligned collections like the Open University's CORE and the Allen Institute's Semantic Scholar. Academy-led efforts to build non-profit models and tools should be transparent, explainable, and auditable.

## Stop Tracking Scholarship

These are early days. The legal uncertainty, the vaporware, the breathless annual-report prose: All of it points to aspiration and C-suite prospecting. We are not yet living in a world of publisher small-language models, trained on our work and behavior.

Still, I am convinced that the big five publishers, plus Clarivate, will make every effort to pad their margins with fresh AI revenue. My guess is that they will develop and acquire their way to a product portfolio up, down, and around the research lifecycle, on Elsevier's existing full-stack model. After all—and depending on what we mean by AI—the commercial publishers have been rolling out AI products for years. Every signal suggests they will pick up the pace, with hype-driven pursuit of GPT-style language models in particular. They will sell their own products back to us, and—I predict—license our papers out to the big foundation models, court willing.

So it is an urgent task to push back now, and not wait until after the models are trained and deployed. What is needed is a full-fledged campaign, leveraging activism and legislative pressure, to challenge the commercial publishers' extractive agenda. One crucial framing step is to treat the impending AI avalanche as continuous with—as an extension of—the publishers' in-progress mutation into surveillance-capitalist data businesses. The surveillance publisher era was symbolically kicked off in 2015, when Reed-Elsevier adopted its "shorter, more modern name" RELX Group to mark its "transformation" from publisher to "technology, content and analytics-driven business" (RELX Group 2015, 5). They have made good on the promise, skimming scholars' behavioral cream with product-by-product avidity. Clarivate and Elsevier's peers have followed their lead.

Thus the turn to AI is more of the same, only more so. The publishers' cocktail of probability, prediction, and profit is predicated on the same process: extract our scholarship and behavior, then sell it back to us in congealed form. The stakes are higher given that some of the publishers are embedded in data-analytics conglomerates—RELX (Elsevier) and Informa (Taylor & Francis), joined by publisher-adjacent firms like Clarivate and Thomson Reuters. Are the companies cross-pollinating their academic and "risk solutions" businesses? RELX's LexisNexis sold face-tracking and other surveillance tools to the US Customs and Border Protection last year, as *The Intercept* reported (Biddle 2023). As SPARC (the library alliance) put it in its recent report on Elsevier's ScienceDirect platform: "There is little to nothing to stop vendors who collect and track

[library] patron data from feeding that data—either in its raw form or in aggregate—into their data brokering business" (Yoose and Shockey 2023, 33).

So far the publishers' data hoovering has not galvanized scholars to protest. The main reason is that most academics are blithely unaware of the tracking—no surprise, given scholars' too-busy-to-care ignorance of the publishing system itself. The library community is far more attuned to the unconsented pillage, though librarians—aside from SPARC (2021) and the Library Freedom Project (n.d.)—have not organized on the issue (see Bettinger et al. 2023). There have been scattered notes of dissent, including a Stop Tracking Science petition, and an outcry from Dutch scholars on a 2020 data-and-publish agreement with Elsevier (Knecht 2020), largely because the company had baked its prediction products into the deal. In 2021, the German national research foundation, Deutsche Forschungsgemeinschaft (DFG), released its own report-cum-warning—"industrialisation of knowledge through tracking," in the report's words (Ausschuss für Wissenschaftliche 2021, 8). Sharp critiques from, among others, Björn Brembs (2021), Leslie Chan (Chen and Chan 2021), Michael Freiberg (2022), Renke Siems (2021), Lai Ma (2023), and Sarah Lamdan (2022) have appeared at regular intervals.

None of this has translated into much, not even awareness among the larger academic public. A coordinated campaign of advocacy and consciousness-raising should be paired with high-quality, in-depth studies of publisher data harvesting—on the example of SPARC's recent ScienceDirect report (Yoose and Shockey 2023). Any effort like this should be built on the premise that another scholarly-publishing world is possible. Our prevailing joint-custody arrangement—for-profit publishers and non-profit universities—is a recent and reversible development. There are lots of good reasons to restore custody to the academy. The latest is to stop our work from fueling the publishers' AI profits.

## Acknowledgments

An earlier version of this essay was published as "Large Language Publishing" in *Upstream* on January 2, 2024, which carries a CC BY 4.0 license.

## Competing Interests

The author declares that they have no competing interests.

## References

Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme. 2021. *Data Tracking in Research: Aggregation and Use or Sale of Usage Data by Academic Publishers. A Briefing Paper of the Committee on Scientific Library Services and Information Systems of the Deutsche Forschungsgemeinschaft*. DFG, German Research Foundation. https://doi.org/10.5281/zenodo.5937995.

Ben-Porat, Guy. 2024. "Introducing the Clarivate Academic AI Platform." Press release, May 21, 2024. https://clarivate.com/blog/introducing-the-clarivate-academic-ai-platform/. Archived at: https://perma.cc/6K4Z-U98B.

Bettinger, Eliza C., Meryl Bursic, and Adam Chandler. 2023. *Disrupting the Digital Status Quo: Why and How to Staff for Privacy in Academic Libraries*. Licensing Privacy Project. https://publish.illinois.edu/licensing-privacy/files/2023/06/Whitepaper-on-Privacy-Staffing-Licensing-Privacy.pdf. Archived at: https://perma.cc/T7EF-RLN6.

Biddle, Sam. 2023. "LexisNexis Sold Powerful Spy Tools to U.S. Customs and Border Protection." *The Intercept*, November 16, 2023. https://theintercept.com/2023/11/16/lexisnexis-cbp-surveillance-border/.

Brembs, Björn. "Algorithmic Employment Decisions in Academia?" *björn.brembs.blog*, September 23, 2021. http://bjoern.brembs.net/2021/09/algorithmic-employment-decisions-in-academia/. Archived at: https://perma.cc/J6K8-NNPR.

Brewster, Freddy. 2023. "Big Tech Is Lobbying Hard to Keep Copyright Law Favorable to AI." *Jacobin*, November 21, 2023. https://jacobin.com/2023/11/artificial-intelligence-big-tech-lobbying-copyright-infringement-regulation/. Archived at: https://perma.cc/5MBV-A2LX.

Brown, Pete. 2024. "Licensing Deals, Litigation Raise Raft of Familiar Questions in Fraught World of Platforms and Publishers." *Columbia Journalism Review*, May 22, 2024. https://www.cjr.org/Tow_Center/Licensing-Deals-Litigation-Raise-Raft-of-Familiar-Questions-in-Fraught-World-of-Platforms-and-publishers.php. Archived at: https://perma.cc/WTM5-3UB3.

Cader, Michael. 2024. "Former Scribd Co-Founder Launches AI Licensing Company for Books." Publishers Lunch, June 25, 2024. https://lunch.publishersmarketplace.com/2024/06/former-scribd-co-founder-launches-ai-licensing-company-for-books/.

Charkin, Richard. 2024. "In Praise of Collective Licensing." *Publishing Perspectives*, July 29, 2024. https://publishingperspectives.com/2024/07/Richard-Charkin-Collective-Licensing/. Archived at: https://perma.cc/MZ9M-NDL7.

Chen, George, and Leslie Chan. 2021. "University Rankings and Governance by Metrics and Algorithms." In *Research Handbook on University Rankings: Theory, Methodology, Influence, and Impact*, edited by Ellen Hazelkorn and Georgiana Mihut. Edward Elgar. https://doi.org/10.4337/9781788974981.

Clarivate. 2023. "Clarivate Announces Partnership with AI21 Labs as Part of Its Generative AI Strategy to Drive Growth." Press release, June 22, 2023. https://ir.clarivate.com/news-events/press-releases/news-details/2023/Clarivate-Announces-Partnership-with-AI21-Labs-as-part-of-its-Generative-AI-Strategy-to-Drive-Growth/default.aspx.

Clarke & Esposito. 2023 "Gemini." *The Brief*, December 29, 2023. https://www.ce-strategy.com/the-brief/gemini/. Archived at: https://perma.cc/Y63C-GQKD.

Conroy, Gemma. 2023. "How ChatGPT and Other AI Tools Could Disrupt Scientific Publishing." *Nature* 622 (7982): 234–36. https://doi.org/10.1038/d41586-023-03144-w.

Copyright Clearance Center. 2024. "CCC Pioneers Collective Licensing Solution for Content Usage in Internal AI Systems." Press release, July 16, 2024. https://www.copyright.com/media-press-releases/ccc-pioneers-collective-licensing-solution-for-content-usage-in-internal-ai-systems/. Archived at: https://perma.cc/GRC2-PMBD.

Criddle, Cristina, and Madhumita Murgia. 2024. "Artificial Intelligence Companies Seek Big Profits From 'Small' Language Models." *Financial Times*, May 20, 2024. https://www.ft.com/content/359a5a31-1ab9-41ea-83aa-5b27d9b24ef9.

Davis, Wes. 2023. "AI Companies Have All Kinds of Arguments Against Paying for Copyrighted Content." *The Verge*, November 4, 2023. https://www.theverge.com/2023/11/4/23946353/generative-ai-copyright-training-data-openai-microsoft-google-meta-stabilityai. Archived at: https://perma.cc/LD4W-DD2T.

"Digital Science Acquires AI Service Writefull." *Research Information*, November 23, 2023. https://www.researchinformation.info/news/digital-science-acquires-ai-service-writefull. Archived at: https://perma.cc/SJY6-DHZA.

Dimensions. 2024. "Discover Dimensions AI Assistant." https://www.dimensions.ai/discover-dimensions-ai-assistant/. Archived at: https://perma.cc/Y94F-H3UU.

Eaton, Lance. 2024. "Academic Fracking: When Publishers Sell Scholars Work to AI." *AI + Education = Simplified*, July 31, 2024. https://aiedusimplified.substack.com/p/academic-fracking-when-publishers.

Elsevier. 2024. "Scopus AI: Trusted Content. Powered by Responsible AI." https://www.elsevier.com/products/scopus/scopus-ai. Archived at: https://perma.cc/VY3V-67DC.

Elsevier. 2023. "ScopusAI: Change the Way You View Knowledge." Video, 2 min., 25 sec. Accessed December 16, 2023. https://www.elsevier.com/products/scopus/scopus-ai.

Esposito, Joseph. 2023. "Who Is Going to Make Money from Artificial Intelligence in Scholarly Communications?" *The Scholarly Kitchen*, July 12, 2023. https://scholarlykitchen.sspnet.org/2023/07/12/who-is-going-to-make-money-from-artificial-intelligence-in-scholarly-communications/. Archived at: https://perma.cc/N4ZS-BJ3H.

Freiberg, Michael. 2022. "Third-Party-Tracking bei Wiley und Springer: Analyse und Ausblick." *ABI Technik* 42 (2): 96–104. https://doi.org/10.1515/Abitech-2022-0017.

Gendron, Yves, Jane Andrew, and Christine Cooper. 2022. "The Perils of Artificial Intelligence in Academic Publishing." *Critical Perspectives on Accounting* 87 (September): 102411. https://doi.org/10.1016/j.cpa.2021.102411.

Gibney, Elizabeth. 2024. "Has Your Paper Been Used to Train an AI Model? Almost Certainly." *Nature* 632 (8026): 715–16. https://doi.org/10.1038/D41586-024-02599-9.

Grynbaum, Michael M., and Ryan Mac. 2023. "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work." *The New York Times*. December 17, 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Hardinges, Jack, Elena Simperl, and Nigel Shadbolt. 2024. "We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models." *Harvard Data Science Review* (Special Issue 5). https://doi.org/10.1162/99608f92.a50ec6e6.

Heikkilä, Melissa. 2024. "AI Companies Are Finally Being Forced to Cough Up for Training Data." *MIT Technology Review*, July 2, 2024. https://www.technologyreview.com/2024/07/02/1094508/Ai-Companies-Are-Finally-Being-Forced-to-Cough-Up-for-Training-Data/. Archived at: https://perma.cc/2RVG-JKX9.

Informa. 2024a. "Market Update: Continuing Momentum and Growth." Press release, May 8, 2024. https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc---market-update.pdf. Archived at: https://perma.cc/2FW5-BZAL.

Informa. 2024b. "Informa PLC 2024 Half-Year Results." Press release, July 24, 2024. https://www.informa.com/globalassets/documents/investor-relations/2024/informa-2024-half-year-results.pdf. Archived at: https://perma.cc/7WL4-UDXC.

Kak, Amba, Sarah Myers West, and Meredith Whittaker. 2023. "Make No Mistake—AI Is Owned by Big Tech." *MIT Technology Review*, December 5, 2023. https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/. Archived at: https://perma.cc/7FJV-PSG4.

Kaufman, Roy. 2023a. "Some Thoughts on Five Pending AI Litigations — Avoiding Squirrels and Other AI Distractions." *The Scholarly Kitchen*, March 7, 2023. https://scholarlykitchen.sspnet.org/2023/03/07/some-thoughts-on-five-pending-ai-litigations-avoiding-squirrels-and-other-ai-distractions/. Archived at: https://perma.cc/QPR3-F5ZS.

Kaufman, Roy. 2023b. "The United States Copyright Office Notice of Inquiry on AI: A Quick Take." *The Scholarly Kitchen*, November 28, 2023. https://scholarlykitchen.sspnet.org/2023/11/28/the-united-states-copyright-office-notice-of-inquiring-on-ai-a-quick-take/. Archived at: https://perma.cc/HX53-Q8AP.

Knecht, Sicco de. 2020. "Dutch Open Science Deal Primarily Benefits Elsevier." *ScienceGuide*, June 29, 2020. https://www.scienceguide.nl/2020/06/open-science-deal-benefits-elsevier/, https://www.scienceguide.nl/2020/06/open-science-deal-benefits-elsevier/. Archived at: https://perma.cc/DR3F-ZFKU.

Koebler, Jason. 2024. "The Backlash Against AI Scraping Is Real and Measurable." *404 Media*, July 23, 2024. https://www.404media.co/the-Backlash-Against-Ai-Scraping-Is-Real-and-Measurable/. Archived at: https://perma.cc/4MHN-CUBC.

Lamdan, Sarah. 2022. *Data Cartels: The Companies That Control and Monopolize Our Information*. Stanford University Press.

Library Freedom Project. n.d. "About Library Freedom Project." https://libraryfreedom.org/lfp-values/. Accessed August 20, 2024. Archived at: https://perma.cc/9QDY-LSRT.

Linacre, Simon. 2024. "Dimensions Research GPT – Evidence-Based Research Insights for ChatGPT Platform Users." Digital Science News Room, February 28, 2024. https://www.digital-science.com/news/dimensions-research-gpt. Archived at: https://perma.cc/UD93-VC4P.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. "S2ORC: The Semantic Scholar Open Research Corpus." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.447.

Longpre, Shayne, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu et al. 2024. *Consent in Crisis: The Rapid Decline of the AI Data Commons. Data Provenance Initiative*. https://doi.org/10.48550/arXiv.2407.14933.

Lawton, George. 2023. "Elsevier Sees Promise in Small Language Models and Graph Data." *diginomica*, March 24, 2023. https://diginomica.com/reed-elsevier-sees-promise-small-language-models-and-graph-data. Archived at: https://perma.cc/9EXT-RSMB.

Ma, Lai. 2023. "The Platformisation of Scholarly Information and How to Fight It." *LIBER Quarterly: The Journal of the Association of European Research Libraries* 33 (1): 1–20. https://doi.org/10.53377/lq.13561.

Matei, Sorin Adam. 2023. "An Academic ChatGPT Needs a Better Schooling." *Times Higher Education*, November 28, 2023. https://www.timeshighereducation.com/blog/academic-chatgpt-needs-better-schooling. Archived at: https://perma.cc/HQQ3-V34R.

Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." *Science* 159 (3810): 56–63. https://doi.org/10.1126/science.159.3810.56.

Milliot, Jim. 2024. "Wiley Looks Ahead After a Transitional Fiscal 2024." *Publishers Weekly*, June 13, 2024. https://www.publishersweekly.com/Pw/by-Topic/Industry-News/Financial-Reporting/Article/95272-Wiley-Looks-Ahead-After-a-Transitional-Fiscal-2024.html. Archived at: https://perma.cc/G9KG-LB5T.

Nicholson, Josh M., Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. "scite: A Smart Citation Index That Displays the Context of Citations and Classifies Their Intent Using Deep Learning." *Quantitative Science Studies* 2 (3): 882–98. https://doi.org/10.1162/qss_a_00146.

Palmer, Kathryn. 2024. "Taylor & Francis AI Deal Sets 'Worrying Precedent' for Academic Publishing." *Inside Higher Ed*, July 29, 2024. https://www.insidehighered.com/News/Faculty-Issues/Research/2024/07/29/Taylor-Francis-Ai-Deal-Sets-Worrying-Precedent. Archived at: https://perma.cc/6XHC-ZYXA.

Paul, Katie. 2024. "AI Dataset Licensing Companies Form Trade Group." *Reuters*, June 26, 2024. https://www.reuters.com/technology/artificial-intelligence/ai-dataset-licensing-companies-form-trade-group-2024-06-26/.

Pooley, Jeff. 2024. "Large Language Publishing." *Upstream*, January 2, 2024. https://doi.org/10.54900/zg929-e9595.

Pooley, Jeff. 2022. "Surveillance Publishing." *The Journal of Electronic Publishing* 25 (1): 39–49. https://doi.org/10.3998/jep.1874.

Powers, Melanie Padgett. "Generative AI Meets Scientific Publishing." *Optica*, October 1, 2023. https://www.optica-opn.org/home/articles/volume_34/october_2023/features/generative_ai_meets_scientific_publishing/. Archived at: https://perma.cc/6A6M-P6SC.

RELX Group. 2015. *Annual Report and Financial Statements 2014*. https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/2014-annual-report.pdf. Archived at: https://perma.cc/E4HA-Y4ZT.

Research Solutions. 2023 "Research Solutions Announces Acquisition of scite." Press release, November 27, 2023. https://www.researchsolutions.com/resources/press-releases/research-solutions-announces-acquisition-of-scite. Archived at: https://perma.cc/35MY-THAJ.

Rossiter, Margaret W. 1993. "The Matthew Matilda Effect in Science." *Social Studies of Science* 23 (2): 325–41. https://doi.org/10.1177/030631293023002004.

Siems, Renke. 2021. "When Your Journal Reads You: User Tracking on Science Publisher Platforms." *Elephant in the Lab*. https://zenodo.org/record/4683778#.Y1A0xi8RpQI.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631 (8022): 755–59. https://doi.org/10.1038/S41586-024-07566-Y.

SPARC. 2021. "Addressing the Alarming Systems of Surveillance Built By Library Vendors." April 9, 2021. https://sparcopen.org/news/2021/addressing-the-alarming-systems-of-surveillance-built-by-library-vendors/. Archived at: https://perma.cc/SRC4-VHLJ.

Springer Nature. 2023a. "Springer Nature Expands Its AI Capability with Acquisition of Slimmer AI's Science Division." Press release, October 25, 2023. https://group.springernature.com/gp/group/media/press-releases/acquisition-slimmer-ai-science-division/26215608. Archived at: https://perma.cc/S84R-45VA.

Springer Nature. 2023b. "Springer Nature Introduces Curie, Its AI-Powered Scientific Writing Assistant." Press release, October 13, 2023. https://group.springernature.com/gp/group/media/press-releases/ai-powered-scientific-writing-assitant-launched/26176230. Archived at: https://perma.cc/QP7P-SQ5S.

Staiman, Avi. 2023. "Will Building LLMs Become the New Revenue Driver for Academic Publishing?" *The Scholarly Kitchen*, August 8, 2023. https://scholarlykitchen.sspnet.org/2023/08/08/will-building-llms-become-the-new-revenue-driver-for-academic-publishing/. Archived at: https://perma.cc/W2UR-DNR4.

The New York Times Company v. Microsoft Corp. et al. Complaint, United States District Court, Southern District of New York (Case 1:23-cv-11195), December 27, 2023. https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf. Archived at: https://perma.cc/E4BA-RBTK.

Van Noorden, Richard. 2023. "ChatGPT-like AIs Are Coming to Major Science Search Engines." *Nature* 620 (7973). https://doi.org/10.1038/d41586-023-02470-3.

Widder, David Gray, Sarah West, and Meredith Whittaker. 2023. "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." Preprint, Social Science Research Network, August 18, 2023. https://doi.org/10.2139/ssrn.4543807.

Wiley. 2024. "Wiley Increases Quarterly Dividend for the 31st Consecutive Year." Press release, June 27, 2024. https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-Increases-Quarterly-Dividend-for-the-31st-Consecutive-Year/default.aspx.

Williams, Tom. 2023. "Publishers Seek Protection from AI Mining of Academic Research." *Times Higher Education*, August 3, 2023. https://www.timeshighereducation.com/news/publishers-seek-protection-ai-mining-academic-research.

Wood, Heloise. 2024. "Wiley and Oxford University Press Confirm AI Partnerships as Cambridge University Press Offers 'Opt-in.'" *The Bookseller*, August 1, 2024. https://www.thebookseller.com/News/Wiley-Cambridge-University-Press-and-Oxford-University-Press-Confirm-Ai-Partnerships. Archived at: https://perma.cc/8E8Y-MSTG.

Yoose, Becky, and Nick Shockey. 2023. "Navigating Risk in Vendor Data Privacy Practices: An Analysis of Elsevier's ScienceDirect." SPARC. https://doi.org/10.5281/zenodo.10078610.

Zhavoronkov, Alex. 2023. "The Unexpected Winners of the ChatGPT Generative AI Revolution." *Forbes*, February 23, 2023. https://www.forbes.com/sites/alexzhavoronkov/2023/02/23/the-unexpected-winners-of-the-chatgpt-generative-ai-revolution/. Archived at: https://perma.cc/98Y6-ZNQQ.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

**How to cite this article:** Pooley, Jefferson. 2024. Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7(1). https://doi.org/10.18357/kula.291

**Submitted:** 24 April 2024 **Accepted:** 18 September 2024 **Published:** 09 October 2024



*KULA: Knowledge Creation, Dissemination, and Preservation Studies* is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS