An academic publisher has struck an AI data deal with Microsoft—without their authors' knowledge

July 23 2024, by Wellett Potter

Credit: Unsplash/CC0 Public Domain

In May, a multibillion-dollar UK-based multinational called Informa announced in a **trading update** that it had signed a deal with Microsoft

involving "access to advanced learning content and data, and a partnership to explore AI expert applications." Informa is the parent company of Taylor & Francis, which publishes a wide range of academic and technical books and journals, so the data in question may include the content of these books and journals.

According to reports published last week, the authors of the content do not appear to have been asked or even informed about the deal. What's more, they say they had no opportunity to opt out of the deal, and will not see any money from it.

Academics are only the latest of several groups of what we might call content creators to take umbrage at having their work ingested by the generative AI models currently racing to hoover up the products of human culture. Newspapers, visual artists and record labels are already taking AI companies to court.

While it's unclear how Informa will react to the rumblings of discontent, the deal is a reminder to authors to be aware of the contractual terms of the publishing agreements they sign.

What's in the Informa deal?

Informa's update stated four focus areas of the Microsoft deal:

- increasing Informa's own productivity
- developing an automated citation tool
- developing AI-powered research assistant software (perhaps like a system being tested by online academic library JSTOR)
- giving Microsoft data access to "help improve relevance and performance of AI systems."

Informa will be paid more than £8 million (A$15.5 million) for initial

access to the data, followed by recurring payments of an unspecified amount for the next three years.

We don't know exactly what Microsoft plans to do with its data access, but a likely scenario is that the content of academic books and articles would be added to the training data of ChatGPT-like generative AI models. In principle this should make the output of the AI systems more accurate, though existing AI models have faced heavy criticism, not only for [regurgitating training data](#) without citation (which can be viewed as a kind of [plagiarism](#)), but also for [inventing false information](#) and [attributing](#) it to real sources.

However, the update also says "the agreement protects [intellectual property rights](#), including limits on verbatim text extracts and alignment on the importance of detailed citation references."

The "limits on verbatim text extracts" mentioned likely pertains to the [US doctrine of fair use](#), which permits certain uses of copyright-protected material.

Many generative AI companies are currently facing copyright infringement lawsuits over their use of training data, and their defenses are likely to rely on claiming fair use.

The "importance of detailed citation references" may pertain to the concept of attribution in copyright. This is a [moral right](#) possessed by authors. It provides that the creator of the work should be known and attributed as the author when their work is reproduced.

How does scholarly publishing usually work?

Most academics do not receive payment or make any profit from most of their scholarly publishing. Rather, writing journal and conference

papers is usually considered part of the scope of work within a full-time, tenured position. Publication builds an academic's credibility and promotes their research.

The basic process often goes like this: an author researches and writes an original article and submits it to a journal publisher for peer review. Most peer reviewers and editorial board members also receive no payment for their work.

In fact, some journals may require authors to pay an "article processing charge" to cover editing and other costs. This can be thousands of dollars for an open access publication. Generally speaking, the more prestigious the publication, the higher the charge.

If an article passes peer review, the author will be asked to sign a publishing agreement. The terms may cover logistical arrangements such as when the article will be published, the format (print, online or both), and the division of royalties (if applicable). There will also be arrangements regarding copyright and ownership of the article.

An author usually must also grant exclusive rights to the publisher to distribute and publish the article. This may mean the author cannot publish the article elsewhere, and the publisher may also be able to sub-license the article to a third party, such as an AI company.

Sometimes publishers require an author to assign copyright in the article to them via a permanent copyright transfer agreement.

Essentially, this means the author grants all of their authorial rights as copyright holder in the work to the publisher. The publisher can then reproduce, communicate, distribute or license the work to others as they wish.

It is possible to only assign limited rights, rather than all rights, and this is something authors should consider.

Content mining

It is vital that authors understand the implications of licensing and assignment and to contemplate precisely what they are agreeing to when they sign a contract. In light of the recent trend of [publishers entering into agreements with generative AI companies](#), publishers' AI policies should also be closely scrutinized.

In the US, a standard [collective licensing solution for content use in internal AI systems](#) has recently been released, which sets out rights and remuneration for copyright holders. Similar licenses for the use of content for AI systems will likely enter the Australian market very soon.

The types of agreements being reached between academic publishers and AI companies have sparked bigger-picture concerns for many academics. Do we want scholarly research to be reduced to content for [AI knowledge mining](#)? There are no clear answers about the ethics and morals of such practices.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation