FRANKONFRAUD

The Big Shady Scrape – Are AI Companies Stealing Data To Train Models?

Publishers Cry Foul as AI Firms Allegedly Ignore 'Do Not Crawl' Signals

In a digital landscape increasingly shaped by artificial intelligence, a new battle is brewing between AI companies and content publishers.

Multiple AI firms have been accused of circumventing a long-standing web standard designed to protect publishers' content, raising alarms about the future of online information and the sustainability of journalism.

The Robots.txt Dilemma

At the heart of the controversy is the Robots Exclusion Protocol, commonly known as "robots.txt." This protocol, dating back to the mid-1990s, has long been respected as a gentlemen's agreement of the internet, allowing website owners to indicate which parts of their sites should not be crawled by search engines and other automated systems.

However, according to a letter sent to publishers by content licensing startup TollBit, "numerous" AI agents are now bypassing this protocol. This move potentially allows these companies unfettered access to valuable content without permission or compensation.

A David vs. Goliath Battle

The accusations come amid a broader debate over the value of content in the age of generative AI. Publishers, already grappling with declining revenues in the digital age, see this as yet another threat to their business models.

"Without the ability to opt out of massive scraping, we cannot monetize our valuable content and pay journalists," warns Danielle Coffey, president of the News Media Alliance, a trade group representing over 2,200 U.S.-based publishers. "This could seriously harm our industry."

The Perplexity Precedent

While TollBit's letter does not name specific offenders, the issue gained public attention through a recent dispute between AI search startup Perplexity and Forbes. The business media publisher accused Perplexity of plagiarizing its investigative stories in AI-generated summaries without attribution or permission.

A subsequent investigation by Wired magazine suggested that Perplexity was likely bypassing robots.txt restrictions, setting a concerning precedent for the industry.

The Legal Gray Area

The robots.txt protocol, while widely respected, lacks clear legal enforcement mechanisms. Some groups, including the News Media Alliance, suggest there may be legal recourse for publishers, but the path forward remains uncertain.

This legal ambiguity has led to a patchwork of responses from publishers. Some, like The New York Times, have taken the route of suing AI companies for copyright infringement. Others are opting to negotiate licensing agreements, though disagreements over content valuation persist.

As AI continues to reshape the digital landscape, the outcome of this conflict may well determine the future of online content, journalism, and the delicate balance between technological progress and intellectual property rights.