Lawsuit Takes Aim at the Way A.I. Is Built

A programmer is suing Microsoft, GitHub and OpenAI over artificial intelligence technology that generates its own computer code.

- Give this article

Tom Smith, a veteran programmer, shows how Codex can instantly generate computer code from a request in plain English. Credit Credit... Jason Henry for The New York Times

Nov. 23, 2022

In late June, Microsoft released a new kind of artificial intelligence technology that could generate its own computer code.

Called Copilot, the tool was designed to speed the work of professional programmers. As they typed away on their laptops, it would suggest ready-made blocks of computer code they could instantly add to their own.

Many programmers loved the new tool or were at least intrigued by it. But Matthew Butterick, a programmer, designer, writer and lawyer in Los Angeles, was not one of them. This month, he and a team of other lawyers filed a lawsuit that is seeking class-action status against

Microsoft and the other high-profile companies that designed and deployed Copilot.

Like many cutting-edge A.I. technologies, Copilot developed its skills by analyzing vast amounts of data. In this case, it relied on billions of lines of computer code posted to the internet. Mr. Butterick, 52, equates this process to piracy, because the system does not acknowledge its debt to existing work. His lawsuit claims that Microsoft and its collaborators violated the legal rights of millions of programmers who spent years writing the original code.

The suit is believed to be the first legal attack on a design technique called “A.I. training,” which is a way of building artificial intelligence that is poised to remake the tech industry. In recent years, many artists, writers, pundits and privacy activists have complained that companies are training their A.I. systems using data that does not belong to them.

Matthew Butterick, a programmer and lawyer, said he was concerned that work he had done was being improperly employed in new artificial intelligence systems. Credit... Tag Christof for The New York Times

The lawsuit has echoes in the last few decades of the technology industry. In the 1990s and into the 2000s, Microsoft fought the rise of open source software, seeing it as an existential threat to the future of the company's business. As the importance of open source grew, Microsoft embraced it and even acquired GitHub, a home to open source programmers and a place where they built and stored their code.

Nearly every new generation of technology — even online search engines — has faced similar legal challenges. Often, “there is no statute or case law that covers it,” said Bradley J. Hulbert, an intellectual property lawyer who specializes in this increasingly important area of the law.

The suit is part of a groundswell of concern over artificial intelligence. Artists, writers, composers and other creative types increasingly worry that companies and researchers are using their work to create new technology without their consent and without providing compensation. Companies train a wide variety of systems in this way, including art generators, speech recognition systems like Siri and Alexa, and even driverless cars.

Copilot is based on technology built by OpenAI, an artificial intelligence lab in San Francisco backed by a billion dollars in funding from Microsoft. OpenAI is at the forefront of the increasingly widespread effort to train artificial intelligence technologies using digital data.

After Microsoft and GitHub released Copilot, GitHub’s chief executive, Nat Friedman, tweeted that using existing code to train the system was “fair use” of the material under copyright law, an argument often used by companies and researchers who built these systems. But no court case has yet tested this argument.

“The ambitions of Microsoft and OpenAI go way beyond GitHub and Copilot,” Mr. Butterick said in an interview. “They want to train on any data anywhere, for free, without consent, forever.”

Mr. Butterick and a team of other lawyers are suing Microsoft and other developers of Copilot. Credit... Mike Segar/Reuters

In 2020, OpenAI unveiled a system called GPT-3. Researchers trained the system using enormous amounts of digital text, including thousands of books, Wikipedia articles, chat logs and other data posted to the internet.

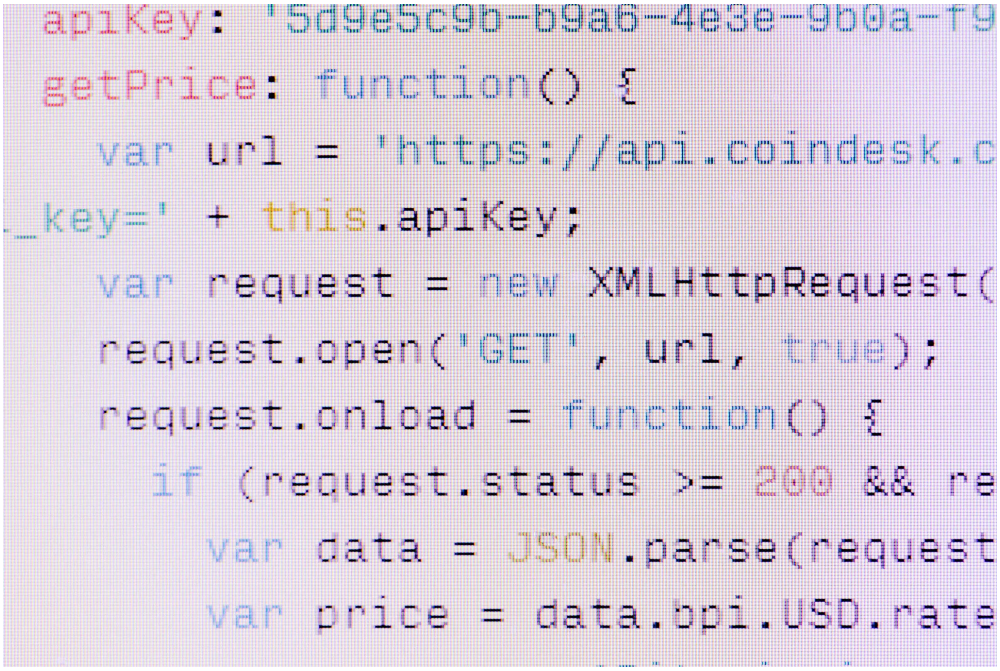
By pinpointing patterns in all that text, this system learned to predict the next word in a sequence. When someone typed a few words into this “large language model,” it could complete the thought with entire paragraphs of text. In this way, the system could write its own Twitter posts, speeches, poems and news articles.

Much to the surprise of the researchers who built the system, it could even write computer programs, having apparently learned from an untold number of programs posted to the internet.

So OpenAI went a step further, training a new system, Codex, on a new collection of data stocked specifically with code. At least some of this code, the lab later said in a research paper detailing the technology, came from GitHub, a popular programming service owned and operated by Microsoft.

This new system became the underlying technology for Copilot, which Microsoft distributed to programmers through GitHub. After being tested with a relatively small number of programmers for about a year, Copilot rolled out to all coders on GitHub in July.

For now, the code that Copilot produces is simple and might be useful to a larger project but must be massaged, augmented and vetted, many programmers who have used the technology said. Some programmers find it useful only if they are learning to code or trying to master a new language.



Codex became the building block for Copilot. Credit... Jason Henry for The New York Times

Still, Mr. Butterick worried that Copilot would end up destroying the global community of programmers who have built the code at the heart of most modern technologies. Days after the system's release, he published a blog post titled: "This Copilot Is Stupid and Wants to Kill Me."

Mr. Butterick identifies as an open source programmer, part of the community of programmers who openly share their code with the world. Over the past 30 years, open source software has helped drive the rise of most of the technologies that consumers use each day, including web browsers, smartphones and mobile apps.

Though open source software is designed to be shared freely among coders and companies, this sharing is governed by licenses designed to ensure that it is used in ways to benefit the wider community of programmers. Mr. Butterick believes that Copilot has violated these licenses and, as it continues to improve, will make open source coders obsolete.

After publicly complaining about the issue for several months, he filed his suit with a handful of other lawyers. The suit is still in the earliest stages and has not yet been granted class-action status by the court.

To the surprise of many legal experts, Mr. Butterick's suit does not accuse Microsoft, GitHub and OpenAI of copyright infringement. His suit takes a different tack, arguing that the companies have violated GitHub's terms of service and privacy policies while also running afoul of a federal law that requires companies to display copyright information when they make use of material.

Mr. Butterick and another lawyer behind the suit, Joe Saveri, said the suit could eventually tackle the copyright issue.

Joe Saveri is one of the lawyers involved in the lawsuit. Credit... Tag Christof for The New York Times

Asked if the company could discuss the suit, a GitHub spokesman declined, before saying in an emailed statement that the company has been “committed to innovating responsibly with Copilot from the start, and will continue to evolve the product to best serve developers across the globe.” Microsoft and OpenAI declined to comment on the lawsuit.

Under existing laws, most experts believe, training an A.I. system on copyrighted material is not necessarily illegal. But doing so could be if

the system ends up creating material that is substantially similar to the data it was trained on.

Some users of Copilot have said it generates code that seems identical — or nearly identical — to existing programs, an observation that could become the central part of Mr. Butterick's case and others.

Pam Samuelson, a professor at the University of California, Berkeley, who specializes in intellectual property and its role in modern technology, said legal thinkers and regulators briefly explored these legal issues in the 1980s, before the technology existed. Now, she said, a legal assessment is needed.

"It is not a toy problem anymore," Dr. Samuelson said.