



Artificial intelligence (AI)

🕒 This article is more than 1 year old

'I didn't give permission': Do AI's backers care about data law breaches?

Regulators around world are cracking down on content being hoovered up by ChatGPT, Stable Diffusion and others

Alex Hern and Dan Milmo

Mon 10 Apr 2023 11.10 CEST

Cutting-edge artificial intelligence systems can help you [escape a parking fine](#), write an [academic essay](#), or fool you into believing [Pope Francis is a fashionista](#). But the virtual libraries behind this breathtaking technology are vast - and there are concerns they are operating in breach of personal data and copyright laws.

The enormous datasets used to train the latest generation of these [AI systems, like those behind ChatGPT and Stable Diffusion](#), are likely to contain billions of images scraped from the internet, millions of pirated ebooks, the entire proceedings of 16 years of the European parliament and the whole of English-language Wikipedia.

But the industry's voracious appetite for big data is starting to cause problems, as regulators and courts around the world crack down on researchers hoovering up content without consent or notice. In response, [AI labs are fighting to keep their datasets secret](#), or even daring regulators to push the issue.

In Italy, ChatGPT has been banned from operating after the country's data protection regulator said there was no legal basis to justify the collection and "massive storage" of personal data in order to train the GPT AI. On Tuesday, the Canadian privacy commissioner followed suit with an investigation into the company in response to a complaint alleging "the collection, use and disclosure of personal information without consent".

Britain's data watchdog expressed its own concerns. "Data protection law still applies when the personal information that you're processing comes from publicly accessible sources," said Stephen Almond, the director of technology and innovation at the Information Commissioner's Office.

Michael Wooldridge, a professor of computer science at the University of Oxford, says "large language models" (LLMs), such as those that underpin OpenAI's ChatGPT and Google's Bard, Hoover up colossal amounts of data.

"This includes the whole of the world wide web - everything. Every link is followed in every page, and every link in those pages is followed ... In that unimaginable amount of data there is probably a lot of data about you and me," he says, adding that comments about a person and their work could also be gathered by an LLM. "And it isn't stored in a big database somewhere - we can't look to see exactly what information it has on me. It is all buried away in enormous, opaque neural networks."

Wooldridge says copyright is a "coming storm" for AI companies. LLMs are likely to have accessed copyrighted material, such as news articles. Indeed the GPT-4-assisted chatbot attached to Microsoft's Bing search engine cites news sites in its answers. "I didn't give explicit permission for my works to be used as training data, but they almost certainly were, and now they contribute to what these models know," he says.

"Many artists are gravely concerned that their livelihoods are at risk from generative AI. Expect to see legal battles," he adds.

Lawsuits have emerged already, with the stock photo company Getty Images suing the British startup Stability AI - the company behind the AI image generator Stable Diffusion - after claiming that the image-generation firm violated copyright by using millions of unlicensed Getty Photos to train its system. In the US a group of artists is suing Midjourney and Stability AI in a lawsuit that claims the companies "violated the rights of millions of artists" in developing their products by using artists' work without their permission.



📷 A sketch drawn by Kris Kashtanova that the artist fed into the AI program Stable Diffusion and transformed into the resulting image using text prompts. Photograph: Kris Kashtanova/Reuters

Awkwardly for Stability, Stable Diffusion will occasionally spit out pictures with a Getty Images watermark intact, examples of which the photography agency included in its lawsuit. In January, [researchers at Google](#) even managed to prompt the Stable Diffusion system to recreate near-perfectly one of the unlicensed images it had been trained on, [a portrait of the US evangelist Anne Graham Lotz](#).

Copyright lawsuits and regulator actions against OpenAI are hampered by the company's absolute secrecy about its training data. In response to the Italian ban, Sam Altman, the chief executive of OpenAI, which developed ChatGPT, said: "We think we are following all privacy laws." But the company has refused to share any information about what data was used to train GPT-4, the latest version of the underlying technology that powers ChatGPT.

Even in its "[technical report](#)" describing the AI, the company curtly says only that it was trained "using both publicly available data (such as internet data) and data licensed from third-party providers". Further information is hidden, it says, due to "both the competitive landscape and the safety implications of large-scale models like GPT-4".

Others take the opposite view. EleutherAI describes itself as a "non-profit AI research lab", and was founded in 2020 with the goal of recreating GPT-3 and releasing it to the public. To that end, the group put together the Pile, an 825-gigabyte collection of datasets gathered from every corner of the internet. It includes 100GB of ebooks taken from the pirate site bibliotik, another 100GB of

computer code scraped from Github, and a collection of 228GB of websites gathered from across the internet since 2008 - all, the group acknowledges, without the consent of the authors involved.

Sign up to TechScape

 Free weekly newsletter

A weekly dive in to how technology is shaping our lives

Enter your email address

Sign up

Privacy Notice: Newsletters may contain info about charities, online ads, and content funded by outside parties. For more information see our [Privacy Policy](#). We use Google reCaptcha to protect our website and the Google [Privacy Policy](#) and [Terms of Service](#) apply.

Eleuther argues that the datasets in the Pile have all been so widely shared already that its compilation “does not constitute significantly increased harm”. But the group does not take the legal risk of directly hosting the data, instead turning to a group of anonymous “data enthusiasts” called the Eye, whose [copyright takedown policy](#) is a video of a choir of clothed women pretending to masturbate their imaginary penises while singing.

Some of the information produced by chatbots has also been false. ChatGPT has falsely accused a US law professor, Jonathan Turley, of George Washington University, of sexually harassing one of his students - citing a news article that didn't exist. **The Italian regulator had also referred to the fact that ChatGPT's responses do not “always match factual circumstances” and “inaccurate personal data are processed”.**

An annual report into progress in AI showed that commercial players were dominating the industry, over academic institutions and governments.

According to the [2023 AI Index report](#), compiled by California-based Stanford University, last year there were 32 significant industry-produced machine-learning models, compared with three produced by academia. Up until 2014, most of the significant models came from the academic sphere, but since then the cost of developing AI models, including staff and computing power, has risen.

“Across the board, large language and multimodal models are becoming larger and pricier,” the report said. An early iteration of the LLM behind ChatGPT, known as GPT-2, had 1.5bn parameters, analogous to the neurons in a human brain, and cost an estimated \$50,000 to train. By comparison, Google's PaLM had 540bn parameters and cost an estimated \$8m.

This has raised concerns that corporate entities will take a less measured approach to risk than academic or government-backed projects. Last week a letter whose signatories included Elon Musk and the Apple co-founder Steve Wozniak [called for](#)

