# OpenAI's hunger for data is coming back to bite it

## The company's AI services may be breaking data protection laws, and there is no resolution in sight.

Melissa Heikkilä

OpenAI has just over a week to comply with European data protection laws following a temporary ban in Italy and a slew of investigations in other EU countries. If it fails, it could face hefty fines, be forced to delete data, or even be banned.

But experts have told MIT Technology Review that it will be next to impossible for OpenAI to comply with the rules. That's because of the way data used to train its AI models has been collected: by hoovering up content off the internet.

In AI development, the dominant paradigm is that the more training data, the better. OpenAI's GPT-2 model had a data set consisting of 40 gigabytes of text. GPT-3, which ChatGPT is based on, was trained on 570 GB of data. OpenAI has not shared how big the data set for its latest model, GPT-4, is.

But that hunger for larger models is now coming back to bite the company. In the past few weeks, several Western data protection authorities have started investigations into how OpenAI collects and processes the data powering ChatGPT. They believe it has scraped people's personal data, such as names or email addresses, and used it without their consent.

The Italian authority has blocked the use of ChatGPT as a precautionary measure, and French, German, Irish, and Canadian data regulators are also investigating how the OpenAI system collects and uses data. The European Data Protection Board, the umbrella organization for data protection authorities, is also setting up an EU-wide task force to coordinate investigations and enforcement around ChatGPT.

Italy has given OpenAI until April 30 to comply with the law. This would mean OpenAI would have to ask people for consent to have their data scraped, or prove that it has a "legitimate interest" in collecting it. OpenAI will also have to explain to people how ChatGPT uses their data and give them the power to correct any mistakes about them that the chatbot spits out, to have their data erased if they want, and to object to letting the computer program use it.

If OpenAI cannot convince the authorities its data use practices are legal, it could be banned in specific countries or even the entire European Union. It could also face hefty fines and might even be forced to delete models and the data used to train them, says Alexis Leautier, an AI expert at the French data protection agency CNIL.

OpenAI's violations are so flagrant that it's likely that this case will end up in the Court of Justice of the European Union, the EU's highest court, says Lilian Edwards, an internet law professor at Newcastle University. It could take years before we see an answer to the questions posed by the Italian data regulator.

## High-stakes game

The stakes could not be higher for OpenAI. The EU's General Data Protection Regulation is the world's strictest data protection regime, and it has been copied widely around the world. Regulators everywhere from Brazil to

California will be paying close attention to what happens next, and the outcome could fundamentally change the way AI companies go about collecting data.

In addition to being more transparent about its data practices, OpenAI will have to show it is using one of two possible legal ways to collect training data for its algorithms: consent or "legitimate interest."

It seems unlikely that OpenAI will be able to argue that it gained people's consent when it scraped their data. That leaves it with the argument that it had a "legitimate interest" in doing so. This will likely require the company to make a convincing case to regulators about how essential ChatGPT really is to justify data collection without consent, says Edwards.

OpenAI told us it believes it complies with privacy laws, and in a blog post it said it works to remove personal information from the training data upon request "where feasible."

The company says that its models are trained on publicly available content, licensed content, and content generated by human reviewers. But for the GDPR, that's too low a bar.

"The US has a doctrine that when stuff is in public, it's no longer private, which is not at all how European law works," says Edwards. The GDPR gives people rights as "data subjects," such as the right to be informed about how their data is collected and used and to have their data removed from systems, even if it was public in the first place.

## Finding a needle in a haystack

OpenAI has another problem. The Italian authority says OpenAI is not being transparent about how it collects users' data during the post-training phase, such as in chat logs of their interactions with ChatGPT.

“What’s really concerning is how it uses data that you give it in the chat,” says Leautier. People tend to share intimate, private information with the chatbot, telling it about things like their mental state, their health, or their personal opinions. Leautier says it is problematic if there’s a risk that ChatGPT regurgitates this sensitive data to others. And under European law, users need to be able to get their chat log data deleted, he adds.

OpenAI is going to find it near-impossible to identify individuals’ data and remove it from its models, says Margaret Mitchell, an AI researcher and chief ethics scientist at startup Hugging Face, who was formerly Google’s AI ethics co-lead.

The company could have saved itself a giant headache by building in robust data record-keeping from the start, she says. Instead, it is common in the AI industry to build data sets for AI models by scraping the web indiscriminately and then outsourcing the work of removing duplicates or irrelevant data points, filtering unwanted things, and fixing typos. These methods, and the sheer size of the data set, mean tech companies tend to have a very limited understanding of what has gone into training their models.

Tech companies don’t document how they collect or annotate AI training data and don’t even tend to know what’s in the data set, says Nithya Sambasivan, a former research scientist at Google and an entrepreneur who has studied AI’s data practices.

Finding Italian data in ChatGPT’s vast, unwieldy training data set will be like finding a needle in a haystack. And even if OpenAI managed to delete users’ data, it’s unclear if that step would be permanent. Studies have shown that data sets linger on the internet long after they have been deleted, because copies of the original tend to remain online.

"The state of the art around data collection is very, very immature," says Mitchell. That's because tons of work has gone into developing cutting-edge techniques for AI models, while data collection methods have barely changed in the past decade.

In the AI community, work on AI models is overemphasized at the expense of everything else, says Mitchell: "Culturally, there's this issue in machine learning where working on data is seen as silly work and working on models is seen as real work."

Sambasivan agrees: "As a whole, data work needs significantly more legitimacy."