

Google confirms it's training Bard on scraped web data, too

On Monday, *Gizmodo* spotted that the search giant updated its privacy policy to disclose that its various AI services, such as Bard and Cloud AI, may be trained on public data that the company has scraped from the web.

"Our privacy policy has long been transparent that Google uses publicly available information from the open web to train language models for services like Google Translate," said Google spokesperson Christa Muldoon to *The Verge*. "This latest update simply clarifies that newer services like Bard are also included. We incorporate privacy principles and safeguards into the development of our AI technologies, in line with our AI Principles."

[a: hover\]: text-gray-63](#) [\[&a: hover\]: shadow-underline-black dark: \[a: hover\]: text-gray-bd dark: \[a: hover\]: shadow-underline-gray \[a: shadow-underline-gray-63 dark: \[a: text-gray-bd dark: \[a: shadow-underline-gray\]](#)>Image: Google

Following the update on July 1st, 2023, Google's privacy policy now says that "Google uses information to improve our services and to develop new products, features, and technologies that benefit our users and the public" and that the company may "use publicly available information to help train Google's AI models and build products and features like Google Translate, Bard, and Cloud AI capabilities."

You can see from the policy's revision history that the update provides some additional clarity as to the services that will be trained using the collected data. For example, the document now says that the information may be used for "AI Models" rather than "language models," granting Google more freedom to

train and build systems beside LLMs on your public data. And even that note is buried under an embedded link for “publically accessible sources” underneath the policy’s “Your Local Information” tab that you have to click to open the relevant section.

The updated policy specifies that “publicly available information” is used to train Google’s AI products but doesn’t say how (or if) the company will prevent copyrighted materials from being included in that data pool. Many publicly accessible websites have policies in place that ban data collection or web scraping for the purpose of training large language models and other AI toolsets. It’ll be interesting to see how this approach plays out with various global regulations like GDPR that protect people against their data being misused without their express permission, too.

A combination of these laws and increased market competition have made makers of popular generative AI systems like OpenAI’s GPT-4 extremely cagey about where they got the data used to train them and whether or not it includes social media posts or copyrighted works by human artists and authors.

The matter of whether or not the fair use doctrine extends to this kind of application currently sits in a legal gray area. The uncertainty has sparked various lawsuits and pushed lawmakers in some nations to introduce stricter laws that are better equipped to regulate how AI companies collect and use their training data. It also raises questions regarding how this data is being processed to ensure it doesn’t contribute to dangerous failures within AI systems, with the people tasked with sorting through these vast pools of training data often subjected to long hours and extreme working conditions.

Gannett, the largest newspaper publisher in the United States, is suing Google and its parent company, Alphabet, claiming that advancements in AI technology have helped the search giant to hold a monopoly over the digital ad market. Products

like Google's AI search beta have also been dubbed "plagiarism engines" and criticized for starving websites of traffic.

Meanwhile, Twitter and Reddit – two social platforms that contain vast amounts of public information – have recently taken drastic measures to try and prevent other companies from freely harvesting their data. The API changes and limitations placed on the platforms have been met with backlash by their respective communities, as anti-scraping changes have negatively affected the core Twitter and Reddit user experiences.

<https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping>