Proskauer

As Generative AI Training Lawsuits Mount, Some Providers Offer New Assurances

New Media and Technology Law Blog on **July 20, 2023**

In April, we wrote about how OpenAI had eased the procedure by which ChatGPT users can opt out of their inputs being used for model training purposes (click here for that post). While neither web scraping nor the collection of user data to improve services are new concepts, AI did not attract much attention for these practices in prior years. However, with the runaway success of generative AI ("GAI") tools like ChatGPT, customers ranging from individual consumers to large businesses are starting to take notice of GAI developers' data supply.[1]

In 2023, we've seen several lawsuits against OpenAI and other GAI providers alleging unauthorized and improper use of plaintiffs' proprietary data as GAI model training material, with claims variously based on copyright, contract, and privacy law. And lawsuits aren't the only way that GAI providers have lately faced increased scrutiny over how and where they obtain training data to develop their GAI products. For example, Reddit recently announced a plan to begin charging for access to its API, which is generally how GAI providers import its data into their models (*i.e.*, Reddit has decided that user posts shouldn't be given away for free to GAI providers whose products might undermine the popularity of its platform). On top of these new hurdles, the FTC is reportedly looking into OpenAI's collection of user data (among other issues, such as publication of false information and potentially anti-competitive practices surrounding GAI).

In light of these recent events, it is perhaps not surprising that some GAI providers have revised certain provisions in the terms and conditions for their tools, in an apparent attempt to reassure their users about how user data may – or more precisely, may not be – used. For example, Microsoft has updated its default commercial terms for its Azure OpenAI service (which provides licensed access to OpenAI's GPT models) to explicitly state that user inputs are not used for training, and GitHub has done the same for its GAI coding tool, Copilot. OpenAI has made a similar update to its template Enterprise Agreement. Even Anthropic (provider of ChatGPT competitor Claude), the newest player on the scene whose terms assert a broad right to use user data to develop new products and services, explicitly excludes model training. Other providers may follow suit.

Although a pattern is emerging on this topic – and certainly, GAI providers' default positions have also congealed in other areas, such as accuracy and bias disclaimers – there are plenty of other areas where default terms can vary significantly (e.g., terms surrounding ownership of outputs and IP infringement). We will see if some of these currently varied default terms begin to homogenize over time, as GAI providers compete to offer not only the best tools, but also the best terms, to business customers. [2]

---

[1] Note that generally, GAI models are "trained" first, and then made available to users in a relatively static form (at least until the next model is released). However, models can also be "fine-tuned" with additional data after receiving initial training. Thus, data rightsholders have two overlapping but distinct areas of concern. We use "training" to refer to both initial training and any subsequent fine-tuning.

[2] Note that even seemingly congealed terms of use may change over time, especially in an evolving space like GAI, and users should never assume they know what position a given set of terms takes on an issue – or assume terms in place today are the same tomorrow – without confirming.

View original.

Related Professionals

- **Peter J. Cramer**
  Associate