Proskauer

An Update on the Latest GenAI Class Action

July 6, 2023

This webinar explored the possible implications of two new putative class action litigations brought against OpenAI in connection with its generative AI ("GenAI") offerings. These two lawsuits are the latest cases filed against OpenAI (note: other prior litigations were discussed in the first webinar in this series, "An Overview of Key IP Issues in AI"). And such cases likely will not be the last [indeed, days after this webinar, on July 7, 2023, a second copyright lawsuit was filed against OpenAI on behalf of book authors].

As explained by the presenters, the two recently-filed cases brought against OpenAI are different in nature: the first (the "Tremblay" case) is a copyright case filed on behalf of a class of authors whose works were allegedly infringed upon when OpenAI used them to train its GenAI models; and the second (the "P.M." case) is a sprawling 157-page complaint against OpenAI that focuses on privacy issues related to how OpenAI collects and uses data to train its GenAI models as well as to OpenAI's relationships with users.

The Tremblay copyright case

The presenters noted that the Tremblay copyright case is similar to other ongoing copyright cases against OpenAI (specifically, the infringement cases against StabilityAI regarding its GenAI image generator as well as cases involving GitHub's software code generator), as all of those cases allege that OpenAI committed infringement when it used copyrighted content to train its models. In essence, the plaintiffs in the *Tremblay* case allege that ChatGPT was trained, without permission, on databases that contained a large volume of copyrighted books that included the plaintiffs' works and other examples of "high-quality longform writing" that is desirable for AI training purposes.

One such database referenced in the complaint is, for example, the BookCorpus database that plaintiffs described as a collection of thousands of unique unpublished books from a variety of genres. As pointed out by the presenters, BookCorpus was compiled by researchers in 2015 and the complaint suggests it was used to train an early iteration of ChatGPT back in 2018, potentially raising some statute of limitations defenses in this case and affected the decision not to include the makers of the BookCorpus database as defendants in this suit. Besides BookCorpus, the plaintiffs' complaint references two other books databases (colloquially named Books1 and Books2) that were allegedly used to train ChatGPT-3. The presenters noted that it is not clear what materials populate the Books1 and Books2 databases, with the plaintiffs speculating that Books1 might consist of material from Project Gutenberg (a literature database of public domain works) and Books2 perhaps consisting of material from so-called "shadow library" websites that offer allegedly infringing copies of academic content that are typically offered only behind paywalls. It is likely that a future motion to dismiss by OpenAI may bring some clarity to these referenced databases (or others) purportedly used to train ChatGPT.

In *Tremblay*, the plaintiffs allege that evidence of infringement can be found simply in ChatGPT's ability to summarize copyrighted books with relative accuracy, thus suggesting that ChatGPT ingested and copied these works as part of its training. The presenters noted that the plaintiff's' argument here is an interesting one that will likely be one of the major factual issues during discovery (if the case proceeds past the initial pleading stages). Indeed, the presenters stated that one of central issues in the case will be whether plaintiffs' works were truly part of ChatGPT's training data; the speakers also stated that plaintiffs' allegations that ChatGPT copied and trained on their works based on the platform's ability to produce an accurate summary may or may not be true, as there are summaries of plaintiffs' works on the web that may have been used by ChatGPT. Thus, the question becomes: did ChatGPT copy and ingest the plaintiffs' works or merely scrape the summaries from the web? The presenters also stated that another important issue will revolve around ChatGPT's training process: is it more like reading and analyzing (akin to a human researcher) or does AI training involve copying and storing the works in some way that implicates copyright law? The presenters pointed out that, notably, the plaintiffs did not allege that the summaries of their works produced by ChatGPT were infringing, rather that infringement occurred during the alleged input and training of ChatGPT and its use of the aforementioned book databases.

In the presenters' minds, the *Tremblay* complaint will likely face a motion to dismiss based on various procedural defenses (including statute of limitations and standing issues), and if the case proceeds to discovery in some form, the parties will likely be sparring over the issue of fair use (with the Second Circuit's *Google Books* case and the recent Supreme Court *Warhol* case being central to both parties' arguments).

The presenters also raised the question of the possible implications of the suit for users of OpenAI's GenAI products (or similar GenAI products) should the *Tremblay* case ultimately succeed on any of the claims. They noted that the GenAI IP-related litigations thus far have focused on the alleged infringing nature of the GenAI training process, so, at this point, there should not be any sizeable risk of direct liability for downstream users of GenAI products. However, the presenters noted that there are unanswered questions about secondary liability and opined that a plaintiff could conceivably come up with a plausible theory, but would have to overcome significant hurdles (e.g., to bring a successful contributory liability claim, a plaintiff would have to show a user had knowledge of direct infringement and materially contributed to or induced the infringement, a high bar particularly in the *Tremblay* case where the GenAI model training on the book databases allegedly occurred years before; similarly, from a vicarious liability perspective, it would be difficult to show that a user had the ability to supervise the AI training process, among other things).

The presenters cautioned that, going forward, as businesses use GenAI products everyday and train their own custom models, such secondary liability theories may be easier to plead depending on the circumstances of a particular case. Thus, the presenters suggested that users of GenAI products should pay close attention to the contractual terms governing GenAI products and where such contracts allocate liability and risk, including related to training and outputs.

The P.M. Privacy Case

Unlike the *Tremblay* case which focused on copyright, the *P.M.* case focuses mainly on privacy issues and outlines some potential doomsday societal issues should AI go wrong. The complaint essentially alleges that the scraping of social media and web-based resources has resulted in OpenAI collecting personal information and sensitive data of individuals and children without consent, thereby violating those persons' privacy and property rights. The presenters noted that the complaint contains 15 privacy claims, including violations of federal electronic privacy laws, the federal Computer Fraud and Abuse Act (CFAA), California privacy laws, Illinois biometric privacy law, New York consumer protection laws, and various other state and common law claims. As part of the complaint's request for relief, the plaintiffs seek many safeguards to prevent privacy violations and societal ills caused by AI. For example, the complaint seeks an injunction barring commercial access to all OpenAI GenAI products and services until certain safeguards are put into place. These measures include the establishment of an AI Council that would approve GenAI products, accountability and transparency protocols, protections to ensure AI products do not surpass human intelligence (aka the "singularity"), and the deletion of personal information from OpenAI's databases and the offering of a data compensation system for individuals whose data was acquired through scraping, among other relief.

The presenters stated that the complaint focuses on web scraping, with the plaintiffs calling it "stolen" data. In their view, this focus on scraping is important beyond this case as many businesses rely on scraping or products that use scraped data in their daily business activities in ways that have nothing to do with AI or with the personal data of individuals. As noted by the presenters, if scraping is relevant to your business, then such entities should watch this case closely as any adverse ruling could result in restrictive practices that affect all types of web scraping.

Commenting on the potential implications of the *P.M.* case, the presenters noted that to the extent any of the privacy claims are sustained, and to the extent that some form of injunctive relief is granted, it's not clear how this will affect these AI products. For example, the speakers pointed out that it may not be technically possible to "un-train" OpenAI's models. As with the *Tremblay* case, for users, contractual issues will be important and enterprises negotiating such agreements should be watching the outcome of these ongoing litigations when considering the allocation of risk and indemnity. Moreover, the presenters noted that enterprises should think ahead and ensure agreements with a provider cover the possible outcome that a court order or settlement compels OpenAI to restrict the functionality of its products.

Additional questions noted by the presenters include: Will such models still be "useful" if there is an adverse litigation result against OpenAI? Can a business pivot away from OpenAI products if the need arises, or will the technology be too ingrained in the business flow, and will any termination remedy be meaningful? Certainly, the presenters stated that at this point there is a bit of uncertainty around these contracting issues and others.

Related Professionals

- **Jeffrey D. Neuburger**
- **David A. Munkittrick**
  Partner