Kluwer Copyright Blog

Generative AI: the US class action against Google Bard (and other AI tools) for web scraping

Gianluca Campus (University of Milan) · Tuesday, October 3rd, 2023

Image by Alexandra_Koch from Pixabay

The US class action against Google Bard (*J.L. v. Alphabet Inc, U.S. District Court for the Northern District of California, No. 3:23-cv-03440*)

In a recent post we analysed a class action filed in the US against Open AI for unauthorized use of copyright works for training of generative AI tools such as ChatGPT (here) ("Generative AI" or "Gen AI"). We have also noted that this was not the only class action filed in the US against Open AI, since a parallel class action was based on alleged data breach (here). Another class action was recently filed against Google (notably by the same law firm which promoted the class actions against Open AI) in the United States District Court – Northern District of California for alleged web scraping (this means covering both copyright and privacy aspects) in the training of its AI tools, Bard, Imagen, MusicLM, Duet AI, and Gemini (here).

Such class actions can be considered part of the evolution of the regulatory landscape dedicated to Generative AI. In fact, in those legal systems (such as the US) where the regulatory approach has been based so far on soft laws and definition of principles (here but see also here for a recent legislative proposal to regulate AI in the US), rather than on strict rules and prescriptions (such as in the EU, which intends to introduce an AI Act based on the EU model for product safety legislation), the outcome of such class actions will be very relevant to address some of the main points of interest relating to the introduction of a disruptive technology such as Generative AI. With regards to the legal side of the Generative AI business model, such class actions will be useful to clarify: (1) if and how the training of the LLM (Large Language Model) can be based on resources available on the Internet and whether any *fair use* doctrine can be invoked for such training; and (2) whether and to what extent a substantial shift of the risks of copyright infringements deriving from input and output onto the users themselves is admissible, a practice that Gen AI producers are adopting via specific clauses in their Terms & Conditions (see Clause 3 "Your Content" in the OpenAI Terms of Use dated March 14, 2023).

The plaintiffs' factual allegations

This class action was filed on 11 July 2023, in the United States District Court Northern District of California by eight claimants identified only by their initials for alleged security and privacy reasons – among them a New York Times best-selling author and investigative journalist, an actor and a professor, with the others mere users of the Google services at stake, on behalf of themselves and other parties in the class action complaint (collectively, the "plaintiffs"), against Google DeepMind, Google LLC, and Alphabet Inc (collectively, the "defendants"). The plaintiffs demand a jury trial to recover equitable relief and various types of damages (including actual, statutory, punitive and exemplary damages) as a result and consequence of the defendants' unlawful conduct.

In the plaintiffs' reasoning, the development of the AI by Google began in 2017, when it introduced the "Transformer" neural network, a revolutionary framework underpinning the LLM. The LLM is "*the very underlying technology that fuels AI chatbots across the AI industry*" (§ I.62). All the products of Google are built with this technology and allegedly using private, personal, and/or copyrighted materials (collectively, the "Products"). The most important Google Products are: (i) Bard, which is a chatbot based on Generative AI and machine learning developed by Google, as a direct competitor to ChatGPT created by OpenAI, and released in beta in the US in March 2023; (ii) Imagen, which is a text-to-image generative AI; (iii) MusicLM, which is a generative AI with text-to-music capabilities; (iv) Duet AI, a generative AI tool which assists users with drafting in Docs and Gmail, image generation in Slides, automatic meeting summaries in Meet, and more; and (v) Gemini, still in development, which is being billed as a highly efficient, multimodal machine-learning model that can decode many data types at once, similar to how humans use different senses in the real world.

For developing its Products, Google's AI model was pre-trained on an estimated 1.56 trillion words of "*public dialog data and web text*," drawn from Infiniset, an amalgamation of various internet content meticulously selected to improve the model's conversational abilities (§ I.76) (see here and here). In addition, the origin of the data used to train LaMDA (here), the language model behind Google Bard, includes the C4 dataset. The C4 dataset, created by Google in 2020, is taken from the Common Crawl dataset, which is an open-source dataset and "*a massive collection of web pages and websites consisting of petabytes of data collected over twelve (12) years, including raw web page data, metadata extracts, and text extracts*" (§ I.78 and here). The Common Crawl dataset is owned by a non-profit, which makes the data available to the public for free — but it is intended to be used for research and education and, according to the plaintiffs, it was never intended to be turned into an AI product for commercial use (see here and here).

Copyright infringements and other legal allegations

According to the plaintiffs' allegations, the defendants' web scraping for LLM training purposes violated their copyright and, moreover, would imply an unauthorized and widespread misappropriation of copyrighted works extending across a wide spectrum of industries that depend on creative content creation. The Products' ability to replicate the writing styles of specific authors, recreate the music and lyrics of specific musicians, and duplicate the works of online content producers, as well as the ability to summarize and reproduce copyrighted materials, arises from the fact that these materials were copied by the defendants without authorization and injected into the underlying LLM as part of its training data (§ I.B.107).

Such conduct would be even more dangerous for the cultural industries, since, despite the existence of numerous lawful ways to acquire training data, the defendants opted instead to pillage the internet for copyrighted works and the resulting impact has not only infringed upon the rights of creators but has created an environment that ultimately could discourage creativity and innovation. It could also undercut the commercial market for books and works already created; this is because, on demand, the Products are able not only to summarize books in detail, chapter by chapter, but also to regenerate the text of books (§ I.B.110-111).

Further, according to the plaintiffs' allegations, the practice of web scraping cannot be considered to fall within the concept of "*fair use*", a critical aspect of copyright law designed to allow limited use of copyrighted material without permission for purposes like commentary, criticism, news reporting, and scholarly reports (see *McGucken vs Pub Ocean Limited*, 42 F.4th 1149 (9th Cir. 2022)). The defendants' wholesale collection and use of copyrighted material, with no option for copyright owners to opt out, would exceed the legal interpretation of "fair use" (see *VHT vs Zillow Group*, 918 F.3d 723, 743 (9th Cir. 2019); *Worldwide Church of God vs Phila. Church of God, Inc.*, 227 F.3d 110, 1118 (9th Cir. 2000) ("*copying an entire work militates against a finding of fair use.*").

In addition to the alleged copyright violations, in the plaintiffs' reasoning, defendants' web scraping violated and continues to violate the plaintiffs' property interests ("*Courts recognize that internet users have a property interest in their personal information and data* (…) *which includes the right to possess, use, profit from, sell, and exclude others from accessing or exploiting that information without consent or remuneration*" (see § I.B.161 recalling the precedent *Calhoun v. Google*, which recognized property interest in personal information). The defendants failed to register as data brokers under applicable laws of California (here). By failing to do so prior to scraping the internet, the defendants did not allow all the class members a right to delete their personal information collected by the defendants, and a right to opt out of the use of that information, which was used to build the Products. The plaintiffs argue that such conduct would replicate that of Clearview. Clearview created AI products using facial recognition technology. To create its product, Clearview scraped billions of publicly available photos from websites and social media platforms. Clearview's illegal scraping practices were subject to administrative fines and regulatory proceedings in the US and in the UK (see here and here).

Measures requested against the Products

The plaintiffs requested as injunctive relief against the use of the Products the following measures (§ 205):

a. Establishment of an independent body of thought leaders (the "AI Council") who shall be responsible for approving uses of the Products before, not after, the Products are deployed for said uses.
b. Implementation of Accountability Protocols that hold the defendants responsible for Products' actions and outputs.
c. Implementation of effective cybersecurity safeguards for the Products as determined by the AI Council.
d. Implementation of Appropriate Transparency Protocols requiring the defendants to clearly and precisely disclose the data they are collecting.
e. The defendants to be required to allow Product users and everyday internet users to opt out of all data collection.
f. The defendants to be required to add technological safety measures to the Products.
g. The defendants to be required to implement, maintain, regularly review and revise as necessary a threat management program designed to appropriately monitor the defendants' information networks for threats.
h. Establishment of a monetary fund (the "AI Monetary Fund" or "AIMF") to compensate class members for the defendants' past and ongoing misconduct.
i. Appointment of a third-party administrator (the "AIMF Administrator") to administer the AIMF to members of the class in the form of "data dividends" as fair and just compensation for the stolen data on which the Products depend.

Conclusion

As compared to the class actions against Open AI, this class action seems to be directed even more precisely to the core issue of the Gen AI tools – their alleged training via resources made public on the internet and/or protected under copyright laws – combining potential legal issues on both the IP and the privacy fronts (not to mention due to bias in the algorithms). Whatever the result of such class actions, this seems to be a timely occasion for parties to clarify and judges to assess legitimacy of Gen AI tools based on a deep analysis of the technical functioning and composition of the training datasets. The fact that this is the main goal of the class action seems to be supported by the proposals introduced by the plaintiffs for a governance scheme for all Gen AI models.

To make sure you do not miss out on regular updates from the Kluwer Copyright Blog, please subscribe here.

Kluwer IP Law

This entry was posted on Tuesday, October 3rd, 2023 at 9:13 am and is filed under Artificial Intelligence (AI), Case Law, Infringement, Legislative process, USA
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.