eTek Law PC
eteklaw.com

Content Licensing for AI Training Datasets and Implications for Commercial Vendors and Open Source Projects

OpenAI and the Associated Press (AP) recently signed a license agreement that allows OpenAI to train ChatGPT on AP content. This is an important development in the AI industry considering pervasive concerns that the data that was used to train most, if not all generative AI engines so far, has been improperly collected (whether by violating website T&Cs through data scraping, by breaching copyright laws or confidentiality agreements, or through non-compliant handling of personal data).

The license arrangement between AP and OpenAI raises a number of issues:

1. This license arrangement shows a path forward for cleaning up AI training datasets from a legal standpoint. But despite the wide range of AP content, this is just a drop in the bucket considering the full scope of the content that was used to train the major generative AI engines. Given the ongoing need for generative AI engines to ingest more and more training data, the question arises whether any particular AI vendor can realistically enter into sufficient content licensing transactions with enough content owners to clean up all or most of their data. Both the cost and the logistical resources needed to achieve that will be daunting.

2. There is a need for an industry-wide solution that would act as a clearinghouse for content from a wide range of content owners for AI licensing purposes. This is a business idea for your next entrepreneurial project! :)

3. The music industry has shown a precedent, although not perfect, for licensing content between a wide range of copyright holders and commercial customers (see for example ASCAP and BMI, each of which collects from users and distributes to content owners over $1B in license fees per year). But the scope of content on the Internet is vastly larger than the global music catalog, so such clearinghouse licensing solutions would have to be scaled up by orders of magnitude for generative AI training.

4. Licensing large volumes of content could create a competitive gap between AI vendors that can afford to do it, and vendors that can't. Also, increasing training costs for generative AI vendors would have to either be passed onto the customer base, or would erode the profit margins of the AI vendors.

5. Widespread content licensing of training datasets for generative AI engines will pose a challenge for Open Source AI engines. Given the decentralized nature of Open Source software and the free distribution models, will content owners seek to obtain injunctions against the Open Source AI engines, pursue the Open Source developers through the legal system, or assert copyright infringement claims against end users? Or will content owners spare Open Source AI engines from license fees given the inherent social and economic benefits of Open Source technology? Or will we have similar industry-wide defensive and offensive conversations as we did years ago when various large companies threatened Open Source software with patent infringement lawsuits? And would content owners be willing to absorb the stigma and backlash from the Open Source industry if they attempt to collect license fees from Open Source generative AI models? The Open Source community has to start thinking seriously about such issues now.

6. AI unlearning is another powerful tool for generative AI engines to deflect and/or mitigate legal threats around training datasets. I recently wrote about AI unlearning here: https://www.linkedin.com/pulse/ai-unlearning-powerful-legal-business-tool-industry-marius-domokos. AI unlearning will likely become an effective tool for AI engines to address material threats against their AI training datasets.

The complexity of the AI ecosystem is continuing to increase fast, both in terms of business use cases and legal issues.

What do you think about content licensing transactions from the perspective of commercial AI vendors and Open Source AI engines?