DAVE SHUMAKER

Report From the Field

Who Owes Whom? CCC Webinar Addresses 'The Heart of the Matter'

Do the creators and users of large language models (LLMs) and associated generative AI applications infringe the copyrights of content creators? If so, what specific acts constitute infringement, and what should be done to remedy the problem? CCC's Feb. 29 town hall, The Heart of the Matter: Copyright, AI Training and LLMs, presented a comprehensive case that AI systems do infringe copyrights and called for a licensing system to compensate content owners.

The presentation wove together technical principles and legal arguments. Technical principles were presented by Babis Marmanis, CCC's EVP and CTO, while Noam Shemtov, professor at Queen Mary University of London's School of Law, and Daniel Gervais, Vanderbilt Law School professor, focused on the relevant legal issues. Catherine Zaller Rowland, CCC's VP and general counsel, moderated.

TECHNICAL PRINCIPLES

Marmanis opened the session with an explanation of the process of developing generative AI systems. These systems, such as the GPT-4 LLM that forms the basis of the popular ChatGPT service, must be trained by "ingesting" large amounts of content. This content is scraped from a variety of web-based sources, which may be accessed with or without appropriate permissions. In any case, the content thus ingested is retained indefinitely. Here, Marmanis drew an important distinction between indexes maintained by Google and other search engines and the storage of content by LLMs. The former, he explained, store content in such a way that they cannot reproduce the original content as it originally existed, while LLMs, because they store the content differently, are capable of exactly reproducing the original. Thus, it can be said that they have made complete and exact copies of everything they have ingested.

LEGAL AND COPYRIGHT ISSUES

With that technical background, the discussion moved to legal and copyright issues. Gervais noted that copyright law and the rights of copyright holders have adapted to new technologies as they have been developed and commercialized over the centuries. He illustrated the point with examples of copyright's evolution from its 18th-century origin in restricting the production and sale of books through the addition of rights to perform and display works and its adaptation to technologies from the

LINKS TO THE SOURCES

The Heart of the Matter: Copyright, AI Training and LLMs
linkedin.com/events/theheartofthematter-copyright-a7157813380118933504/comments

Reuters: "OpenAI Gets Partial Win in Authors' Copyright Lawsuit"
reuters.com/legal/litigation/openai-gets-partial-win-authors-us-copyright-lawsuit-2024-02-13

Reuters: "Exclusive: Reddit in AI Content Licensing Deal With Google"
reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22


Babis Marmanis


Noam Shemtov


Daniel Gervais


Catherine Zaller Rowland

Photos courtesy of CCC

player piano to radio, cable television, and computer software. This evolution continued in the web era with the enactment of two international treaties that defined a new right: the "right of making available."

In all cases, Gervais continued, copyright has balanced the rights of authors to determine the uses of—and be compensated for—their works with the ability of consumers to use works on fair and reasonable terms. Now, he concluded, LLMs have created the need for a new evolution and adaptation of copyright. LLMs can produce commercially competitive content that may displace human authors in some cases and affect their work in other cases, and they can do this because they have ingested the works of human authors.

Noting that both the providers and users of LLMs may be liable for copyright infringements, depending on how the AI systems are used and what is done with their output, the presentation turned to questions of specific rights of both copyright holders and content users. Fair use loomed large among those rights. This principle sets the conditions under which the copying of works does or does not constitute infringement, and it remains an exclusively American legal concept—although other countries have considered it, and some have adopted similar rules.

While the law requires a four-factor test to determine whether a use qualifies as fair, in practice, lawsuits have leaned heavily on the determination of whether the use is "transformative"—i.e., creates a substantially new and different work. The effect of the new work on the market for the original has also been an important factor. These considerations require a case-by-case analysis of the relationship of the copy to the original, and no one-size-fits-all rule exists to simplify the judgment. Gervais concluded by noting that some outputs of AI applications may qualify as fair use, depending on the nature of the output and whether the user is a commercial or nonprofit entity.

The discussion then expanded to the international level. Shemtov contrasted the situation in the European Union and the U.K. with that in the U.S. In the U.S., it has fallen to the judiciary to set policy through verdicts in individual lawsuits, while the Europeans have enacted legislation to govern text and data mining (which is assumed to be equivalent to LLM training). Broadly, the legislation distinguishes between nonprofit and for-profit institutions and uses, giving broader latitude to the former to make use of copyrighted material without compensation.

The legal discussion concluded by making the case for the licensing of copyrighted materials for use in LLMs and AI applications. Essentially, the panelists held that while some products of AI applications may be allowable as exceptions to owners' rights, whether under fair use or other doctrines, many are not. Gervais noted that it will take years to resolve lawsuits already underway in the U.S. courts and that decisions favoring copyright holders could result in massive financial liabilities for AI providers. Thus, he argued, it would make sense to resolve the disputes, eliminate the uncertainties, and reduce the financial stakes by entering into reasonable licensing agreements. Shemtov added that calls for licensing should not be construed as anti-technology or anti-business. On the contrary, he argued, they will facilitate technological progress and business development by establishing clear and predictable rules for the use of copyrighted material.

CONCLUSION

These days, AI seems to be part of every conversation. It affects everything else, and everything else has to take it into account. Copyright is but one of these conversations. Legal actions involving fundamental principles and potentially large dollar amounts are underway—and some have already been decided. As of this writing in March 2024, Reuters reports that a federal judge in California has dismissed a case by copyright holders that the output generated by ChatGPT constitutes infringement. However, Reuters also reports that Reddit and Google have entered into a licensing deal that allows Google to train its LLM on Reddit content for a fee of $60 million per year. CCC has presented one side of the debate, but the eventual outcome will be subject to disputes, litigation, and negotiation for some time to come.

Dave Shumaker is a retired clinical associate professor at The Catholic University of America in Washington, D.C., and a former corporate information manager. He is also the author of *The Embedded Librarian: Innovative Strategies for Taking Knowledge Where It's Needed* (Information Today, Inc., 2012), and he founded SLA's Embedded Librarians Caucus in 2015. Send your comments about this article to itletters@infotoday.com.