21st Century Cyber News: Creatives V. Machine Learning Corporations

LIS461: Data and Algorithms: Ethics and Policy

Shia Aaron Lloyd Fisher

University of Wisconsin, Madison

10 July 2023

Shia Aaron Lloyd Fisher

Paul J. Kelly, Tallal Ahmad, Yinka Ajiobola

Data and Algorithms: Ethics and Policy – SADIE 1

10 July 2023

Creatives V. ML Corps.

There is a disagreement between creatives and the well-known Machine Learning (ML) corporation, OpenAI, over the "fair use," interpretation under the law with respect to published material that is allegedly "ingested," by OpenAI's flagship software, ChatGPT. ChatGPT is an artificial intelligent chat bot that uses ML algorithms able to take user input ("data") to formulate an output based on both the interpolation and the interpretation of the text. The complexity of these AI algorithms is such that they also rely heavily on the data set used for the initial training of a particular build. That is, the software is first fed a lot of data from various sources to constitute a particular build.

Authors Mona Awad and Paul Tremblay allege OpenAI has infringed on their copyrighted material and have filed a lawsuit in the Federal Circuit. The seventeen-page-long complaint was filed in San Francisco federal court in July of 2023. It alleges that "a large language model has copied and ingested the text in its training dataset," which, if true, gives their complaint merit if we assume that explicit, informed consent of content creators ("creatives") is a necessary requirement for ML corporations to use copyrighted material as training data (Trembly

v. OpenAI, 94102). Folks familiar with the matter say this is just the beginning of legal challenges for ML companies as creatives defend themselves against the implications of potential intellectual property ("IP") infringement. The legal question is centered on what "fair use," should mean exactly for language modeling engines.

The moral questions are focused primarily on the harms. Consequentialism may argue for a utilitarian society, one that seeks to strengthen overall well-being. If OpenAI for instance can show the ends justify the means when it comes to how their software was built, then they can also argue that they had no moral requirement to exclude copyrighted material that was used without consent in their training set. Because ChatGPT has improved society significantly, the authors too can benefit. However, there are many objections to this type of argument as this story of algorithms in the news helps exemplify.

First, it treats the work of creatives as an object as opposed to expressions of the subjects who created them. The act of not seeking consent seems to offend a certain level of humanity this 21st century society has negotiated itself into. Afterall, this is filed in a state that exports thousands of titles annually in all mediums. These works all are crafted with the minds and attention of their creatives. For the purposes of this paper let us continue to consider written works. Kantian Deontology is a great rebuttal for the utilitarian argument, as it highlights the principle that one should never treat a human being as a mere means.

Second, even if we assume the intentions were never to infringe, the result is that clearly infringement did occur. The concept of infringement covers any unauthorized action or violation (Legal Information Institute, 2023). The application of a copyright not only acknowledges the right of creatives, the agency over their works, but it also records this acknowledgment thus offering protections against the unauthorized use of their work. With the best of intentions, OpenAI has violated this decree unless the corporation can prove they somehow had permission of the creatives to use their works. Whereas Kantian Deontology honors the existence of rights, we can look here for the moral reasoning behind these protections and understand why some of these safety nets are already codified into law.

Diving deeper, Trembly et. al. may argue that lack of consent may undermine the agency of creatives altogether. Contemporary creatives have the ability to control when and how their published material is presented. The purpose of the publisher is just that, they fix the media and work with distributors to place a limited set of copies in various locations. The act of consuming the digitized material circumvents this level of agency completely, thus treating the authors too as objects. By stripping away their autonomy and using their works without permission, these authors are treated as a mere means for the ends of OpenAI to build their software ChatGPT.

One hurdle for the plaintiffs is how they will justify their moral reasonings behind their allegations before a jury. This technique is important if they want to convince the jury their interpretation of key concepts is the correct one to apply. Lawyers involved in the matter estimate ChatGPT contains some “294,000” titles in the training dataset (The Guardian p. 2, 2023). Demonstrating the OpenAI software operates significantly differently without the works in the training set is unlikely to be successful for the creatives since it does not make up a significant amount of the entire training set on its own. This case is subsequently filed under several provisions from state to federal statutes that deal with copyright infringement, such as violations of section 1202(b) of the Digital Millenium Copyright Act (“DMCA”) which “makes it unlawful to provide or distribute false copyright management information (CMI) with the intent to induce or conceal infringement,” (U.S.C. DCMA, 2023). The more authors “similarly positioned,” that enjoin on this class action lawsuit, the more likely they will be in demonstrating their works significantly alter the software (Trembly v. OpenAI, 94102). From there, they have proven their works were used, finally proving they did not offer their permission will show that OpenAI has violated this federal crime.

OpenAI currently faces other legal challenges as well. On the same day the creatives filed their class action lawsuit, plaintiffs in the Northern District of California launch a 157-page-long scathing lawsuit against the ML corporation, alleging several infractions, to include illegal mishandling of private data. While there are distinctions between published and private information, it is worth noting that published authors offer consent for derivative works often by contract agreement. They are often compensated in the form of royalties and other arrangements in exchange for their IP. Hence it is important for ML entities to interpret “fair use,” as works in public domain only. ML corporations may continue to find themselves in court until a precedence is set or their practices change in the training portion of their latest build. To avoid these types of lawsuits they should implement a way for creatives to participate in training that data on their own volition to ensure respect of their overall autonomy. This will reduce the harms caused by the proliferation of ML systems.

References

Guardian News and Media. (2023, July 5). *Authors file a lawsuit against OpenAI for unlawfully "ingesting" their books*. The Guardian. https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books/. Accessed 10 July 2023.

H.R.2281 - 105th congress (1997-1998): Digital Millennium Copyright Act. (n.d.). https://www.congress.gov/bill/105th-congress/house-bill/2281. Accessed 10 July 2023.

Legal Information Institute. (n.d.). *17 U.S. Code § 1202 - integrity of copyright management information*. Legal Information Institute. https://www.law.cornell.edu/uscode/text/17/1202. Accessed 10 July 2023.

Office, U. S. C. (n.d.). *The Digital Millennium Copyright Act*. The Digital Millennium Copyright Act | U.S. Copyright Office. https://www.copyright.gov/dmca. Accessed 10 July 2023.

Rivera, G. (n.d.). *2 authors say OpenAI "ingested" their books to train ChatGPT. now they're suing, and a "wave" of similar court cases may follow.* Business Insider. https://www.businessinsider.com/openai-copyright-lawsuit-authors-chatgpt-trained-on-books-2023-7. Accessed 10 July 2023.

Trembley v. OpenAI, Creamer 1. (San Francisco Fed. Ct. 94102). https://llmlitigation.com/pdf/03223/tremblay-openai-complaint.pdf. Accessed 10 July 2023.

An update on Artificial Intelligence and the law. JD Supra. (n.d.). https://www.jdsupra.com/legalnews/an-update-on-artificial-intelligence-8543469/. Accessed 10 July 2023.