CASE STUDY: GENERATIVE AI HAS AN INTELLECTUAL PROPERTY PROBLEM

Steve Desilets

MSDS 485: Data Governance, Ethics, and Law

November 19, 2023

1. Introduction

Due to recent technological advancements. generative artificial intelligence (AI) applications have exploded in popularity. For example, the popular text generation application, OpenAI's ChatGPT, has over 180 million users and the text-to-image generative AI application, Stable Diffusion, has over 10 million users (Duarte 2023; Valyeava 2023). While groundbreaking generative AI applications like these have begun fundamentally reshaping how creative works are developed, companies and artists have begun claiming that many of these applications are infringing upon their intellectual property (IP) rights (Valyeava 2023). In the Harvard Business Review's "Generative AI Has An Intellectual Property Problem," the authors discuss the ethics and legality of training generative AI applications with works protected by copyrights or trademarks (Appel, Neelbauer, and Schweidel 2023). In this paper, we review the data, data governance, relevant legislation, and data governance solutions discussed in this case study that advances ideas that could protect IP from infringement by generative AI applications.

2. The Data

In the *Harvard Business Review* paper, the authors focus on any data that meets all three of the following criteria: 1) data that have been leveraged to train commercial generative AI applications, 2) data which are protected by IP laws, and 3) data for which the generative AI application developers have not gained the creator's permission to use. The authors even highlight a specific example of such a scenario – when developers trained the Stable Diffusion application using the LAION-5B dataset that contains over 6 billion tagged images – many of which are protected by copyrights. Despite the ethics and legality of these application training practices being questionable, the quality of the data and analyses are likely quite remarkable since generative AI applications like ChatGPT may be some of the most advanced breakthroughs for humanity recently.

3. Relevance To Data Governance

While there are four Pillars of Data Governance (data stewardship, data quality, master data management, and data governance use cases), the focus of this case study primarily fits into the Data Stewardship Pillar, which implements policies related to data management, legal compliance, and ethics (Eye on Tech 2020). Similarly, when considering the data governance framework recommended by the authors of *Data Governance: The Definitive Guide*, the ethical and legal considerations presented in this article most closely relate to the "Policies" data governance framework component (Eryurek et al. 2021). The reason that this article's subject matter aligns so closely to data stewardship and data governance framework policies is that organizational data governance policies should reflect the ethical and legal standards to which companies want to hold themselves and their operations. As

the authors explain, many companies have not done enough to write and enforce internal data governance policies that protect authors and brands from copyright and trademark infringements during the training process for generative AI applications.

4. Relevant Legislation

While the authors of the *Harvard Business Review* article do not mention any specific laws, they repeatedly refer to the concept of generative AI applications violating US intellectual property laws throughout the piece. Notably, the US has a robust suite of laws for the two types of IP that are the focus of this piece. Trademarks (like photography watermarks) are protected by Lanham Act, and a range of literary, artistic, and scientific works (such as many texts and images) are protected by The Copyright Act of 1976 (World Intellectual Property Organization 2020; Legal Information Institute 2023; Wikipedia 2023). In fact, the US tradition of protecting intellectual property runs so deep that Article 1 Section 8 Clause 8 of the US Constitution specifically granted Congress the power to write legislation to protect IP at the federal level (Congress 2020). While the US has strong IP protection laws in place, the authors explain that it's now up to the judicial branch to uphold those IP protections as some of the first-ever cases to consider generative AI and IP infringement reach judges' courtrooms.

5. Data Governance Solution

The writers of the *Harvard Business Review* article explain that existing data governance practices do not sufficiently satisfy society's legal and ethical norms. Accordingly, the authors call on developers, creators, and businesses to implement new data governance policies to meet societal expectations for IP protections. The authors primarily advocate for developers needing to better comply with IP laws when sourcing data and needing to better track data provenance so that generative AI platforms can properly credit and compensate artists that inspire application outputs. Also, the authors suggest that creators and brands should take steps to better protect their copyrighted and trademarked works, like searching for their pieces in public data lakes and issuing cease-and-desist notices / licensing demand letters to IP violators. Last, the authors say that businesses entering contracts with generative AI companies should contractually demand that developers acquired the appropriate licenses for training data. While all these suggestions seem helpful, in my opinion, the best mechanism of all to protect intellectual property would be for courts to side with plaintiffs in cases like *Anderson v. Stability AI et al.* and for federal judges to clarify the exact boundaries of how protected intellectual property may and may not be leveraged by companies when training generative AI applications.

References

Appel, Gil, Juliana Neelbauer, and David A. Schweidel. 2023. "Generative AI Has an Intellectual Property Problem" *Harvard Business Review*. https://search-ebscohost-com.turing.library.northwestern.edu/login.aspx?direct=true&db=buh&AN=163013740&site=ehost-live

Congress. 2020. "Artl.S8.C8.1 Overview of Congress's Power Over Intellectual Property." *Congress*. https://constitution.congress.gov/browse/essay/artI-S8-C8-1/ALDE_00013060/

Duarte, Fabio. 2023. "Number of ChatGPT Users (Nov 2023)." *Exploding Topics*. https://explodingtopics.com/blog/chatgpt-users

Eryurek, Evren, Uri Gilad, Valliappa Lakshmanan, Anita Kibunguchy-Grant, and Jessi Ashdown. 2021. *Data Governance: The Definitive Guide*. Sebastopol, CA: O'Reilly Media.

Eye on Tech. 2020. "What is Data Governance? How Does it Impact Businesses?" YouTube, 2:16. February 14, 2020. https://www.youtube.com/watch?v=BqdPuwvwPk4

Legal Information Institute. 2023. "Lanham Act." *Cornell University*. https://www.law.cornell.edu/wex/lanham_act#:~:text=The%20Act%20provides%20for%20a,mark%20is%20likely%20to%20occur

O'Brien, Matt. 2023. "Sarah Silverman and novelists sue ChatGPT-maker OpenAI for ingesting their books." *Associated Press*. https://apnews.com/article/sarah-silverman-suing-chatgpt-openai-ai-8927025139a8151e26053249d1aeec20

Valyeava, Alina. 2023. "AI Has Already Created As Many Images As Photographers Have Taken in 150 Years. Statistics for 2023." *Everypixel Journal*. https://journal.everypixel.com/ai-image-statistics#:~:text=So%20far%2C%20we%20have%20two,million%20users%20across%20all%20channels

Wikipedia. 2023. “Copyright Act of 1976.” *Wikipedia*. https://en.wikipedia.org/wiki/Copyright_Act_of_1976

World Intellectual Property Organization (WIPO). 2020. “What is Intellectual Property?” Geneva, Switzerland: *WIPO*. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_450_2020.pdf