Thesis Project Portfolio

Custom Entity Extraction: An End-to-End Machine Learning Pipeline Tailored to Unique Customer Data

(Technical Report)

Intellectual Property in the Age of Generative AI: How Social Group Priorities Surrounding Copyright are Changing in Response to Novel AI Tools

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Madelyn Khoury

Spring, 2024

Department of Computer Science

Table of Contents

Executive Summary

In the last three years, usage of artificial intelligence (AI) has rapidly increased, on both the individual and corporate level. Various entities build, train, own, use the results of, and ultimately benefit from AI systems, but the groups who benefit from AI may not always be the groups whose work allowed the systems to be developed in the first place. My capstone research was motivated by a desire to create better AI trained on a client's behalf. Specifically, my work aimed to allow enterprise customers to extract information from documents more accurately, thus allowing for the menial and error-prone task of processing thousands of PDF documents to be outsourced to an AI system. In my capstone research, I improved a model that was trained on a customer's proprietary data, that would ultimately be used by the customer for their own benefit, and that would itself be owned by the customer. The data that was integral for training the model was a) explicitly provided for the task of training the model, and b) provided by the beneficiary of the model. In contrast, many recent AI tools have been trained on data that does not belong to the company producing the AI tool or to the beneficiary of the tool. The training data is usually not intended to be used for training an AI system, much less provided explicitly for the purpose. My STS research was fueled by a desire to investigate the response of groups whose data has been used to train AI systems. My research examined whether social groups expressed more concern about copyright and intellectual property after AI – particularly generative AI – had become more prevalent. Further, my research questioned whether groups worried more about copyright when the generative AI in question produced works that could compete with – or be inspired by – their own.

For the duration of my capstone research, I worked at Appian Corporation, a producer of enterprise software. Appian hoped to improve the performance of its intelligent document

processing (IDP) capabilities, particularly for documents that are uniquely formatted, which tend to be used by their high-volume customers. To improve accuracy of the entities extracted from customer documents, my team and I developed a custom entity extraction (EE) feature, allowing clients to use their own model (trained on their own proprietary documents) rather than a model shared among all Appian customers. I adapted and streamlined the BROS model—an open-source EE model capable of better predictions—for use by customers. Additionally, I built a Kedro machine learning pipeline to integrate this new model into the overall document-processing workflow. My team and I created an end-to-end inference workflow using the custom EE model. Future work required to complete the custom EE feature includes implementing a training workflow for the model, which will then allow for customers to fully train and use the custom model. Thus, customers are the producers of the data for – as well as the beneficiaries of the results of – the EE models in question. The custom EE feature is widely anticipated by Appian customers.

In fact, few technological artifacts from the last two years have prompted as much discussion as artificial intelligence tools. Particularly, generative artificial intelligence (AI) tools like Midjourney and ChatGPT that are capable of generating a wide range of outputs have become nearly ubiquitous. Yet, the functionality of these tools is only possible due to the massive sets of human-created media used to train them, which raises concerns about copyright. My STS research used the social construction of technology framework to examine how opinions on copyright law differed by social group, particularly focusing on occupations that may be disrupted by generative AI. Specifically, I researched how the priorities of key social groups regarding copyright law have changed in response to the advent of three popular generative AI tools, Midjourney, ChatGPT, and Github Copilot. My research involved a thematic

analysis of social media posts relating to copyright law before and after the release of these tools, which produce text, images, and computer programs, respectively. In doing so, it provided a nuanced view of how the response of key social groups varied based on the type of generative AI tool. Further, I analyzed lawsuits relating to generative AI that had been filed to change or dispute copyright law in order to qualify the actions that social groups took to satisfy their desires surrounding generative AI.

Completing my capstone and STS research simultaneously has provided me with a more nuanced understanding of how proprietary data can – and should – be handled in AI systems. In doing my capstone research, I experienced firsthand how to process training data to create a model, imparting on me an idea of how data can be feasibly handled. I was immersed in the priorities and attitudes of one of the three key social groups of my STS research paper: that of computer programmers. Additionally, I glimpsed how AI is viewed in corporate settings and heard user stories about successes and priorities relating to the use of AI. Meanwhile, my STS research allowed me to explore the priorities of *other* social groups, such as those of artists and authors, which I would not consider myself as belonging to. Additionally, I read scholarly articles as well as news stories while completing my STS research, which exposed me to ideas from academics and groups who are advocating against the use of proprietary information in AI. Thus, completing both research projects allowed me to get a broader picture of the range of opinions around the use of data in AI. Knowing these opinions will help me be more responsible as a creator of AI, a user of AI, and ultimately, a citizen who may advocate for ethical and responsible adoption of AI on a wider scale.