AI, Privacy and Copyright

Social Impact Paper
Chi Xia, anemone@bu.edu

Artificial Intelligence (AI) is increasingly becoming a significant part of modern society, transforming industries, enhancing decision-making, and reshaping the way people interact with technology. AI systems have demonstrated their immense potential to improve efficiency and convenience in everyday life, and have even stepped into the field of art.

AI has a relatively short history, being popular only for a few years, but concerns about privacy and copyright have also grown. The collection, processing, and utilization of personal data by AI systems pose significant risks to individual privacy, raising ethical and legal questions.

AI relies on huge amounts of data to improve efficiency and accuracy. This dependency on data raises concerns about how information is collected, who has access to it, and how it is used.[1] Issues such as unauthorized data sharing, algorithmic bias, and lack of transparency have highlighted the need for stronger privacy protections and ethical guidelines with AI in commerce. Unfortunately, it's unlikely that people have idea how much of their personal data has been included in datasets of AI training systems.

Recently, X has changed its privacy policy to allow third parties to use public data from the platform to train AI models, which has been in effect since November 15, 2024. Users don't have an option to opt out of this unless they quit X.[2] Many painters and illustrators have announced to cease posting their paintings online, or they are finding alternatives of X.

This is reflecting important issues regarding AI and user data:

- Data is becoming more valuable for its potential in AI training.
- Users have weaker control over their personal data.
- More vague boundaries between public and private data.
- Users' awareness of the potential uses of their content.

First, AI training requires massive amounts of data, which makes all types of data generated by individuals, companies, and platforms potentially commercially valuable. For example, a seemingly ordinary social media comment may be used to train a sentiment analysis model, and a shared photo may become a training sample for a computer vision model. This trend of data value is accelerating, but the corresponding value distribution mechanism is far from mature.

At the same time, users' control over their personal data is weakening. Although various privacy protection regulations such as GDPR[3] are constantly being improved, in actual operations, users are often in a passive position. Once data is shared on the Internet, it may be collected and used in various ways. For example, even if users delete old posts on social media, these contents may have been collected and used for AI training by third parties. It is difficult for

users to track and control the flow and use of their data.

The blurring of the boundaries between public and private data is another serious problem. In the digital age, what is "public" data? Does sharing photos on social media mean that you implicitly agree to use them for AI training? The answers to these questions are not clear. The development of technology has made it easier to collect, integrate, and analyze data, and traditional privacy boundaries are being broken.

The improvement of user awareness is crucial but challenging. Most users do not understand how their data may be used, nor do they know the long-term impacts of such use. For example, personal photos shared today may be used to train facial recognition systems, which may affect personal privacy and freedom in the future. Raising user awareness of data use requires joint efforts from platforms, regulators, and educational institutions.

Before more complete regulations are introduced, we obviously cannot rely on companies using AI themselves regulating their use of data. In the fields of art, engineering, and services, some tools to combat AI have already appeared. Here are some examples.

Glaze is a system designed to protect artists by subtly altering artworks to confuse AI models while remaining visually unchanged to humans. It achieves this by embedding modifications that make AI perceive the art as a different style, effectively disrupting unauthorized style mimicry. These changes are resilient to common transformations like resizing or filtering, ensuring robust protection for the original artist's work.[4]

Common Crawl is a non-profit organization that maintains an open repository of web crawl data, freely accessible to the public. Since its inception in 2007, Common Crawl has been collecting and storing petabytes of web data, including raw web page data (WARC), metadata (WAT), and text extractions (WET).[5] However, a report by Human Rights Watch highlighted that over 170 images and personal details of Brazilian children were used without their knowledge or consent in an open-source dataset to train AI models. These images were scraped from sources like mommy blogs and YouTube videos and included in the LAION-5B dataset, which was created using data collected from Common Crawl. This incident raises concerns about the privacy and safety of individuals whose content is collected without their awareness.[6]

GitHub Copilot is an AI-powered code completion tool developed by GitHub in collaboration with OpenAI. It leverages the OpenAI Codex model, which has been trained on billions of lines of publicly available code from various repositories, including those hosted on GitHub.[7] Many developers were unaware that their publicly shared code was used to train Copilot. This lack of transparency has led to concerns about consent and the ethical use of open-source contributions.[8] In response to these concerns, a class-action lawsuit was filed against Microsoft, GitHub, and OpenAI in November 2022. The lawsuit alleges that GitHub Copilot violates the legal rights of developers who posted code under open-source licenses, challenging the legality of using such code for training AI systems without explicit permission.[9]

Although existing countermeasures such as Glaze and digital watermarks provide certain protection capabilities, these measures are essentially passive defense measures and cannot fundamentally solve the problem of unauthorized data collection and use. One popular perspective is to truly solve this problem, we need to establish a brand new data governance system.[10] This system should first be based on a sound legal framework, clearly define the boundaries of data collection and use, require that the source of AI training data must be traceable, and establish a strict punishment mechanism. At the same time, we need to develop a data usage tracking system at the technical level to achieve visualization and controllability of the entire data flow, so that users can truly master the right to use their own data. Platforms need to assume more responsibilities, not only to provide clear data usage instructions, but also to establish a fair data usage compensation mechanism. More importantly, we need to rethink the distribution of data rights and interests and establish a data economy system that can balance innovation needs and personal rights.

REFERENCES

[1] https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/many-executives-uncertain-if-their-organizations-have-ethical-standards-for-generative-ai.html

[2] https://9meters.com/entertainment/social-media/elon-musks-x-can-now-use-your-data-to-train-its-ai

[3] https://gdpr.eu/what-is-gdpr/

[4] https://glaze.cs.uchicago.edu/what-is-glaze.html

[5] https://commoncrawl.org/overview/

[6] https://www.wired.com/story/ai-tools-are-secretly-training-on-real-childrens-faces/

[7] https://github.com/features/copilot/

[8] https://fossa.com/blog/analyzing-legal-implications-github-copilot/

[9] https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data/

[10] https://www.digitalocean.com/resources/articles/ai-and-privacy/