

Fair or Not? The Ethics of Unseen Data in AI Training

Nathan A. C. Hogg

Loyola University of Chicago, Computer Science Department

Comp 317: Social, Legal, and Ethical Issues in Computing

Dr. Smadar K. Bergman

June 22 2024

On November 30, 2022, OpenAI released ChatGPT, a free-to-use public test of their new Large Language Model (LLM). Their training dataset consisted of “text collected from multiple sources on the internet, including Wikipedia articles, books, and other public webpages.” (OpenAI, 2022) This raises the question: were page authors notified that their data was used to train an AI model? The likely answer is no. In the current “AI gold rush,” (Coresite, 2024) fair use guidelines are frequently ignored when training models. Should people be notified that their data is being used, or is anything on the internet fair game? AI companies should be legally required to inform people when their data is used to train an artificial intelligence model in order to conform to fair use, build trust in AI, and avoid potential privacy violations. There are several other factors to investigate, such as what falls under fair use, existing legal protections, impacts on businesses and financial issues, and technological limitations.

Artificial intelligence has a rich history that dates back to the mid-1900s. The term “artificial intelligence” was coined during the Dartmouth Conference in 1956, one of the first events in the field. (Dartmouth, 1956) Early artificial intelligence research focused on symbolic reasoning and problem solving. Alan Turing and Arthur Samuel laid most of the groundwork for the AI we use today. (Tableau, 2023)

Throughout the 1980s and 90s, AI saw significant advancements with the development of machine learning algorithms. These models were able to learn from data, rather than using hardcoded rules. These new abilities meant that more data was needed for the systems to work properly. After the internet boom, the era of big data emerged. The early 2000s saw a vast increase in the amount of data produced by internet users. This gave AI researchers more data to use for training and analysis. The development of deep learning in recent years has allowed researchers to take advantage of this data. Convolutional neural networks (CNNs) and recurrent

neural networks (RNNs) have made speech and image recognition accessible, though they require more data than ever. This need for data has led to AI companies scraping the open internet, often without explicit permission from the data creators.

These developments have brought on concerns about data privacy. The first factor to examine is what exactly falls under existing fair use guidelines when it comes to online content. According to Stanford University, fair use is “any copying of copyrighted material done for a limited and ‘transformative purpose, such as to comment upon, criticize, or parody a copyrighted work.” (Stanford, 2021) By this definition, the use of online materials to train AI models should be fair use. Even so, ChatGPT only provides references when it directly copies and displays something from a website. Thus, some written ideas are not given proper credit when used by ChatGPT. To conform directly with fair use guidelines, ChatGPT should, at the least, inform contributors that their material is being used in order to conform to fair use guidelines.

To investigate the problem at the source, I asked ChatGPT how it transforms the data it is fed and if it believes that the way it uses information warrants a citation. It responded with 5 bullet points labeled the following: preprocessing, tokenization, training with context, fine-tuning, and reinforcement learning. It went on to explain that in a “conversational AI context,” it is challenging to provide citations for every piece of information given the breadth of training data. This makes sense. Theoretically, one paragraph of text generated by ChatGPT could contain information from many sources. It followed up with a paragraph talking about how it can provide citations when accessing real-time information from the web browser. This is also consistent with what I already knew. I then followed up by asking who wrote the Declaration of Independence. ChatGPT gave a factual response with no citation or link to where the information came from. I then asked where it got the information. ChatGPT responded with inline citations

and links for every statement made. It should do this regardless. Currently, the only OpenAI model that can access the web is GPT4o. So, the free-to-use model, GPT3 does not provide any citations.

Another factor to consider is what rights consumers have regarding their data. The California Consumer Privacy Act (CCPA), effective January 1, 2020, is a data privacy law that grants California residents several rights regarding their personal information, like the right to know, right to delete, and right to opt-out.

The right to know guarantees that consumers can request that businesses disclose both the categories of information collected, as well as specific pieces of personal information that have been collected about them. Additionally, they can request the purposes for which the data was collected, and the types of third parties the data is shared with (advertising, analytics, etc.) (Gibson Dunn, 2023)

The right to delete gives consumers the right to request the deletion of personal information that has been collected about them, with two main exceptions: consumers can't request the deletion of information that is "likely to render impossible or seriously impair the ability to complete such research," and they cannot request deletion if they gave informed consent. (Ogletree, 2023)

Finally, the right to opt-out allows consumers to opt-out of the sale of personal information to third parties. Businesses are required by law to include a link on their webpage titled "Do Not Sell My Personal Information" to allow consumers to opt-out of this process. Also, privacy policies must contain steps to opt-out. As of January 1, 2023, the California Privacy Rights Act (CPRA) adds several points to this right, including the right to correct inaccurate

personal information and the right to limit use of personal information, furthering consumers' ability to control what happens with their data. (Ogletree, 2023)

So, how do the CCPA and CPRA impact AI? According to Akin, the CPRA directs the CCPA to issue regulations regarding automated decision-making technologies, or ADT. Specifically, the CCPA is imposing "significant regulation on the use of AI in ADT." (Akin, 2024) They define ADT as "any system, software, or process – including one derived from machine-learning, statistics, or other data-processing or AI – that processes personal information and uses computation as whole or part of a system to make or execute a decision or facilitate human decision making." Any AI software used by Californians that uses or analyzes personal data must comply with both CCPA and CPRA regulations. Therefore, companies and businesses utilizing this technology must provide pre-use notices to consumers about their use of ADT. By design, artificial intelligence analyzes data and makes decisions based on conclusions drawn from that data. Thus, most AIs are considered ADT. This legislation should be adopted in every state to ensure that consumer privacy is protected.

Although there are legal protections in place, the process of informing every user about every bit of their information that has been used to train an AI model can have impacts on the business and financial aspect of the AI market. For starters, the compute power to do so would be considerably higher, using more electricity than just training and serving the model. According to The Verge, training a model like ChatGPT3 uses up enough energy to power 130 US homes for one year. (Vincent, 2024) This figure would only increase, thus furthering both the financial and environmental impact of artificial intelligence.

It could be argued that requiring AI companies to notify users that their data has been used would create more jobs in the AI market, as legal teams would have to be consulted to write

up the documents that adhere to privacy protections. Additionally, a team would have to program the AI model to flag each entry and serve an email to the consumer. This would open the door for economic growth and help rekindle the job market globally.

One key challenge that presents itself is finding efficient ways to notify users that their data is being used. Finding a solution to this could prove to be as difficult as developing the models themselves. Even so, potential solutions include automated notification systems, and utilizing the blockchain.

Automated notification systems are a potential avenue. Users could be informed via email or text-message and given the proper steps to opt-out or request compensation. The notifications could also include things like how and where the data was used. The only downside to this is that more processing power will be required to constantly check the origin of the data.

Another solution relies on the blockchain. According to IBM, blockchain is “a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business world.” (IBM, 2021) Basically, anything stored on the blockchain is read-only, so it cannot be tampered with. Cryptocurrencies like Bitcoin and Ethereum are built on this technology because of this. If I own 50 Bitcoin, it can be proven by looking at the transactions on the chain. Additionally, the blockchain makes it easy to trace the origin, transfer, and movement of data. This ability would add transparency and allow consumers to see first-hand where their data has gone and what has been done with it. Blockchain smart contracts add additional functionality, like automated notifications when new “blocks” are created. Finally, Blockchain can facilitate decentralized data marketplaces, giving users further control over their data. They can choose to share their data with AI companies, under specific terms and perhaps for

compensation. This system fosters a more equitable relationship between consumers and developers.

Not informing users that their data is being used can open the door for numerous privacy violations. Most AI models have protections in place to avoid this, but there is still potential for gaps in security. What if a company's database was unintentionally exposed to the open web and scraped to train a model without their knowledge? Someone could query the model and ask for private information about the company, or even worse, about the company's customers. This is something to consider when using unsupervised training on the open web. According to Hidden Layer's AI Threat Landscape Report for 2024, 77% of businesses reported a breach to their AI in the last year. (Hidden Layer, 2024) This statistic shows that AI is already a target for cyber-attacks. This impact could be lessened significantly if users were made aware of what data is being used in AI models.

I can see how this type of system could be flawed. On top of the energy usage concerns, there are technical limitations as well. It would be difficult to reliably and accurately determine who wrote what information on the internet. Even while writing this paper, I have had issues with citation generators not being able to extract authors from online articles. I am sure that an AI model could be trained to do this, but that just brings the issue full circle. To effectively implement such a system would require further advances in technology.

Given these limitations, it seems logical that AI companies continue operating as they do now and simply put forth a stronger effort to let users know that their data may be used. For example, an AI company could tell a website that they are using their data to train and leave it up to that website to conform to regulations such as CCPA. This seems like a reasonable solution as most sites that are up to date legally already follow these regulations.

The added transparency efforts put forth by AI companies would likely impact trust in the new technology. Since it is a relatively new technology, there is still a lot of uncertainty surrounding it. Being open about where data is coming from, as well as how it is used would be an effective way to nudge the technology into everyday life, rather than expecting people to blindly adopt it. According to a Forbes survey, 65% of consumers say they maintain trust in businesses that are adopting AI technologies. (Forbes) This is a step in the right direction, as AI continues to evolve and solidify its place in our day-to-day lives.

In conclusion, AI companies should be required to inform users that their data is being used to train their models. After investigating several factors, such as what falls under fair use, existing legal protections, impacts on businesses and financial issues, and technological limitations, it seems that this is the right choice to protect consumers, build trust in future AI developments, and propel the AI industry forward. The California Consumer Privacy Act has already laid the framework for standardized data protection laws, so implementing this policy would be simple enough if adapted on a national scale. Two potential solutions to the issue of informing the users are using automated email systems and utilizing blockchain technology. Requiring that AI companies inform users about data usage aligns with existing legal frameworks and promotes user trust. It is a necessary step to ensure the responsible development and implementation of AI technologies. Moving forward, putting an emphasis on transparency and accountability will not only protect user privacy, but it will also pave the way for a more secure and trustworthy AI-driven future.

Works Cited

Dartmouth. (n.d.). Artificial Intelligence (AI) coined at Dartmouth. Retrieved from

<https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth>

Edsall, T. (2024a). Will A.I. be a creator or a destroyer of worlds? Retrieved from

<https://www.nytimes.com/2024/06/05/opinion/will-ai-be-a-creator-or-a-destroyer-of-worlds.html?searchResultPosition=2>

Edsall, T. (2024b). Will A.I. be a creator or a destroyer of worlds? Retrieved from

<https://www.nytimes.com/2024/06/05/opinion/will-ai-be-a-creator-or-a-destroyer-of-worlds.html?searchResultPosition=2>

Font, J. (2024). 2023: The year of the AI gold rush. Retrieved from

<https://www.coresite.com/blog/2023-the-year-of-the-ai-gold-rush>

Haan, K. (2024). 24 top AI statistics and Trends in 2024. Retrieved from

https://www.forbes.com/advisor/business/ai-statistics/#ai_trust_and_user_sentiment_section

Hidden Layer. (2024). Hiddenlayer's 2024 AI threat landscape report. Retrieved from

<https://hiddenlayer.com/threatreport2024/>

IBM. (2021). What is blockchain? Retrieved from

[https://www.ibm.com/topics/blockchain#:~:text=IBM-,What%20is%20blockchain%3F,patents%2C%20copyrights%2C%20branding\).](https://www.ibm.com/topics/blockchain#:~:text=IBM-,What%20is%20blockchain%3F,patents%2C%20copyrights%2C%20branding).)

- Johri, S. (2023). The making of chatgpt: From data to dialogue. Retrieved from <https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/>
- Kessler, A. (2024). The dial-toning of ai. Retrieved from <https://www.wsj.com/articles/the-dial-toning-of-ai-b160be07>
- Kohne, N. G., Hold, J., Kurzweil, R. C., & Reed, M. A. (2024). Proposed California regulations for Automated Decision-making. Retrieved from <https://www.akingump.com/en/insights/blogs/ag-data-dive/proposed-california-regulations-for-automated-decision-making>
- Lessig, L. (2024). Opinion: The risks of AI could be catastrophic. we should empower company workers to warn US. Retrieved from <https://www.cnn.com/2024/06/06/opinions/artificial-intelligence-risks-chat-gpt-lessig/index.html>
- Mitchell, S. (2023). California Superior Court halts enforcement of certain California Privacy Regulations. Retrieved from <https://www.gibsondunn.com/california-superior-court-halts-enforcement-of-certain-california-privacy-regulations/>
- OpenAI. (2022). Introducing ChatGPT. Retrieved from <https://openai.com/index/chatgpt>
- Perry, B. W. (2023). California court delays enforcement of new CCPA regulations until March 2024. Retrieved from <https://ogletree.com/insights-resources/blog-posts/california-court-delays-enforcement-of-new-ccpa-regulations-until-march-2024/>
- Stim, R. (2021). What is fair use? Retrieved from <https://fairuse.stanford.edu/overview/fair-use/what-is-fair-use/>

Vincent, J. (2024). How much electricity does AI consume? Retrieved from

<https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption>

What is the history of Artificial Intelligence (AI)? (n.d.). Retrieved from

<https://www.tableau.com/data-insights/ai/history#:~:text=Birth%20of%20AI%3A%201950%2D1956&text=into%20popular%20use.-,Dates%20of%20note%3A,ever%20learn%20the%20game%20independently.>