ALM | LAW.COM

Page Printed From:
https://www.law.com/newyorklawjournal/2024/05/17/the-ai-revolution-comes-with-data-privacy-risks-what-consumers-should-know/

The 'AI Revolution' Comes With Data Privacy Risks: What Consumers Should Know

This article covers the legal and privacy considerations related to the rise of artificial intelligence. The authors discuss what AI is and how companies obtain the data for those sets along with the risks consumers face in having their data collected and ingested into AI systems, specifically for training purposes. They also identify possible ways consumers can protect their data.

May 17, 2024 at 10:00 AM

Privacy

By Carol C. Villegas, Michael P. Canty, Danielle Izzo and Gloria J. Medina | May 17, 2024 at 10:00 AM

Tech companies developing artificial intelligence ("AI") systems have consumed nearly all publicly available data on the internet, but this data, alone, is not enough to train their systems. Now, with little public data left to ingest, the tech industry is turning to consumers to collect and use their personal data to train their AI systems—a highly invasive data collection practice. Consumers should take note: these data collection practices pose a variety of serious privacy issues.

In recent months, artificial intelligence and the "AI Revolution" have been at the forefront of business, media, and government attention. AI technology is rapidly evolving at a pace far faster than the development of applicable laws and regulations.

Big Tech companies are releasing extensive AI search features across platforms and enhancing their targeted marketing technology. For example, Google recently launched a new generative AI feature for Chrome, expanding Google's browsing and personalization features. Meta introduced an "Ask Meta AI anything" search function on its Instagram and Facebook social media platforms.

Additionally, OpenAI released the ChatGPT artificial intelligence search assistant, which has garnered tremendous public attention. Now, other tech companies are working to keep pace.

These tech companies are scraping every corner of the internet to obtain the massive amounts of data required to train and launch their AI systems. Indeed, companies require billions of datapoints to build and train the algorithms and machine learning models underlying AI tools. *See* Karen Hao, *The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why.*, MIT Tech. Rev. (Oct. 5, 2021), https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/; *see, e.g.*, Aatish Bhatia, *Watch an A.I. Learn to Write by Reading Nothing but Jane Austen,* N.Y. Times (Apr. 27, 2023), https://www.nytimes.com/interactive/2023/04/26/upshot/gpt-from-scratch.html. Companies need far more; real-time data is placing consumers and their personal data at risk, which begs the question: what are the privacy costs to consumers?

Understanding AI Systems

To assess the privacy risks associated with AI technology, it is first important to understand the process for developing AI systems.

AI is powered by algorithms, mathematical procedures or rule sets that are designed to analyze data, discern patterns, and reach related conclusions. *See Algorithm*, Merriam Webster Dictionary Online (last visited May 3, 2024), https://www.merriam-webster.com/dictionary/algorithm. An "algorithm is simply a strand of coded instructions for completing a task or solving a problem," but when used to power AI, algorithm structures are far more complex, drawing conclusions based on billions of data points and, on an automated basis, incorporating each conclusion into the next analysis—thus, building "knowledge." *How do Algorithms Work?*, Univ. of York (last visited May 3, 2024), https://online.york.ac.uk/how-do-algorithms-work/#:~:text=An%20algorithm%20is%20a%20coded,to%20produce%20the%20desired%20result

Machine learning models require an enormous amount of data in order to build this "knowledge." *See* Danilo Bzdok *et al.*, *Machine Learning: A Primer*, 14 Nature Methods 1119 (2017); Cecilia Kang *et al.*, *Four Takeaways on the Race to Amass Data for A.I.*, N.Y. Times (Apr. 6, 2024), https://www.nytimes.com/2024/04/06/technology/ai-data-tech-takeaways.html ("A.I. models become more accurate and more humanlike with more data.").

Data is used to train the AI algorithms through a repetitive "trial and error" process, thus exposing the algorithms to various data points for training. In training, algorithms are presented with massive data sets and are tasked with drawing conclusions based on the data. *See* Samuel R. Bowman, *Eight Things to Know about Large Language Models (Unpublished Manuscript)*, NYU Courant Inst. of Mathematical Scis. (2023), https://cims.nyu.edu/~sbowman/eightthings.pdf. Human employees or reviewers conduct quality control checks on the algorithms' conclusions or "outputs" to train the models to distinguish between correct and incorrect outputs.

This training builds the artificial knowledge. *See, e.g.* Munsif Vengattil & Paresh Dave, *Facebook 'Labels' Posts by Hand, Posing Privacy Questions*, Reuters (May 6, 2019), https://www.reuters.com/article/us-facebook-ai/facebook-labels-posts-by-hand-posing-privacy-questions-idUSKCN1SC01T/ (discussing algorithm training).

Where Do Companies Get Training Data?

As discussed above, companies go to great lengths to obtain the massive data sets required to train their nascent AI systems. Recent reports reveal that OpenAI (developer of the ChatGPT AI system) used its technology to scrape any data available on the internet to collect and use to train its AI systems, but this data alone is not enough. As a result, OpenAI is developing speech recognition tools to transcribe audio and collect information from YouTube videos and other video sources. *See* Cade Metz *et al.*, *How Tech Giants Cut Corners to Harvest Data for A.I.*, N.Y. Times (Apr. 8, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

Worse than internet scraping, some companies purchase data sets from data brokers, such as LexisNexis, containing consumers' personal data amassed from the businesses they engage with, including car companies. *See* Alice Holbrook, *When LexisNexis Makes a Mistake, You Pay For It*, Newsweek (Sept. 26, 2019), https://www.newsweek.com/2019/10/04/lexisnexis-mistake-data-insurance-costs-1460831.html (reporting that LexisNexis "aggregates and sells consumer data…helping other companies figure out whether to renew your insurance, approve your loan or offer you a job," etc.); Kashmir Hill, *Automakers Are Sharing Consumers' Driving Behavior with Insurance Companies*, N.Y. Times (March 13 2024), https://www.nytimes.com/2024/03/11/technology/carmakers-driver-tracking-insurance.html.

Other companies, like Meta, intercept data by embedding mobile app and web tracking tools, like the Meta pixel and SDKs, to surreptitiously intercept users highly sensitive information, including location and health data. *See* Kristin Cohen, *Location, Health, and Other Sensitive Information: FTC Committed to Fully Enforcing the Law Against Illegal Use* *and Sharing of Highly Sensitive data*, Fed. Trade Comm'n (July 11, 2022), https://www.ftc.gov/business-guidance/blog/2022/07/location-health-and-other-sensitive-information-ftc-committed-fully-enforcing-law-against-illegal (explaining how SDKs collect sensitive information); *see also In re Meta Pixel Healthcare Litig.*, No. 22-cv-03580 (N.D. Cal.) (alleging Meta's pixel tool intercepts and transmits sensitive patient health information from medical websites).

Increasing Public Privacy Concerns

The data collection practices described above raise serious questions about individuals' data privacy. Indeed, companies have collected billions of data points across the many mobile apps and websites consumers use daily.

According to the Pew Research Center, 81% of American consumers fear that data collected by companies that work with AI will be used in ways that make them uncomfortable. Michelle Faverio, *Key Findings About Americans and Data Privacy*, Pew Rsch. Ctr. (Oct. 18, 2023), https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/.

And consumers are right to worry. A recent *New York Times* investigation found that "three major players in this race [for more data,] OpenAI, Google, and Meta[,]…[are] *willing to do almost anything* to get their hands on [] data, including ignoring, and in some cases, violating corporate rules and wading into a legal gray area." Cade Metz, *A.I.'s Original Sin*, N.Y. Times (April 16, 2024), https://www.nytimes.com/2024/04/16/podcasts/the-daily/ai-data.html?showTranscript=1 (emphasis added).

Risks of "Data at All Costs"

Companies have collected billions of data points from across the internet and are still seeking more. Now, consumers' most sensitive data is at risk of being used to train AI algorithms. Last year, pharmaceutical company GoodRx was fined $1.5 million by the Federal Trade Commission for "shar[ing] customers' health data with Google, Facebook and other third parties…[even after] pledging not to share user data." Ruth Reader, *FTC Cracking Down on Companies That Share Customers' Health Data*, Politico (Feb. 1, 2023), https://www.politico.com/news/2023/02/01/ftc-cracking-down-on-companies-that-share-customers-health-data-00080672. And earlier this year, the public learned that Google collected "billions of personal records" from Chrome Incognito users after assuring them that their browsing information would not be tracked. Michael Liedtke, *Google Will Purge Billions of Files Containing Personal Data in Settlement of Chrome Privacy Case*, Associated Press (Apr. 1, 2024), https://apnews.com/article/google-chrome-privacy-lawsuit-settlement-203cc5063f1a1d4013de1900d9376814.

Consumers also face the risk of having their data reviewed by human employees during the AI training process. As discussed above, AI algorithms are trained with massive data sets. However, these data sets require human annotation before they can be used in training. This means that human employees must review, categorize, and annotate data in preparation for AI training—exposing consumers' personal data to unknown company employees. *See* Josh Dzieza, *AI Is a Lot of Work*, The Verge (Jun. 20, 2023), https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots.

Worse yet, consumers often cannot rely on terms of service and privacy policies to sufficiently disclose companies' true data collection and use practices. The FTC recently took action against several companies for misrepresenting their data practices to consumers. Most notably, in 2023, the FTC banned a mental health services app BetterHelp from sharing user's sensitive mental health information with third parties, including Meta and Snapchat, after assuring users that "it would not use or disclose their personal health data except for limited purposes, such as to provide counseling services." *FTC to Ban BetterHelp from Revealing Consumers' Data, Including Sensitive Mental Health Information, to Facebook and Others for Targeted Advertising*, Fed. Trade Comm'n (Mar. 2, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/03/ftc-ban-betterhelp-revealing-consumers-data-including-sensitive-mental-health-information-facebook.

The FTC also took action against Epic Games, Inc. (creator of the video game Fortnite) for failing to take any steps to obtain consent from parents for its data sharing practices, despite having internal evidence that many of its users were under 13 years of age. *Fortnite Video Game Maker Epic Games to Pay More Than Half a Billion Dollars over FTC Allegations of Privacy Violations and Unwanted Charges*, Fed. Trade Comm'n (Dec. 19, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/12/fortnite-video-game-maker-epic-games-pay-more-half-billion-dollars-over-ftc-allegations.

Protect Data and Privacy

At this point in the "AI Revolution," data has been collected from millions of individuals and used without their knowledge or consent. What can consumers do to remedy these invasions of privacy and protect their data moving forward?

First, consumers can proactively protect their data by "opting out" of data sharing at the data broker level. Data brokers, such as LexisNexis, provide consumers with the ability to complete opt-out forms to stop certain personal data from being shared. *See, e.g.*, LexisNexis Opt-Out Form, LexisNexis (last visited May 3, 2024), https://optout.lexisnexis.com/; Thomas Claburn, *How to Spot OpenAI's Crawler Bot and Stop It Slurping Sites for Training Data*, The Register (Aug. 8, 2023), https://www.theregister.com/2023/08/08/openai_scraping_software/.

Consumers may also consider enabling certain privacy settings like Apple's "Ask App Not to Track" feature. *See If an App Asks to Track Your Activity*, Apple, Inc. (last visited May 3, 2024), https://support.apple.com/en-us/102420. However, there are no guarantees that Apple is able to enforce this setting across all apps and, at best, these purported protections are only limited to information accessed through Apple's devices.

Consumers can also seek out browsing and messaging platforms that prioritize privacy, like Brave, non-profit browser Tor, and private messaging service, Signal.

But is this enough? With the number of apps and websites in existence surreptitiously collecting data, consumers are left guessing as to which apps and websites track what—and *how* their personal data is actually used.

One driver of systemic change across the industry that is available to all consumers looking to protect their privacy, learn how their data is being used, and seek redress if their privacy rights are violated, is the legal system. Consumers can seek remedies for invasions of their privacy by filing claims in lawsuits and arbitrations.

Given that many claims are statutory in nature—meaning that the violation carries a penalty requiring an actual dollar payment (anywhere from hundreds to thousands of dollars per claim)—consumers bringing a critical mass of claims can have beneficial attitude- and policy-shifting effects on companies that collect data in an unauthorized manner.

And beyond seeking monetary compensation, consumers can also seek injunctive relief, forcing companies to change their policies and behaviors. In dealing with the future of AI and data collection practices, injunctive relief is particularly attractive, allowing consumers to reshape companies' data collection and use practices on a systemic level for the benefit of all future consumers.

Carol C. Villegas *and* Michael P. Canty *are partners at Labaton Keller Sucharow.* Danielle Izzo and Gloria J. Medina are associates at the firm.