PRIVACY

TRAINING DATA SETS SEEM TO INCLUDE PERSONAL DATA

This includes data both scraped from the web and provided by the customers of the AI vendor in question.

MODELS CAN'T "UNLEARN", YET

Once a model has trained on a data set, removing that data from the model is difficult. "Machine Unlearning" is still immature, and it's uncertain whether it can be made to work on models like GPT-4.

LANGUAGE MODELS ARE VULNERABLE TO MANY PRIVACY ATTACKS

Attackers can discover whether specific personal data was in the training data set. They can often reconstruct and extract specific data. Some attacks let you infer the specific properties of the data set, such as the gender ratios of a medial AI.

HOSTED SOFTWARE HAS FEWER PRIVACY GUARANTEES

Many major AI tools are *hosted*, which limits the privacy assurances they can make. Pasting confidential data into a ChatGPT window *is effectively leaking it*. Do not enter private or confidential data into hosted AI software.

DATA IS OFTEN REVIEWED BY UNDERPAID WORKERS

Even if personal data in the training set doesn't end up in the model itself, much of that data is reviewed by a small army of underpaid workers.

AI VENDORS ARE BEING INVESTIGATED

Privacy regulators are looking into AI industry practices. That includes most major European countries, the EU itself, Canada, four regulatory bodies in the US, and more. The US FTC has forced tech companies in the past to delete models that were trained on unauthorised personal data.

REFERENCES

For more, see chapters 10 and 12 in *The Intelligence Illusion*

Al-Sibai, Noor. "Amazon Begs Employees Not to Leak Corporate Secrets to ChatGPT." *Futurism*, 2023. https://futurism.com/the-byte/amazon-begs-employees-chatgpt.

Barr, Kyle. "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery." *Gizmodo*, March 2023. https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv, October 2021. https://doi.org/10.48550/arXiv.2110.01963.

Bourtoule, Lucas, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. "Machine Unlearning." arXiv, December 2020. https://doi.org/10.48550/arXiv.1912.03817.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting Training Data from Diffusion Models." arXiv, January 2023. https://doi.org/10.48550/arXiv.2301.13188.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. "Extracting Training Data from Large Language Models." arXiv, June 2021. https://doi.org/10.48550/arXiv.2012.07805.

Coles, Cameron. "3.1% of Workers Have Pasted Confidential Company Data into ChatGPT." *Cyberhaven*, February 2023. https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/.

Council, Stephen. "OpenAI Admits Some Premium Users' Payment Info Was Exposed." SFGATE, March 2023. https://www.sfgate.com/tech/article/chatgpt-openai-payment-data-leak-17858969.php.

Di, Jimmy Z., Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. "Hidden Poison: Machine Unlearning Enables Camouflaged Poisoning Attacks." arXiv, December 2022. https://doi.org/10.48550/arXiv.2212.10717.

Edwards, Benj. "Artist Finds Private Medical Record Photos in Popular AI Training Data Set." *Ars Technica*, September 2022. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/.

Feiner, Lauren. "U.S. Regulators Warn They Already Have the Power to Go After A.I. Bias — and They're Ready to Use It." CNBC, April 2023. https://www.cnbc.com/2023/04/25/us-regulators-warn-they-already-have-the-power-to-go-after-ai-bias.html.

Fraser, David. "Federal Privacy Watchdog Probing OpenAI, ChatGPT Following Complaint CBC News." CBC, April 2023. https://www.cbc.ca/news/politics/privacy-commissioner-investigation-openai-chatgpt-1.6801296.

"FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI." *Federal Trade Commission*, April 2023. https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

"FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology." *Federal Trade Commission*, May 2021. https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology.

Gal, Uri. "ChatGPT Is a Data Privacy Nightmare, and We Ought to Be Concerned." *Ars Technica*, February 2023. https://arstechnica.com/

information-technology/2023/02/chatgpt-is-a-data-privacy-nightmare-and-you-ought-to-be-concerned/.

Inan, Huseyin A., Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. "Training Data Leakage Analysis in Language Models." arXiv, February 2021. https://doi.org/10.48550/arXiv.2101.05405.

"Intelligenza Artificiale: Il Garante Blocca ChatGPT. Raccolta Illecita Di Dati Personali. Assenza Di Sistemi Per La Verifica Dell'età Dei Minori," March 2023. https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870847.

Kan, Michael. "OpenAI Confirms Leak of ChatGPT Conversation Histories." PCMAG, March 202AD. https://www.pcmag.com/news/openai-confirms-leak-of-chatgpt-conversation-histories.

Kumar, Vinayshekhar Bannihatti, Rashmi Gangadharaiah, and Dan Roth. "Privacy Adhering Machine Un-Learning in NLP." arXiv, December 2022. https://doi.org/10.48550/arXiv.2212.09573.

Lee, Dave. "Someone Asked ChatGPT If It Had a Signal… And It Gave Him *MY NUMBER*." *Twitter*, February 2023. https://twitter.com/DaveLeeFT/status/1626288109339176962.

Lomas, Natasha. "FTC Settlement with Ever Orders Data and AIs Deleted After Facial Recognition Pivot." *TechCrunch*, January 2021. https://techcrunch.com/2021/01/12/ftc-settlement-with-ever-orders-data-and-ais-deleted-after-facial-recognition-pivot/.

McGowan, Emma. "Is ChatGPT's Use of People's Data Even Legal?" February 2023. https://blog.avast.com/chatgpt-data-use-legal.

Milmo, Dan. "ChatGPT Reaches 100 Million Users Two Months After Launch." *The Guardian*, February 2023. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

Mukherjee, Supantha, Elvira Pollina, and Rachel More. "Italy's ChatGPT Ban Attracts EU Privacy Regulators." *Reuters*, April 2023. https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/.

Pan, Xudong, Mi Zhang, Shouling Ji, and Min Yang. "Privacy Risks of General-Purpose Language Models." In 2020 IEEE Symposium on Security *and* Privacy (SP), 1314–31, 2020. https://doi.org/10.1109/SP40000.2020.00095.

Perrigo, Billy. "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer." *Time*, January 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Raieli, Salvatore. "Machine Unlearning: The Duty of Forgetting." *Medium*, September 2022. https://towardsdatascience.com/machine-unlearning-the-duty-of-forgetting-3666e5b9f6e5.

"Regulatory Framework Proposal on Artificial Intelligence Shaping Europe's Digital Future," February 2023. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Rigaki, Maria, and Sebastian Garcia. "A Survey of Privacy Attacks in Machine Learning." arXiv, April 2021. https://doi.org/10.48550/arXiv.2007.07646.

Shepardson, David, Diane Bartz, and Diane Bartz. "US Begins Study of Possible Rules to Regulate AI Like ChatGPT." *Reuters*, April 2023. https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models." arXiv, March 2017. https://doi.org/10.48550/arXiv.1610.05820.

Simonite, Tom. "Now That Machines Can Learn, Can They Unlearn?" *Wired*. Accessed February 21, 2023. https://www.wired.com/story/machines-can-learn-can-they-unlearn/.

Sterling, Toby. "European Privacy Watchdog Creates ChatGPT Task Force." *Reuters*, April 2023. https://www.reuters.com/technology/

european-data-protection-board-discussing-ai-policy-thursday-meeting-2023-04-13/.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. "Ethical and Social Risks of Harm from Language Models." arXiv, December 2021. https://doi.org/10.48550/arXiv.2112.04359.

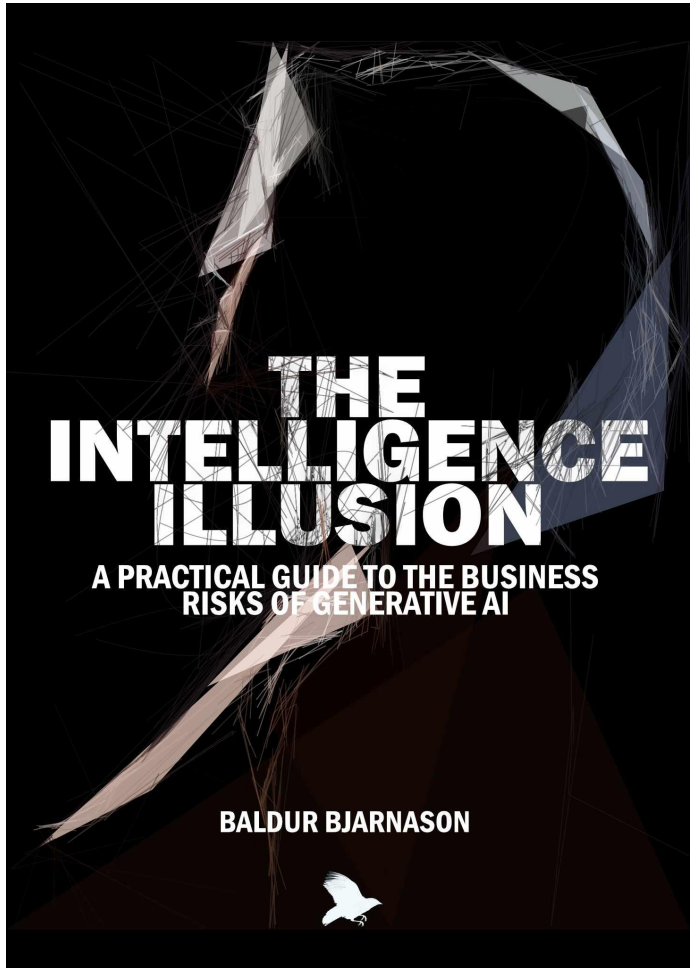
Wiggers, Kyle. "Addressing Criticism, OpenAI Will No Longer Use Customer Data to Train Its Models by Default." *TechCrunch*, March 2023. https://techcrunch.com/2023/03/01/addressing-criticism-openai-will-no-longer-use-customer-data-to-train-its-models-by-default/.

Writer, Robert LemosContributing, Dark ReadingMarch 07, and 2023. "Employees Are Feeding Sensitive Business Data to ChatGPT." *Dark Reading*, March 2023. https://www.darkreading.com/risk/employees-feeding-sensitive-business-data-chatgpt-raising-security-fears.

Xiang, Chloe. "OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit." *Vice*, February 2023. https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

Xiang, Chloe, and Emanuel Maiberg. "ISIS Executions and Non-Consensual Porn Are Powering AI Art." *Vice*, September 2022. https://www.vice.com/en/article/93ad75/isis-executions-and-non-consensual-porn-are-powering-ai-art.

Yeom, Samuel, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting." arXiv, May 2018. https://doi.org/10.48550/arXiv.1709.01604.



GENERATIVE AI: WHAT YOU NEED TO KNOW