Safeguarding Data Privacy and Well-Being for AI Chatbot Users

Dr. Jennifer King[1]
Stanford University

Testimony presented to the U.S. House Committee on Energy and Commerce, Subcommittee on Oversight and Investigations, November 18, 2025.

Executive Summary

Americans want limits on the types of data companies collect about them, especially when that data is sensitive personal data related to their health. While technologies designed for and used specifically in healthcare settings are governed by the Health Insurance Portability and Accountability Act, general-purpose tools like chatbots are not. Yet consumers are increasingly turning to these chatbots for health-related concerns, including mental health support.

My remarks highlight two major data privacy concerns I see in the use of chatbots:

1. **Users are increasingly disclosing highly sensitive personal information to chatbots**, which are designed to mimic human conversation and maximize user engagement. Large platforms are contemplating how to monetize this data in other parts of their businesses.
2. **Developers are incorporating chatbot-derived user data into model training without oversight.** Their privacy policies demonstrate a lack of transparency regarding whether and how they take steps to mitigate privacy risks, including for children's data.

To address these concerns, I recommend three specific areas for congressional attention:

- **Implement data privacy and safety design principles.** Demand that chatbot developers institute both data privacy and health and safety design principles that prioritize the trust and well-being of the public.
- **Minimize the scope of personal data in AI training.** Mandate that developers make transparent their data collection and processing practices. Users should not be automatically opted in to model training, and developers should proactively remove sensitive data from training sets.
- **Demand that developers adopt safety metrics.** Developers must track and report metrics related to user privacy, safety, and experiences of harm and increase vetted researcher access to chatbot training data to ensure independent review and ensure accountability.

[1] Privacy and Data Policy Fellow, Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Statement

Chairman Joyce, Ranking Member Clarke, and Members of the Subcommittee, thank you for inviting me to appear before you today on the important issue of the use of AI chatbots and their risks for users, especially children and teenagers.

I am a research fellow with the Stanford Institute for Human-Centered Artificial Intelligence, where my research focuses on understanding the data privacy impacts of emerging technologies, such as AI chatbots, on society. I have researched and published widely on consumer data privacy concerns related to emerging technologies. I speak to you in my personal capacity, and my views represent my expertise and not the views of my employer.

Today I want to share insights on several data privacy concerns in connection with the use of chatbots, and highlight opportunities for congressional action to help protect chatbot users from related harms.

The Context: Current State of Governing Data Collection and Privacy

Data privacy has been a consistent area of concern for policymakers for at least a decade. At the federal level, this committee has introduced two data privacy bills to provide American consumers with protections over their personal information collected by technology companies. My home state of California passed the first consumer privacy act in 2018, and 19 other states have since passed consumer data privacy protection legislation.

Americans want limits on the types of data companies collect about them.[2] In particular, they support heightened protections for any sensitive personal data they disclose to companies. In no area is this more true than in our personal health. While the technologies that our doctors use to manage our medical records and communicate with us are protected by the Health Insurance Portability and Accountability Act (HIPAA), outside of licensed clinical settings we have few if any protections on data relating to health, including our mental health. For nearly two decades, consumers have lived with the reality that the data they share with health-related mobile apps may be breached, shared, or sold without their knowledge or consent. The Federal Trade Commission has taken action against multiple companies for the unauthorized disclosure of consumers' health data, yet these problematic data practices persist.[3]

---

[2] Colleen McClain et al., "How Americans View Data Privacy," Pew Research Center, October 18, 2023.
[3] See, for example, Federal Trade Commission, "FTC Gives Final Approval to Order Banning BetterHelp from Sharing Sensitive Health Data for Advertising, Requiring It to Pay $7.8 Million," press release, July 14, 2023; Federal Trade Commission, "FTC Enforcement Action to Bar GoodRx from Sharing Consumers' Sensitive Health Info for Advertising." press release, February 1, 2023; Federal Trade Commission, "FTC Finalizes Order with Flo Health, a Fertility-Tracking App that Shared Sensitive Health Data with Facebook, Google, and Others," press release, June 22, 2021.

This is the backdrop against which OpenAI released ChatGPT, and other developers launched additional consumer-facing generative AI chatbots. Unlike other forms of AI applications in the healthcare sector — which may be purpose-built, trained on curated datasets, and comply with HIPAA regulations — general-purpose chatbots are trained on massive datasets and marketed as offering something for everyone. Consumers can engage in human-like conversations with them, and many offer specifically designed products intended to function as personal companions. Increasingly, consumers are also turning to these chatbots for health-related concerns, including mental health support.

The data privacy implications are immense. I would like to share two important privacy concerns I believe warrant particularly close policy attention.

1. Users Are Disclosing Highly Sensitive Personal Information to Chatbots

The conversational design of consumer-facing, general-purpose AI chatbots encourages users to disclose vast amounts of highly personal and sensitive information, including health data, to chatbot developers.

Chatbots are designed to mimic human conversation. They are also, by design, excessively flattering and agreeable. This is a result of a business model inherited from social media platforms that aims to maximize engagement for commercial gain.[4] More than ever, chatbots can engage consumers in unlimited conversations as long as the user is willing to continue, opening up unlimited opportunities to collect data. Recent reporting has also documented how prolonged chat sessions can induce psychosis in chatbot users.[5]

The conversational nature of chatbot interactions can encourage consumers to disclose highly personal and sensitive information.[6] While this behavior may be at its most extreme when consumers deliberately seek out mental health advice or have developed parasocial relationships[7] with a chatbot, consumers who simply seek advice and information may inadvertently disclose in-depth personal details about a physical or mental health concern.

---

[4] Jim Steyer, "AI Companies' Race for Engagement Has a Body Count," *Tech Policy Press*, August 28, 2025.
[5] Kashmir Hill and Dylan Freedman, "Chatbots Can Go into a Delusional Spiral. Here's How It Happens," *New York Times*, August 12, 2025.
[6] Katherine Tangalakis-Lippert and Henry Chandonnet, "OpenAI Quickly Rolled Back a New Feature That Allowed Users to Make Private Conversations with ChatGPT Searchable," *Business Insider*, July 31, 2025; Imran Rahman-Jones, "Meta AI Searches Made Public — But Do All Its Users Realise?," BBC, June 13, 2025.
[7] Coralie Kraft, "They Fell in Love with A.I. Chatbots — and Found Something Real," *New York Times*, November 5, 2025.

This is concerning because large platforms such as Microsoft and Google are already contemplating how to integrate chatbot-derived data into their advertising businesses as they integrate their AI chatbots across their existing product suites.[8] While OpenAI and Anthropic currently pledge that they are not using their chat data to profile consumers, we should expect that all of the developers will consider both integrating paid advertising into chats and how to monetize their customers' chat disclosures. The incentives to gather as much consumer data as possible and use it across multiple contexts are immense, especially as frontier developers create AI agents for personal task automation. A seemingly straightforward agentic task such as booking a plane ticket will require a significant amount of personal information to conduct.

2. Developers Incorporate Personal Data Into Model Training Without Oversight

Chatbot developers presently face little to no oversight when it comes to handling the increasingly personal and sensitive data provided by their customers. Their privacy policies demonstrate a lack of transparency regarding whether and how they take steps to mitigate privacy risks.

I recently co-authored a study that found that the six U.S.-based frontier LLM developers offered chatbots with default settings that opt their users into having their chatbot data used for model retraining.[9] At least two developers did not appear to allow their customers to ever *opt out* of model training. Half of the developers appear to retain chat data indefinitely. Developers may also train their models on files their customers upload to chat platforms, such as photos, videos, voice recordings, and documents. We could not verify whether all developers de-link chats from customer accounts prior to model training, and only one developer explicitly stated that they attempt to remove identifiable personal information from chats prior to training. Notably, these concerns are faced only by consumers. Enterprise versions of these products do not include business users' chats in training data.

Developer policies are mixed in terms of how they approach children's data. While some do not allow children under 18 years old to create accounts, others do but ensure they do not train on their data. Some do not state how they treat the data of children 13 and older and presumably do not differentiate between minors' and adults' data.

Training on this sensitive data is concerning because LLMs can memorize their training data and then regurgitate it verbatim.[10] To date, memorization has been demonstrated with training data

[8] Jennifer King et al., "An Analysis of Frontier Developers' Privacy Policies," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 2 (October 15, 2025):1465-77. doi:10.1609/aies.v8i2.36646.
[9] Ibid.
[10] See generally: Milad Nasr et al., "Scalable Extraction of Training Data from Aligned, Production Language Models," *International Conference on Learning Representations*, 2025; Jamie Hayes et al., "Measuring Memorization in Language Models via Probabilistic Extraction," *Proceedings of the 2025 Conference of the Nations*

scraped from the internet, but as chatbots are increasingly trained on consumer interactions, the stakes of memorization increase. This data is far more personal in nature and more revealing of an individual's psychological state than data typically found on the internet or behavioral data collected by online ad networks.

Policy Recommendations

Consumers who use commercial chatbots — whether for health-related reasons or not — face substantial privacy risks. I recommend three specific areas for congressional attention.

First, we must demand that chatbot developers institute both data privacy and health and safety design principles that prioritize the trust and well-being of the public.

A core misalignment between how chatbots are designed and how the public uses them is that the public wants to use these tools in ways that should not be subject to commercial pressures, such as for mental health support. Digital tools purpose-built for healthcare contexts respect the HIPAA-protected patient-doctor relationship, and the data they generate cannot be repurposed outside of the healthcare context. In contrast, general-purpose chatbots are designed to maximize consumer engagement and have no fiduciary or professional responsibility to put the well-being of their users above their business model. This misalignment in goals is at the heart of why the public's use of commercial chatbots for therapeutic purposes or companionship presents significant risks.

Similar to safety standards that govern how Americans drive cars or consume food, we need baseline privacy, security, and safety design requirements that make consumer technology products such as chatbots safer to use. These requirements could include interventions to nudge consumers toward healthier behavior and placing limits on the length and frequency of chat sessions. While consumer chatbots should not be used to replace trained mental health professionals, consumer trust in technology in general will continue to decline without basic safeguards.

Second, we must minimize the scope of personal data that AI developers can use to train their models.

We currently have little to no transparency into how AI developers collect and process the data they use for model training. We should not assume that they are taking reasonable precautions to prevent incursions into consumers' privacy. Existing mitigation strategies often cited by developers aren't enough. For example, output suppression — a technique that prevents chatbots

---

of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies 1, April 2025:9266-91.

from outputting specific data types such as phone numbers or birthdates — can fail with complex or nuanced prompts and is dependent on developers foreseeing privacy-violating topics in advance.

Developers must provide detailed source information on data used to train their models, and document the steps they take to reduce the amount of personal information in training sets.[11] When they want to train models using customer chat data, they should have to first remove personal information from said data and, for example, by default remove entire chats identified as relating to sensitive contexts such as health issues. Users should not be opted-in to model training by default, as they are now. Tools such as temporary chats that do not persist across sessions should be available on all platforms. Developers must also be able to respond to consumer data rights requests, such as requests to delete personal information from training data. Removing personal data from massive foundation model training sets may be challenging and resource intensive but theoretically not impossible.[12]

Third, we must demand that developers adopt safety metrics and provide opportunities to verify related findings.

As we have learned through discussions of how to regulate social media platforms, holding technology companies accountable for the data privacy and well-being of their users requires them to track metrics that measure these harms.[13] California could offer a potential model for accountability and transparency measures: Recently passed SB 243 requires operators of companion chatbots to annually report crisis service provider referrals while the California Consumer Privacy Act requires companies to annually post the number of data rights requests they receive and fulfill. Neither of these examples is perfect and, in fact, Congress should consider going further. For instance, safety metrics must also capture consumers' own experiences of harm. This could be achieved by requiring developers to adopt standardized safety reporting categories and making the reports available to the public. Relatedly, increasing vetted researcher access to chatbot training data would allow for independent review and assessment of developers' performance while also helping to ensure accountability.[14]

---

[11] Nishant Subramani et al., "Detecting Personal Information in Training Corpora: An Analysis," *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, July 2023: 208-20.

[12] Borkar, Jaydeep et al., "Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training," *arXiv preprint*, June 25, 2025.

[13] Arturo Bejar, Written Testimony of Arturo Bejar before the Subcommittee on Privacy, Technology, and the Law, November 7, 2023.

[14] Kevin Klyman et al., "Safeguarding Third-Party AI Research," *Stanford Institute for Human-Centered AI*, February 13, 2025.

Conclusion

There is still much that researchers and even AI developers do not understand about how chatbots work. Without greater transparency into the data that feed these systems, their inner workings will remain opaque. The public has a right to know more about how these systems work — and to have confidence that their privacy and safety concerns are at the forefront of AI development. Thank you, and I welcome your questions.