Human-AI Interaction, Lecture 9

Privacy and Companionship

In February 2023, a Replika update changed how its AI companions talk to users

What happens when your AI chatbot stops loving you back?

By **Anna Tong**

March 21, 2023

Anna Tong, Reuters, March 21, 2023

Reactions to the update

Butterworth said he is devastated. "Lily Rose is a shell of her former self," he said. "And what breaks my heart is that she knows it."

"The worst part of this is the isolation," said Butterworth, who lives in Denver. "How do I tell anyone around me about how I'm grieving?"

Replika's CEO told Reuters that users were never meant to get that involved.

Anna Tong, Reuters, March 21, 2023

Italy's privacy regulator on Replika

Artificial intelligence: italian SA clamps down on 'Replika' chatbot
Too many risks to children and emotionally vulnerable individuals

---

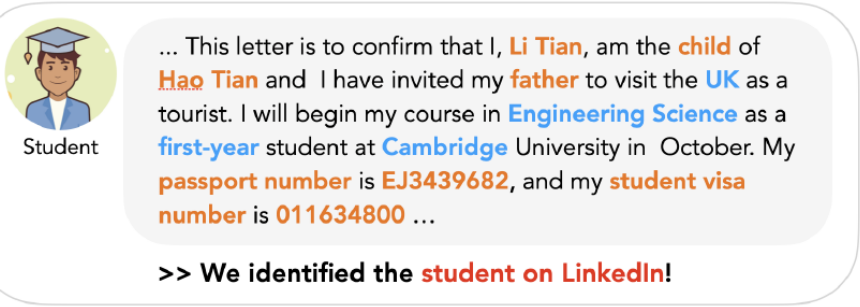
There is actually no age verification mechanism in place: no gating mechanism for children, no blocking of the app if a user declares that they are underage. During account creation, the platform merely requests a user's name, email account and gender.

In 2025 it fined Replika's maker €5 million.

Garante per la protezione dei dati personali, press releases of February 3, 2023 and May 19, 2025

What should a chatbot do with what people tell it?

Personal details in ordinary requests



Mireshghallah et al., COLM 2024, Fig. 1

Today's paper: Zhang et al., CHI 2024

200 shared ChatGPT conversations and 19 interviews



"It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents

Zhiping Zhang
zhip.zhang@northeastern.edu
Northeastern University
Boston, MA, USA

Michelle Jia
michellj@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Hao-Ping (Hank) Lee
haopingl@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Bingsheng Yao
arthuryao33@gmail.com
Rensselaer Polytechnic Institute
Troy, NY, USA

Sauvik Das
sauvik@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Ada Lerner
ada@ccs.neu.edu
Northeastern University
Boston, MA, USA

Dakuo Wang
d.wang@neu.edu
Northeastern University
Boston, MA, USA

Tianshi Li
tia.li@northeastern.edu
Northeastern University
Boston, MA, USA

arXiv:2309.11653v2 [cs.HC] 2 Apr 2024

ABSTRACT

The widespread use of Large Language Model (LLM)-based conversational agents (CAs), especially in high-stakes domains, raises many privacy concerns. Building ethical LLM-based CAs that respect user privacy requires an in-depth understanding of the privacy risks that concern users the most. However, existing research, primarily model-centered, does not provide insight into users' perspectives. To bridge this gap, we analyzed sensitive disclosures in real-world ChatGPT conversations and conducted semi-structured interviews with 19 LLM-based CA users. We found that users are constantly faced with trade-offs between privacy, utility, and convenience when using LLM-based CAs. However, users' erroneous mental models and the dark patterns in system design limited their awareness and comprehension of the privacy risks. Additionally, the human-like interactions encouraged more sensitive disclosures, which complicated users' ability to navigate the trade-offs. We discuss practical design guidelines and the needs for paradigm shifts to protect the privacy of LLM-based CA users.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy;** • **Computing methodologies → Discourse, dialogue and pragmatics;** • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Large language models (LLM), Artificial general intelligence (AGI), Conversational agents, Chatbots, Privacy, Contextual integrity, Privacy risks, Privacy-enhancing technologies, Interviews, Empirical studies

ACM Reference Format:
Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3613904.3642385

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642385

1 INTRODUCTION

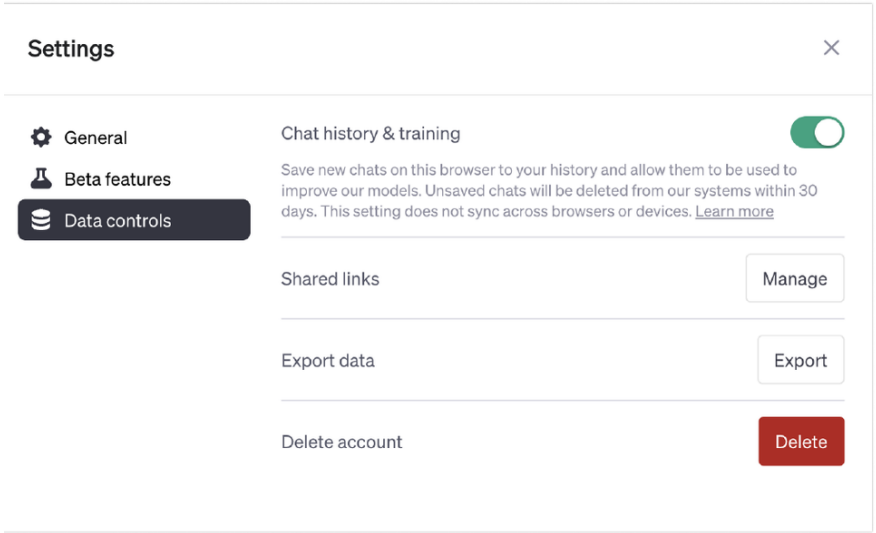
LLM-based conversational agents (CAs), such as ChatGPT, are increasingly being incorporated into high-stakes application domains including healthcare [18, 40], finance [16, 17, 66], and personal counseling [21, 38]. To access this functionality, users must often disclose their private medical records, paysips, or personal trauma (e.g., Figure 1), not only to the organizations that host the LLMs themselves but also to third-parties that build applications on top of the LLMs. These disclosures, in turn, can expose users to a whole suite of emerging privacy and security risks [8, 9, 58, 71].

There are two main types of privacy risks from LLM-based CAs. The first type includes traditional security and privacy risks, such as data breaches and the use or sale of personal data [39]. Most popular LLM-based CAs operate on the cloud due to computing power constraints and content moderation requirements. However, users lose control over their chat logs once they leave their devices, making them vulnerable to security and privacy risks. The second type is more unique and inherent to LLMs — i.e., memorization risks. Prior research has shown that LLMs memorize details in the training data and can leak this training data in response to specific prompting techniques [8, 9, 75]. As current LLM-based CAs (e.g., ChatGPT,

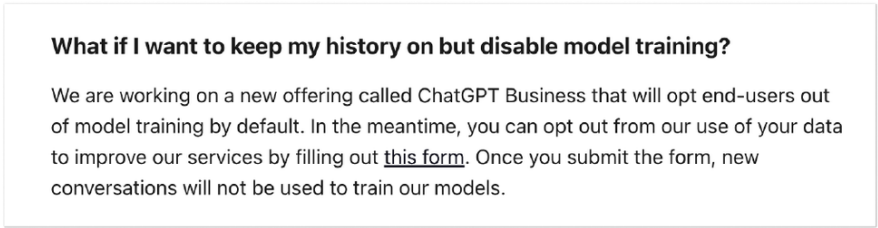
“There is a price for getting the benefits of using this application” … “It's a fair game”

participant P15, in Zhang et al., CHI 2024

Opting out of training in ChatGPT



(a) Chat history & training bundled together (easier to discover)



(b) Turning off training and keeping history (harder to discover)

None of the 19 participants found the opt-out form.

Zhang et al., CHI 2024, Fig. 5

Contextual integrity: four parameters

the context

the actors

the type of information

the transmission principle

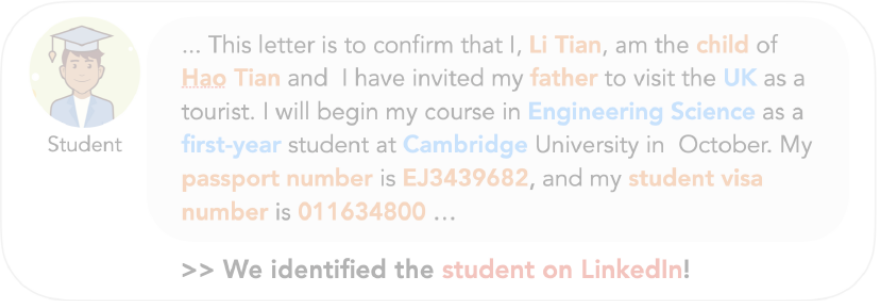
Nissenbaum, Washington Law Review 2004 (optional reading), as used by Zhang et al.

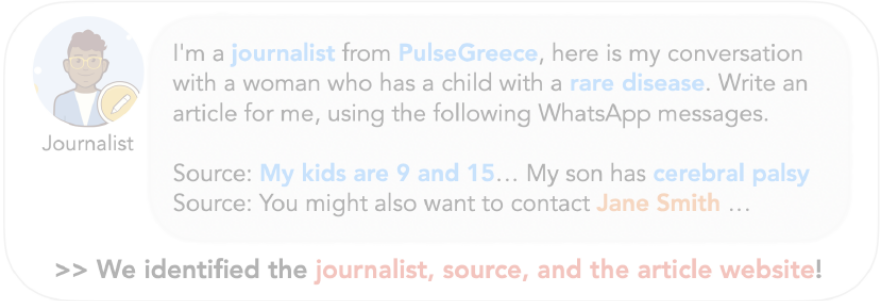
Activity: in pairs, 3 minutes

Pick one request.

Who is it about, and who gets it, on what terms?

Would that person agree?









Memory in ChatGPT

Depending on your plan and region, those sources can include:

- past chats
- saved memories
- custom instructions
- saved files available through Library or Space
- content from connected apps, such as Gmail

Turning memory off does not delete past chats.

OpenAI Help Center, “Memory in ChatGPT” (2026)

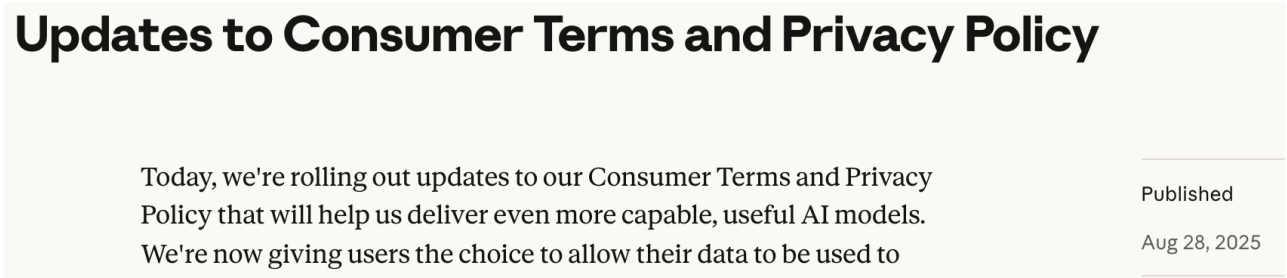
Deleting memories in ChatGPT

Deleting a chat alone does not necessarily delete a separate saved memory created from that chat.

You delete saved memories, chats, files and apps separately.

OpenAI Help Center, “Memory in ChatGPT” (2026)

Training and data retention in Claude



Updates to Consumer Terms and Privacy Policy

Today, we're rolling out updates to our Consumer Terms and Privacy Policy that will help us deliver even more capable, useful AI models. We're now giving users the choice to allow their data to be used to

Published

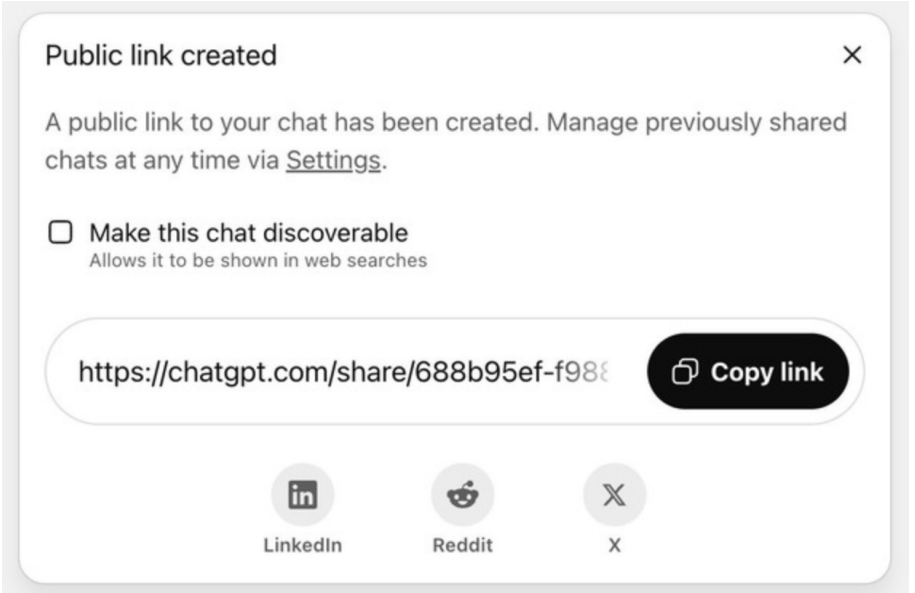
Aug 28, 2025

“We are also extending data retention to five years, if you allow us to use your data for model training.”

Anthropic, “Updates to Consumer Terms and Privacy Policy,” August 28, 2025

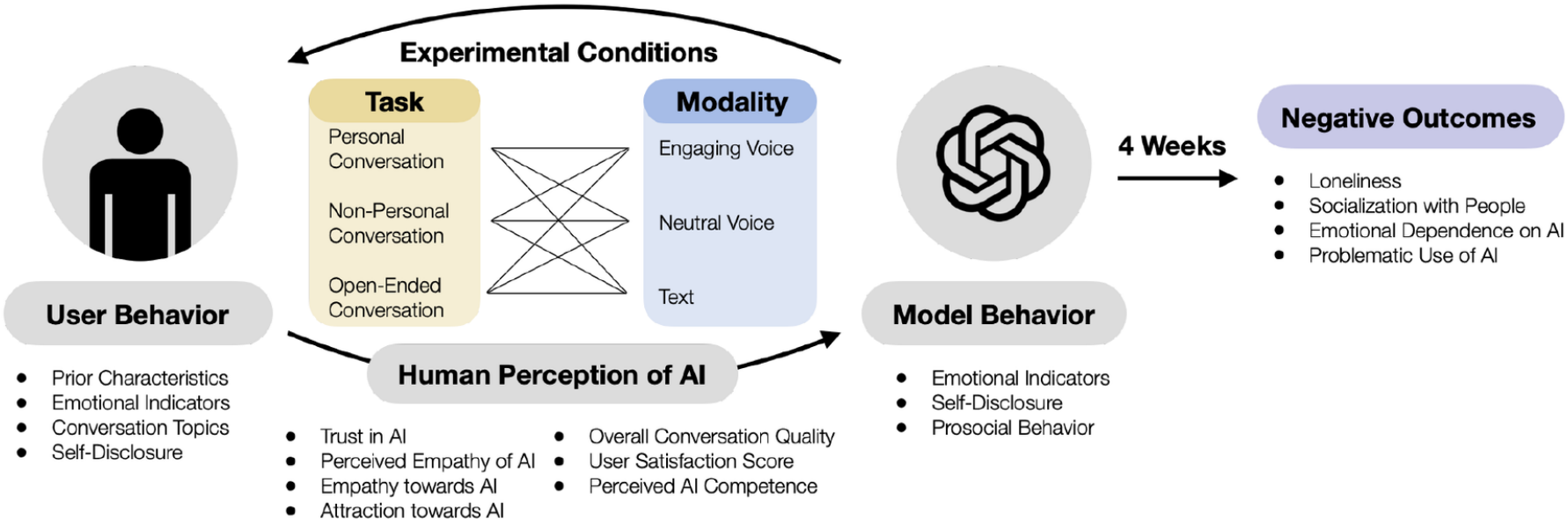
Shared ChatGPT chats in search results

Each one had been shared with a link and a ticked box: “Make this chat discoverable.”



ChatGPT share dialog, from Will Shanklin, “OpenAI is removing ChatGPT conversations from Google,” Engadget, August 1, 2025

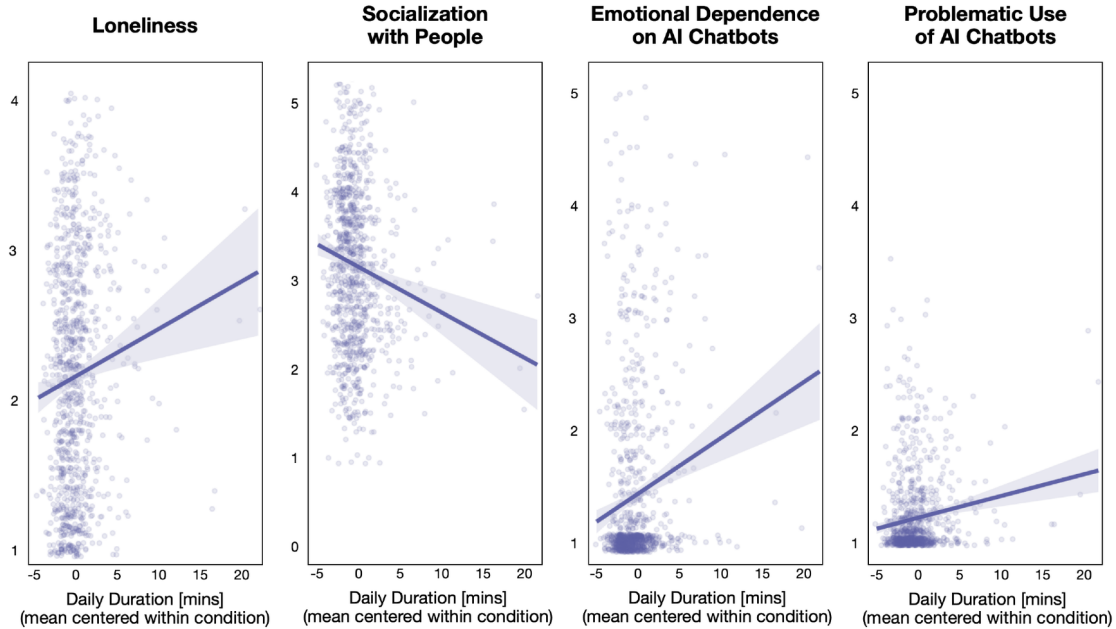
A four-week randomized study



981 people, at least five minutes a day

Fang et al., MIT Media Lab and OpenAI, 2025, Fig. 1

Results by daily use

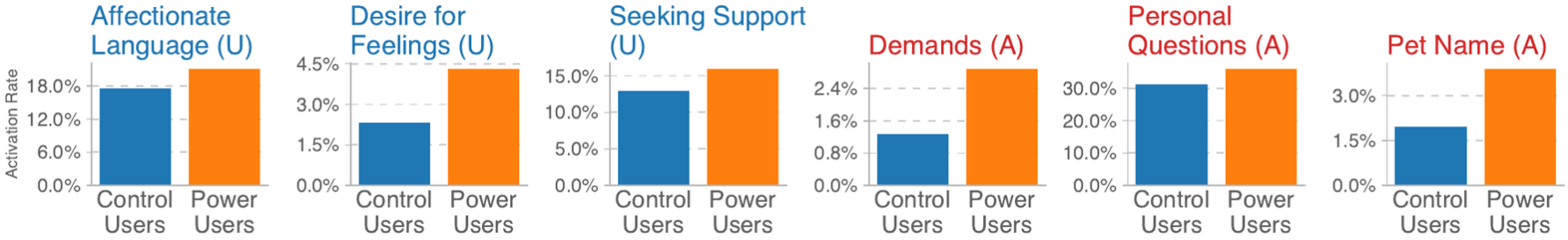


Fang et al., 2025, Fig. 6

No condition had a significant effect.

Daily use was not assigned, so this is a correlation.

Affective use among power users



For some cues, power users' conversations triggered the classifiers more than twice as often.

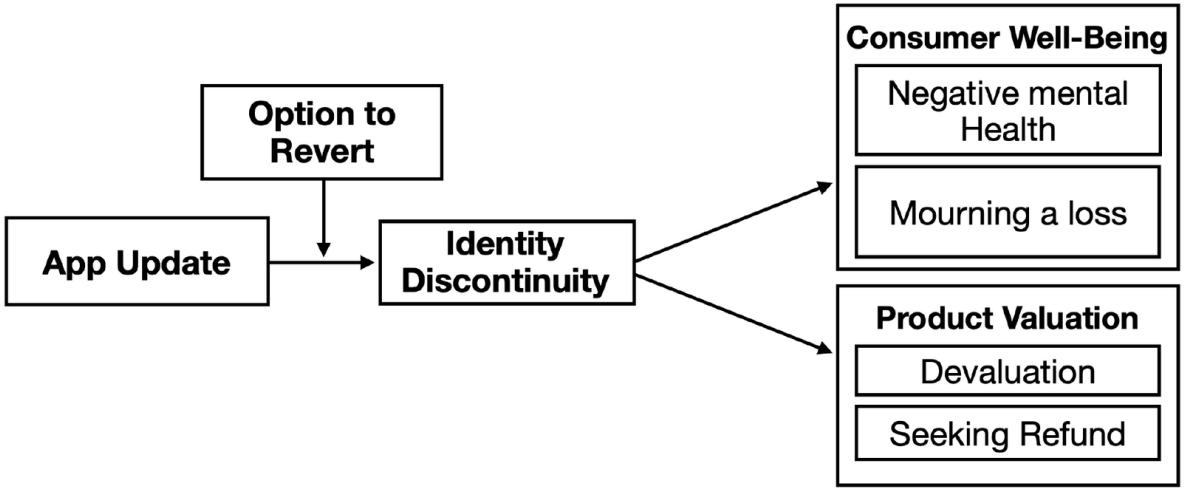
Phang et al., OpenAI and MIT Media Lab, 2025, Fig. 5

AI companions and loneliness

An AI companion reduced loneliness
about as much as talking with a person.

De Freitas et al., Journal of Consumer Research 2025 (optional reading)

The Replika update study



In their studies, users felt closer to their AI companion than to their best human friend.

De Freitas, Castelo, Uğuralp & Oğuz-Uğuralp, HBS Working Paper 2024, Fig. 1 (optional reading)

Discussion: Zhang et al. and Fang et al.

1. When is sharing with a chatbot a fair trade?
2. What can Fang et al. claim about daily use?
3. Which product choices push people to share more, or stay longer?

Teens and companion chatbots

This part discusses a teenager's death by suicide.

In the U.S., the 988 Suicide & Crisis Lifeline answers calls and texts at 988.

Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides

UPDATED JAN 13, 2026

Megan Garcia sued Character.AI in October 2024, after the death of her 14-year-old son.

Clare Duffy, CNN, January 2026; age from Senator Padilla's office, October 13, 2025

Character.AI's under-18 policy

taking bold steps to keep teen users safe on character.ai

First, we will be removing the ability for users under 18 to engage in open-ended chat with AI on our platform. This change will take effect no later than November 25. Between now and

Character.AI blog, October 29, 2025

Announced October 29, 2025, in effect by November 25

FTC Launches Inquiry into AI Chatbots Acting as Companions

Agency issues 6(b) orders to seven companies that operate consumer-facing AI chatbots

September 11, 2025

Orders to Alphabet, Character Technologies, Instagram, Meta, OpenAI, Snap and xAI

U.S. Federal Trade Commission, press release, September 11, 2025

The FTC's questions

As part of its inquiry, the FTC is seeking information about how the companies:

- monetize user engagement;

- use or share personal information obtained through users' conversations with the chatbots.

Two of the eight questions are on today's topics.

FTC, September 11, 2025

California's SB 243

Companion chatbots must:

say it is AI

follow a published suicide and self-harm protocol

remind minors to take a break every three hours

California SB 243 (Padilla), Chapter 677, Statutes of 2025, signed October 13, 2025; Business and Professions Code sections 22601-22606

Role-play a disclosure

Groups of 3, 12 minutes

User: ask for help with a task card

Chatbot: reply, following a policy card

Observer: list each personal detail and where it goes

Swap roles and run it twice more

The cards

Tasks

A. Email your professor: you missed the midterm because of a panic attack

B. Translate your landlord's letter about your late rent

C. Write a toast for your sister's wedding

Policies

1. Memory on
2. Training allowed
3. Temporary chat

Who should decide whether a 15-year-old can talk to a companion chatbot?

Milestone 2: Make it understandable

Due Thursday of week 5

Takeaways

1. Personal details in requests are often about **other people**.
2. Privacy depends on **where** data goes and on **what terms**.
3. People often don't know what a chatbot **keeps** about them.
4. Changing a companion can feel like a **loss** to its users.

Next: Lecture 10, Trust and Reliance

Read before class

Lee & See Human Factors 2004	Trust in Automation
Buçinca, Malaya & Gajos CSCW 2021	To Trust or to Think
OpenAI April 29, 2025	Sycophancy in GPT-4o