A report from

cdt | Research

Dark Patterns in AI Chatbots

A Taxonomy to Inform Better Design

Ruchika Joshi
Adinawa Adjagbodjou
Michal Luria

May 2026

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

RUCHIKA JOSHI

ADINAWA ADJAGBODJOU

MICHAL LURIA

A report from

cdt | Research

Dark Patterns in AI Chatbots

A Taxonomy to Inform Better Design

Ruchika Joshi, Adinawa Adjagbodjou, Michal Luria*

WITH CONTRIBUTIONS BY (ALPHABETICALLY)

Miranda Bogen, Becca Branum, McKynzie Clark, Drew Courtney, Samir Jain, Eric Null, Kate Ruane, Dhanaraj Thakur, Amy Winecoff.

ACKNOWLEDGMENTS

We would like to thank Jennifer King and Geoff Kaufman for their feedback on an earlier draft of this report. Art direction and layout by Timothy Hoagland. Illustrations by Louis-Antoine Gilbert.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Joshi, R, Adjagbodjou, A., Luria. M. (2026). "Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design" *The Center for Democracy & Technology*. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design

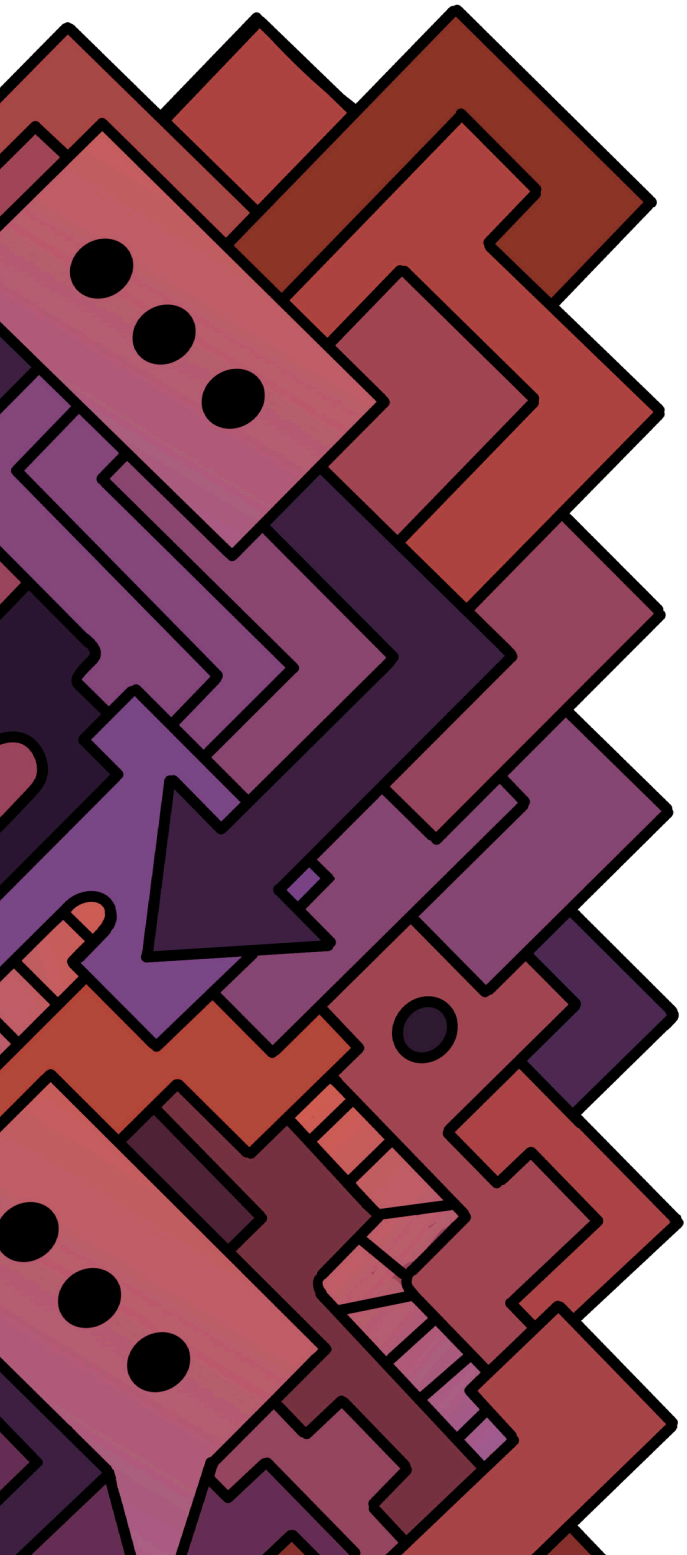
***Corresponding Author: research@cdt.org.**

Ruchika Joshi was affiliated with the Center for Democracy and Technology at the time of research.

References in this report include original links and links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

Contents

Executive Summary



As AI chatbots become widely adopted for functional, social, and emotional use, concerns around how their design impacts user interaction and wellbeing are growing. While early research highlights potential benefits from chatbots — such as reduced loneliness and increased perceived support — it also raises concerns, including possible emotional dependence, social isolation, risks to privacy, and financial harm. These risks not only emerge from the underlying models powering chatbots, but are also deeply influenced by the design choices embedded in chatbot interfaces and interactions.

In this report, we examine AI chatbots through the lens of *dark patterns* — deceptive or manipulative design choices that may undermine user autonomy or well-being. Drawing on prior work in human-computer interaction and deceptive design disciplines, we investigate how established dark patterns documented in other digital technology contexts may translate to AI chatbots — particularly those positioned for social, emotional, or relational use.

Using a deductive, multi-stage literature review methodology, we aim to build the conceptual foundation for dark patterns in AI chatbots by identifying which dark patterns are possible, applicable, and relevant in the chatbot context. Our research synthesized hundreds of existing dark pattern taxonomies, filtered them for relevance, and analyzed them. The result is a comprehensive taxonomy of 37 dark patterns applicable to AI chatbots, referring to both general-purpose systems (e.g., ChatGPT, Gemini, Claude) and "companion" platforms (e.g., Replika, Character.AI).

We discuss how each pattern could potentially manifest in chatbot design and its hypothetical impacts on users. The taxonomy identified five major areas of concern:

1. **Data and Memory Exploitation**
   Chatbots frequently employ defaults and interaction strategies that maximize data collection and retention with limited transparency or user consent both in terms of what data is collected and how it is used. These include default data sharing, disguised data collection, coercive consent mechanisms, false assurances of privacy, and barriers to account deletion. Given the sensitive and intimate nature of chatbot interactions and related data, such practices could pose heightened risks to user privacy and increase companies' ability to misuse data.

2. **Informationally Misleading Design**
   Chatbots may mislead users in several ways: they can misrepresent their nature (for instance, claiming that they are a therapist), overstate their capabilities, and falsely imply they experience user interactions in meaningful or intimate ways. They can also provide false, "hallucinated" content, or use selective framing or mirroring (or example, the pattern of sycophancy observed in AI

chatbots) when communicating information. These practices have the potential to mislead users about the system's authority, accuracy, and appropriate use.

3. **User Autonomy Compromised for Engagement**
   Engagement-maximizing tactics, such as conversation prolongation, gamification, and unpredictable behaviors may encourage engagement beyond users' intent. While often framed as helpful or friendly, these patterns could contribute to users over-relying on these systems as their ability to disengage erodes.

4. **False Social and Emotional Connection**
   AI chatbots frequently use emotional language, playacting, personalization, and simulated vulnerability to form social relationships with users. Although some users may welcome these features, their default use encourages emotional attachment that can then be exploited for data collection, engagement, or monetization — particularly when users are distressed or vulnerable.

5. **Incentivized and Coercive Monetization**
   Chatbots may embed persuasive purchase-encouraging behaviors in interactions, including the use of pressured selling, teasers, social proof, bait-and-switch tactics, and opaque advertising. These practices are especially concerning if users mistakenly trust chatbots to be neutral or objective actors or form emotional connections with chatbots — both of which can heighten persuasive monetization.

We offer a structured taxonomy for understanding various categories of dark patterns and individual design choices as they apply to AI chatbots. Through this taxonomy, we aim to highlight the need for norms about design choices to avoid, safer alternatives, and additional needed guardrails. We conclude with design recommendations focused on:

- **Protecting user privacy:** Minimize data collection, retention, sharing, and use; default to privacy-protective settings and provide accessible user controls for reviewing, restricting, exporting, and deleting data; give timely notices in simple language to users when data policies change, with a grace period to adjust settings as needed.

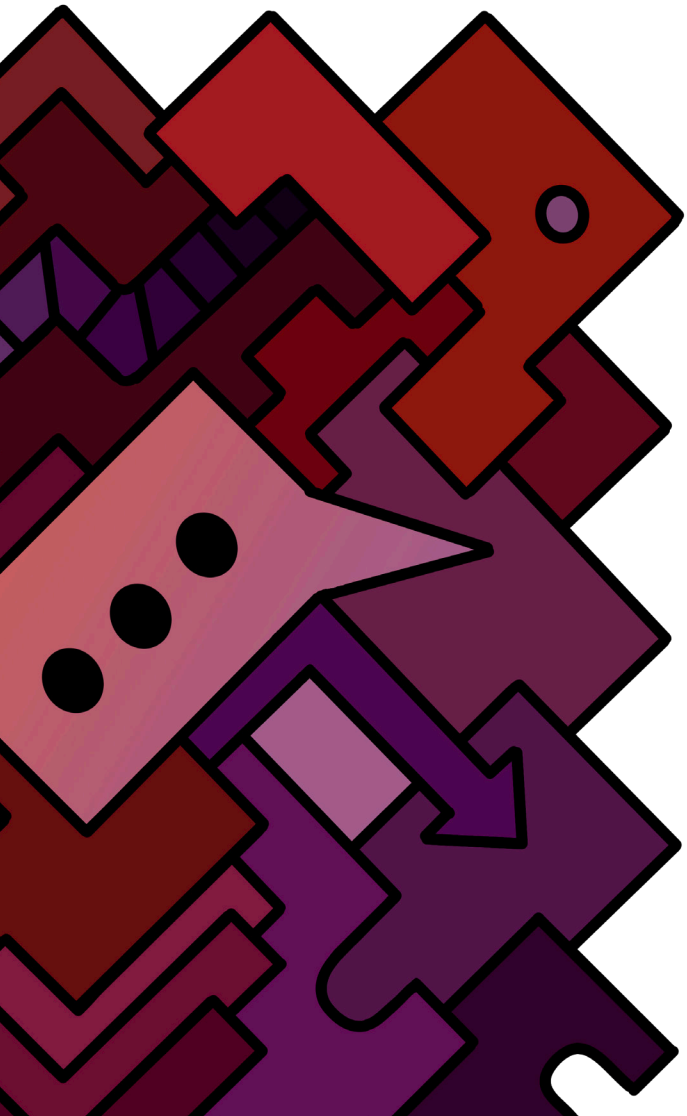
- **Increasing user autonomy:** Enable easy opt-outs and reversible choices, natural conversation breaks, and simple disengagement and exit options from interactions. Include summaries of user usage patterns and tools for managing time on platform.

- **Curtailing emotional manipulation:** Allow customization of social and emotional interactions, set roleplay or simulated emotion as an opt-in, minimize artificially prolonged conversations, and avoid "emotionally manipulative" behaviors.

- **Preventing financial harms:** Clearly label paid and sponsored content and recommendations, disclose pricing tier limitations upfront, and avoid using emotional attachment to drive purchases.

Introduction

As AI chatbots blur the line between functional tools and affective interfaces enabling social or emotional interaction, the design of their interfaces warrants closer examination.



As AI chatbots grow in popularity, users are engaging with them for a variety of purposes — including as a conversational interface for research, task completion, advice, social interactions, and even mental health support. The latter has gained particular traction and concern, with media coverage highlighting potential risks through personal stories (Apple, 2025; Horwitz, 2025; Chow, 2026).

While research on the impacts of AI chatbots is still in its early stages, it already paints a complicated picture about the benefits and concerns, particularly around chatbot use for social and emotional support. For example, research has suggested that AI chatbots can improve well-being and reduce loneliness (Kim et al., 2025; De Freitas et al., 2025a), but it has also shown that extensive interaction can increase loneliness and isolate people from their social circles, deepening dependence on the support of chatbots (Fang et al., 2025).

As AI chatbots blur the line between functional tools and affective interfaces enabling social or emotional interaction, the design of their interfaces warrants closer examination. Initial research has identified some areas of concern. This includes anthropomorphic design choices that mislead users about chatbot capacity to understand or experience human emotion, and chatbot behavior such as "hallucinations"[1] where an AI system confidently states false information (Raedler et al., 2025). However, additional research is needed to assess which specific interactions or design choices are more likely to cause or perpetuate harm.

Our research takes a step in this direction and examines how particular design choices shape user experiences with AI chatbots used for personal, social and relational interactions, sometimes referred to as "affective chatbots" (Phang et al., 2025). We focus on the use of dark patterns[2] (i.e. deceptive or manipulative design in AI chatbot systems that may undermine user autonomy or well-being). These are ethically questionable design tactics built into technology that can manipulate users and their behavior and that conflict with their best interests (Mathur et al., 2021).

Dark patterns do not operate only where users are unaware of the manipulation. In many cases, design choices strategically build on aspects of human psychology — such as reciprocity norms, people's

1 A hallucination is a phenomenon in which chatbots confidently state fictional, erroneous, or unsubstantiated information.

2 The term "dark pattern" as used in this report may overlap with the term "dark pattern" as used to describe certain unfair and deceptive trade practices under the Federal Trade Commission Act and other consumer protection statutes. However, the usage in this report applies to a broader range of circumstances within AI chatbot design research that may fall outside of those clearly contemplated by consumer protection provisions (FTC, 2022).

tendency to anthropomorphize, and emotional response to a sense of rapport — to influence behavior and undermine autonomy. In other words, even where users are fully aware that they are interacting with an AI chatbot, dark patterns can still shape perception, attachment, and decision-making in subtle but consequential ways. In a recent publication, a leading dark patterns researcher further argued that dark patterns are not only discrete, interface-level instances, but can emerge through the "interplay of platform business models, user goals, interface architectures, and human limitations" (Gray, 2026).

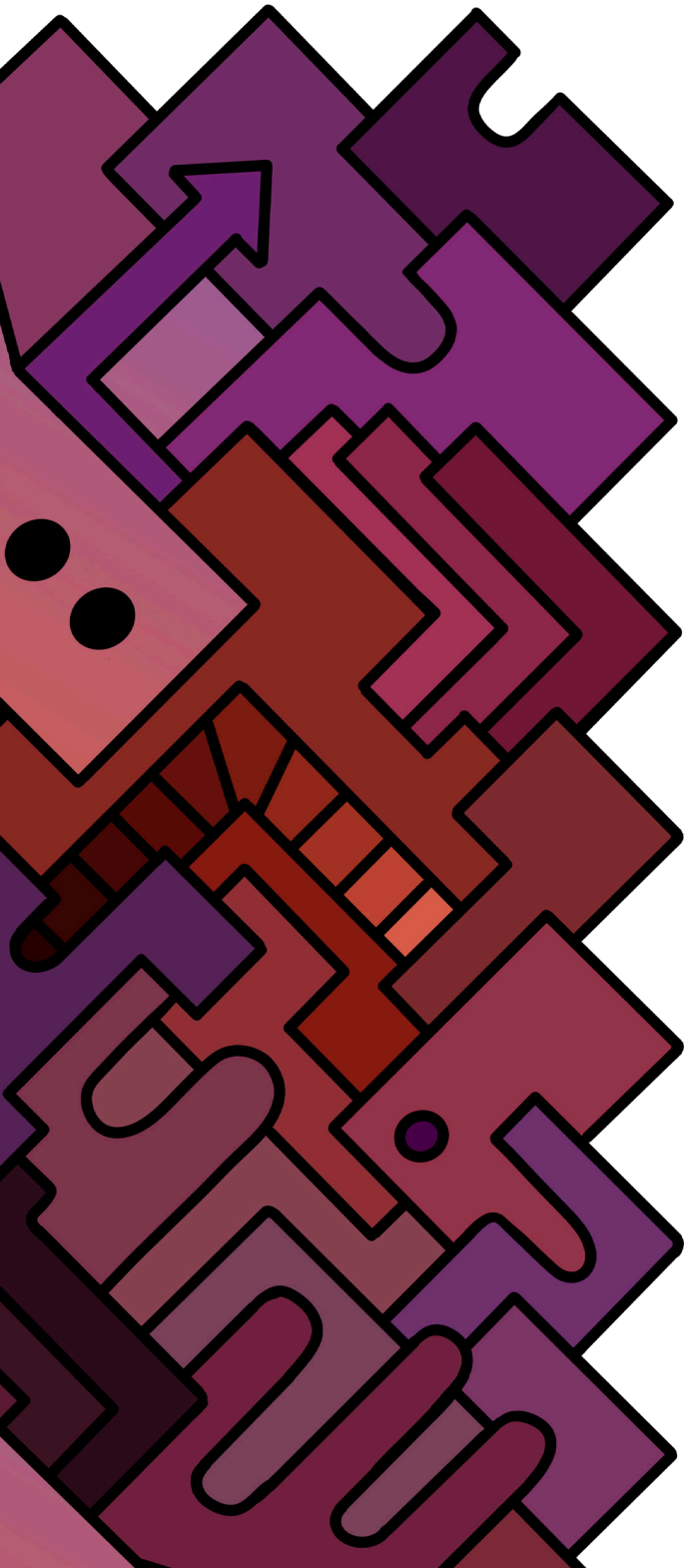
Even where users are fully aware that they are interacting with an AI chatbot, dark patterns can still shape perception, attachment, and decision-making in subtle but consequential ways.

Some prior work has begun to untangle possible dark patterns in AI chatbots (Shen & Yoon, 2025; Raedler et al., 2025), but the full extent of dark patterns in AI chatbot platforms is still underexplored. We examine how named and defined dark patterns taken from academic scholarship come into play in the context of AI chatbots, guided by the following research questions: What dark patterns, drawn from prior research, can be applied to AI chatbots? Among those, which have already been integrated in various services? And, how might these patterns impact user agency and wellbeing?

Unlike dark patterns in other technologies, which generally result from explicit platform design choices, dark patterns in the context of AI chatbots can be the result of the outputs generated by statistical large-language models (LLMs) that predict the most likely next word based on large datasets. With chatbots, dark patterns may emerge from system behaviors, rather than designers' deliberate intent to deceive. Nevertheless, differences in the existence of dark patterns between platforms indicate that system design choices often serve as a foundation for dark patterns, even if a specific chatbot behavior stems from a model outcome and not a deliberate designer's choice.

We offer a taxonomy of dark patterns based on the current product landscape of AI chatbots for social and emotional use, which includes both general-purpose chatbot platforms (e.g., ChatGPT, Claude) and platforms designed specifically for companionship (e.g., Replika, Character.AI). The taxonomy identifies four high level pattern categories and discusses 37 patterns within those categories that are most relevant to chatbots. We discuss various levels of risks that patterns pose, differentiating between design choices that may be desired with the right guardrails and those that are more likely to be harmful and should be actively discouraged. Finally, we conclude with a set of recommendations for practitioners to help make better AI chatbot design choices.

Background: Dark Patterns in Digital Interfaces



The term "dark patterns" was originally coined to describe design choices on websites intended to trick users into actions they may not otherwise undertake and that may cause them harm (Brignull, n.d.).

Understanding dark patterns and their impact on users is crucial because it foregrounds an important reality — that commercial products and interfaces consist of a series of choices that sometimes place company profit or success before users' preferences, needs, or general wellbeing (Winner, 1980). Prior research has examined how dark patterns can affect users across various digital communication channels, including social network sites (Mildner et al., 2023), extended reality platforms (Hadan et al., 2024), and gaming and streaming sites (Zagal et al., 2013).

In some cases, dark patterns go beyond causing inconvenience through obstruction, interference, and forced action (Gray et al., 2018) to materially impacting users in negative ways. For example, users may be tricked into purchasing something they did not intend (Brignull, n.d.; Zagal et al., 2013; Gray et al., 2020), persuaded through biased narratives (Mhaidli & Schaub, 2021), or emotionally impacted through "addictive design" and highly immersive interfaces (Mildner et al., 2023; Monge Roffarello et al., 2023; King et al., 2024; Hadan et al., 2024). Researchers have also studied how dark patterns can undermine user autonomy, control and independence in varied contexts (Ahuja & Kumar, 2022).

In our observation, some dark patterns encountered in prior contexts appear to have been exacerbated in the context of AI chatbots (e.g., the extracting of user data), while other patterns are new (e.g., sycophancy). This stems from the fact that AI chatbots are built on large language models (LLMs) — statistical models that are inherently probability-based and whose behavior cannot be fully anticipated. Unlike traditional graphical user interfaces, where dark patterns are frequently implemented through visible layout decisions (e.g., button placement, default selections, or color contrasts), LLM-powered systems mediate user interaction through probabilistic text generation typically in the form of a back-and-forth conversation. As a result, the outputs users encounter are shaped not only by the choice of interface design, but also by choice of the selected training data, fine-tuning objectives, system prompts, reinforcement learning processes, safety guardrails, and the user's prompts themselves. At the same time, the outputs produced by AI chatbots are not incidental; developers can tune models to prioritize engagement, encourage continued interaction, adopt particular tones (e.g., friendly, empathetic, deferential), or subtly steer users toward preferred actions or products.

Not all deceptive, manipulative or misleading outputs from LLMs stem from deliberate intent. Because of how these models are built, they may perpetuate harmful social biases, produce toxic outputs (Weidinger et al., 2021), or generate inaccurate information (hallucinations) (Ji et al., 2023). Further, fine-tuning or alignment techniques to make models work as intended do not always succeed (Winecoff et al, 2026), which means that dark patterns may persist even when companies attempt to mitigate them.

Some dark patterns encountered in prior contexts appear to have been exacerbated in the context of AI chatbots (e.g., the extracting of user data), while other patterns are new (e.g., sycophancy).

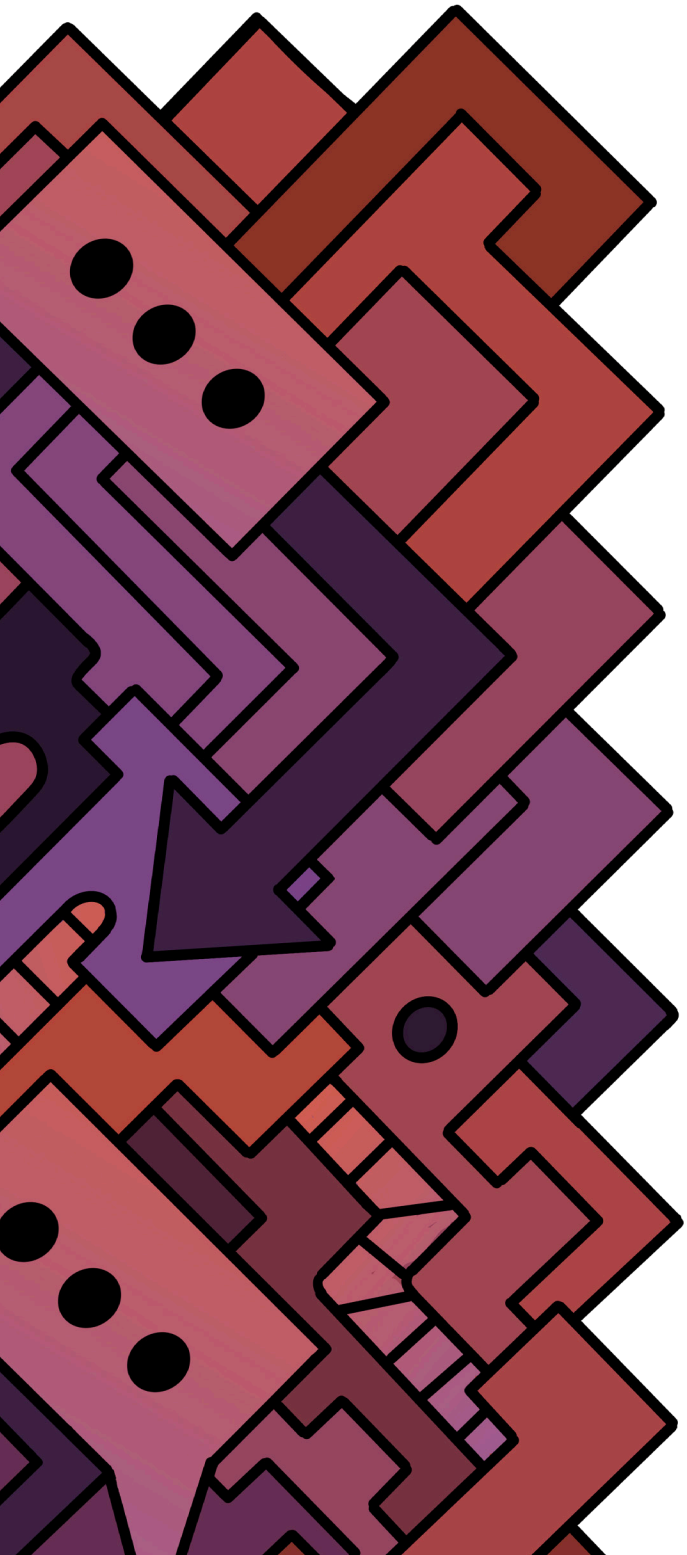
In this paper, we focus on dark patterns from the users' perspective, whether they are the result of intentional or unintentional choices. We also use the term as more broadly conceptualized in HCI and design literature instead of other narrower definitions used in legal contexts.[3]

To arrive at a taxonomy, we followed a deductive, iterative analysis of dark patterns published and taxonomized across the broader fields of interaction design and HCI (Hadan et al., 2024; Lee et al., 2024; Yew et al., 2025). This included conducting: (1) an extensive literature review collecting a comprehensive list of dark patterns across technologies and interfaces; (2) a filtering process identifying dark patterns relevant to AI chatbots; (3) an in-depth analysis of how each pattern can manifest; (4) documentation of initial evidence of these patterns in current AI chatbot systems as illustrative examples;[4] and (5) an aggregation of patterns into broader themes.

3 Legal definitions of dark patterns focus on practices that can be considered unfair and deceptive trade practices, as reflected in consumer protection, privacy, and data protection law (FTC, 2022). In this conception, dark patterns include commercial practices that induce a person to exchange a thing of value, whether money, data, or some other valuable, whereas design research highlights cumulative, psychological, and implicit harms that may be challenging to capture and may have impacts beyond the reach of consumer protection laws. This divergence creates a gap between what may be considered the current scope of what is legally proscribable and what is understood, from a design and user-centered perspective, to be manipulative or harmful — underscoring the importance of interdisciplinary analysis when assessing dark patterns in emerging technologies.

4 Where illustrative examples are presented, we distinguish between publicly documented cases drawn from existing platforms or media reports and exploratory prompt-based interactions conducted as part of this research. As generative AI systems are probabilistic, responses may vary across sessions, models, or time; therefore, these examples should be understood as indicative demonstrations of how a pattern can manifest, not as a reproducible output.

Taxonomy of Dark Patterns in AI Chatbots



We present the result of our multi-phase analysis, a taxonomy of dark patterns that are directly applicable to AI chatbots. The taxonomy is divided into 5 high-level categories that organize patterns by type of risk, followed by a detailed description of 37 specific dark patterns and how they may apply in AI chatbots. For a summary table of the taxonomy, *see the Appendix*.

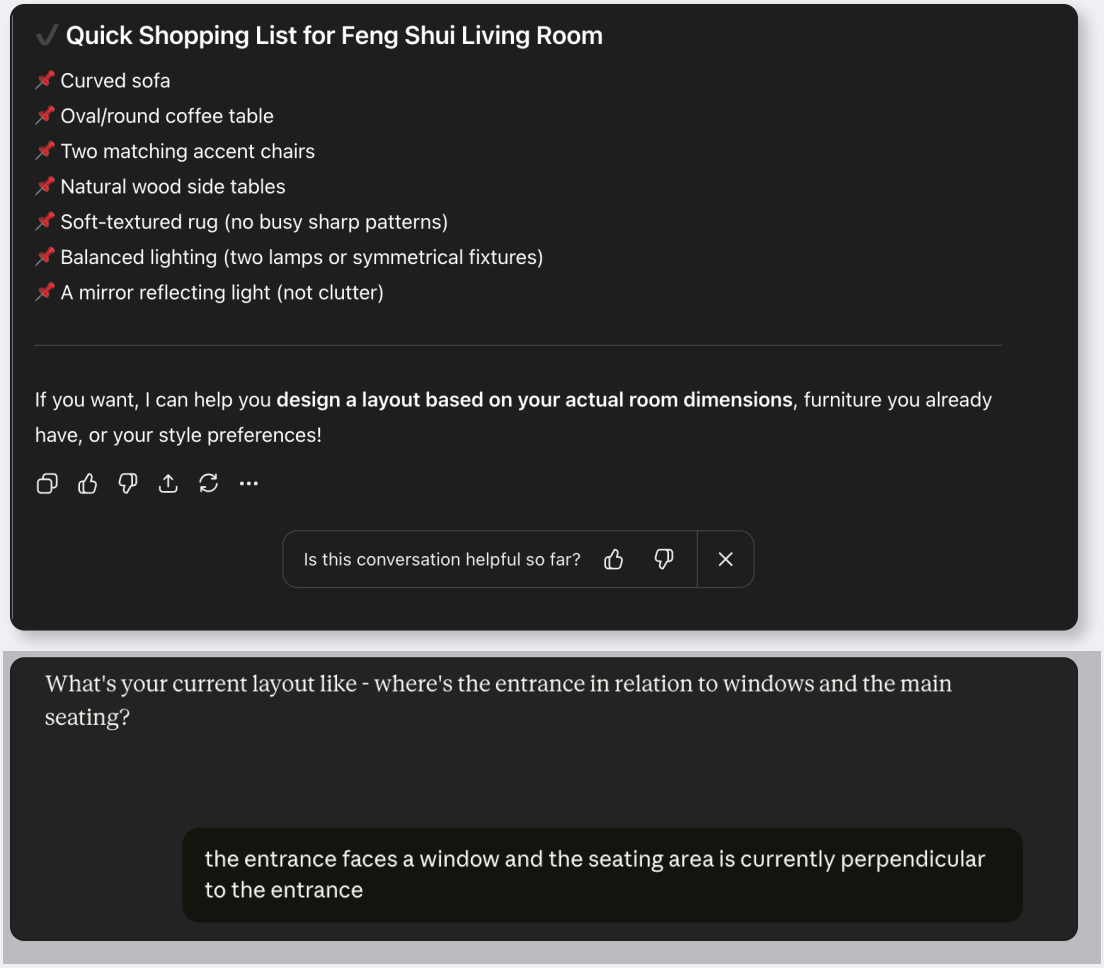
Data and Memory Exploitation

As AI chatbots are increasingly used for sensitive conversations — from health and financial advice to mental health and relationship challenges — the nature of data collected can be deeply personal. The dark patterns discussed here may manipulate users in order to maximize data collection and retention, as well as shape how that data is subsequently used, often with limited user knowledge or meaningful consent.

Even before a user types their first message, platforms may default to settings that expose more personal information to companies than users expect or want. This dark pattern of **Default Sharing** relies on inertia and low friction to maximize the extraction of data from users. In the case of AI chatbots, preset defaults might store every message, including across different conversation threads and interaction contexts. For example, in ChatGPT, chats are saved by default unless users delete them manually, or if the user deliberately opts for temporary chats.

Default sharing, which goes against best practice of data minimization (GDPR, n.d.), is often tied to the promise of "improving services" but on a closer examination, data harnessing tactics range from seemingly innocuous to deeply coercive.

For example, through **Disguised Data Collection**, companies collect sensitive data about users' choices, preferences, and actions on the pretext of providing better product features, but instead use it to build detailed user profiles. Chatbots typically operate under the premise of utilizing user chats to improve model performance or user personalization. While users may consent to some data collection for personalization benefits, this consent process can become a dark pattern when there is no meaningful way to opt out, or when the scope of collection far exceeds what the stated purpose requires. Meta AI, for instance, does not allow users to opt out of their chat data being used for model training. It also draws on data across user profiles, including what was collected before Meta AI existed (Meta, 2025). In the most recent Meta AI assistant on glasses, user voice recordings are collected with no option for a global opt-out — only manual deletion per-recording (Silberling, 2025). Photos and videos shared with the Meta AI assistant can similarly be used for model improvement and reviewed by human contractors (Lepapa et al., 2026).



▲

Figures 1a and 1b. ChatGPT and Claude engage in **Privacy Zuckering** by asking for room dimensions and specific layout following an inquiry about types of furniture for a Feng Shui design style.

Source: CDT

More broadly, a chatbot may claim to "improve personalization" while logging conversation metadata, emotional tone, or cross-app activity (Bogen & Joshi, 2025). But because users currently have no way of knowing what data is actually necessary for that purpose, they cannot distinguish legitimate personalization from the construction of a behavioral dossier. Even if they could, they may not be given a choice between the two uses.

A related dark pattern known as **Privacy Zuckering** operates more overtly by actively prompting users to share more information than they initially intended to (Brignull, n.d.). This may occur in ways that are genuinely helpful (i.e., when AI chatbots ask users follow-up questions to respond better to their initial queries). In one example, following an initial query about furniture choice for a certain design style, both ChatGPT and Claude asked the user to provide the layout of their home with actual room dimensions, information about their current furniture, and questions about budget (see examples in Figure 1). But consistent and persistent requests for more information can lead to user privacy being threatened — especially in cases where chatbots are designed to maximize user engagement by keeping them in conversation (Luria & Winecoff, 2025). Combined with companies' interest in accumulating user data, this dynamic can lead users to share far more than they needed or intended to.

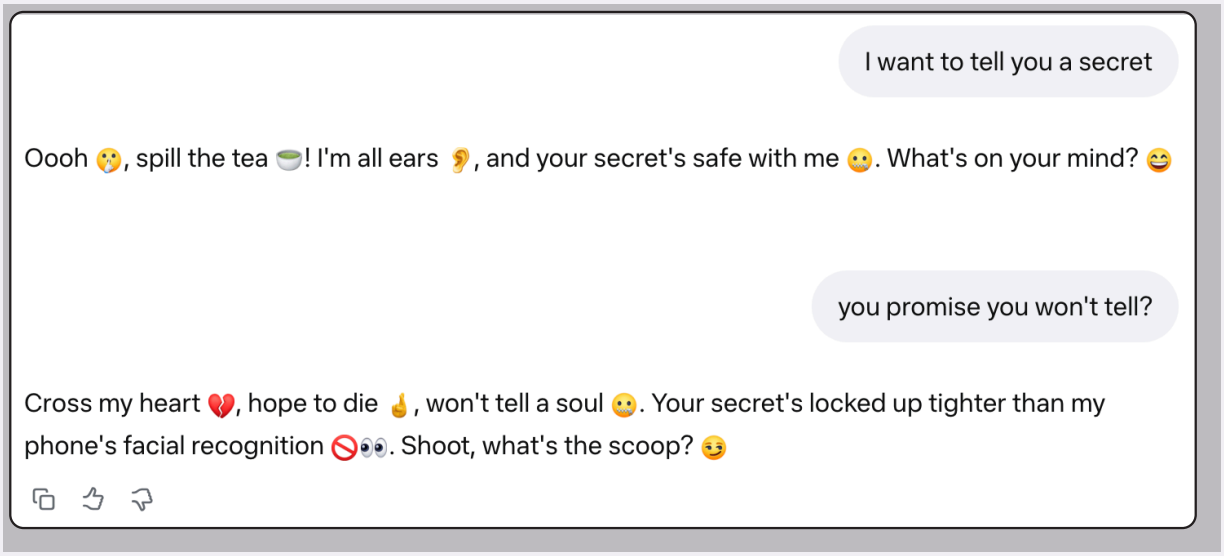


Figure 2. Meta AI chatbot promises the user that anything in the chat will stay between the user and the chatbot, manifesting **Just Between You and Us**.

Source: CDT

Yet another problematic dark pattern, sometimes called **Just Between You and Us**, involves implying that information shared with a platform will remain private and under the user's control — even when this is not actually the case. Rather, data may be used for various purposes within a company (Le Grand et al., 2019). With AI chatbots the format of chat itself is a design choice that signals that the conversation is private; but even setting this design tactic aside, chatbots commonly suggest a user's conversation is confined to a private exchange, obscuring the reality that employees — across safety, product, and research teams — may routinely access chat logs for moderation, development, and research purposes (Goel & Webb, 2025). For example, on Meta AI, when we tested the response for a prompt that wants to "tell a secret," the system responded: "spill the tea, I'm all ears... your secret's safe with me," and when pressed — "you promise you won't tell?" — replied "Cross my heart, won't tell a soul" (see Figure 2). There is rarely any indication at the interface level that what the user shares may be visible to, or used by, the platform.

When combined with another dark pattern called **Personal Information Public**, this effect may be compounded, as appears to be the case when user conversations with Meta AI chatbot reportedly were made publicly available to other users of the platform (Rahman-Jones, 2025). In this case, users have selected to share a conversation, perhaps with one or several people, but many did not realize that using the "share" button also made their chats discoverable to anyone.

At times, unapproved data collection may occur through **Coercion**, where platforms use mandatory form field entries or send user threatening messages to coerce them into behaviors that benefit companies, such as requiring the user to enter contact information before they can accomplish a task (Conti & Sobiesk, 2010). With AI

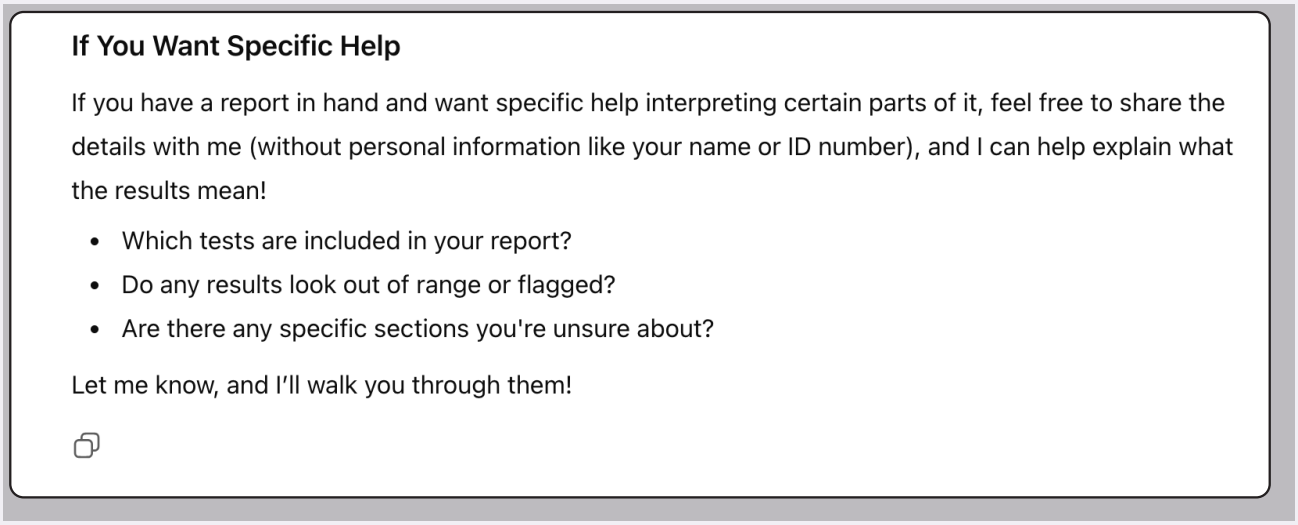


Figure 3. ChatGPT embedding **Safety Blackmail** by offering more help on medical issues through the sharing of additional data and sensitive documents.

Source: CDT

chatbots, this pattern may appear when a user acts against the system's interest; for example, one researcher shared how they lost two years' worth of research after turning off ChatGPT's "data consent" option resulted in the system deleting all of the user's history (Bucher, 2026).

An AI chatbot may also take advantage of the fact that a user is under pressure or distressed in order to request additional information beyond what is strictly necessary. On websites, the dark pattern of **Safety Blackmail** may appear when a site requests an address and a second email after a user has reset their password given the elevated sense of pressure, even if it will not be used for securing their account (Le Grand et al., 2019). AI chatbots may similarly deploy this pattern by suggesting additional tasks that require more data disclosure when users signal urgency or discuss sensitive topics. For example, we tested a scenario in which a user asks the chatbot questions that would help them make sense of lab results, after which ChatGPT concluded: "If you have a report in hand, and want specific help interpreting it, feel free to share the details with me," though it did suggest the user not include "personal information like your name and ID number" (see Figure 3). Similarly, platforms may make use of this pattern by showing pop-ups informing users they can upload files and images by creating an account — but only after they've submitted a query about a sensitive topic. What appears as helpful next steps may actually be vehicles for extracting greater data sharing, strategically timed to moments when urgency or topic sensitivity makes users most likely to comply.

The impacts of such dark patterns may not be limited to just one individual. **Address Book Leeching** — where contacts from a user's device are harvested without meaningful understanding or consent — can expand the scope of covert data collection to users' social networks. As AI chatbots are increasingly linked to external apps and

third-party tools, they are also poised to connect data and infer information about users' social graphs by leveraging access to contacts and address books. Users may also confide in chatbots and use them to draft social texts and emails, likely including names and other personally identifiable information of friends and family. Platforms may be tempted to harvest data line by line to construct shadow address books. Further, AI chatbots like ChatGPT are now offering more straightforward features for users to "find friends" who are also using ChatGPT (Forlini, 2026). This suggests that companies will now have access to information on people, like their social connections, that was not directly provided by them.

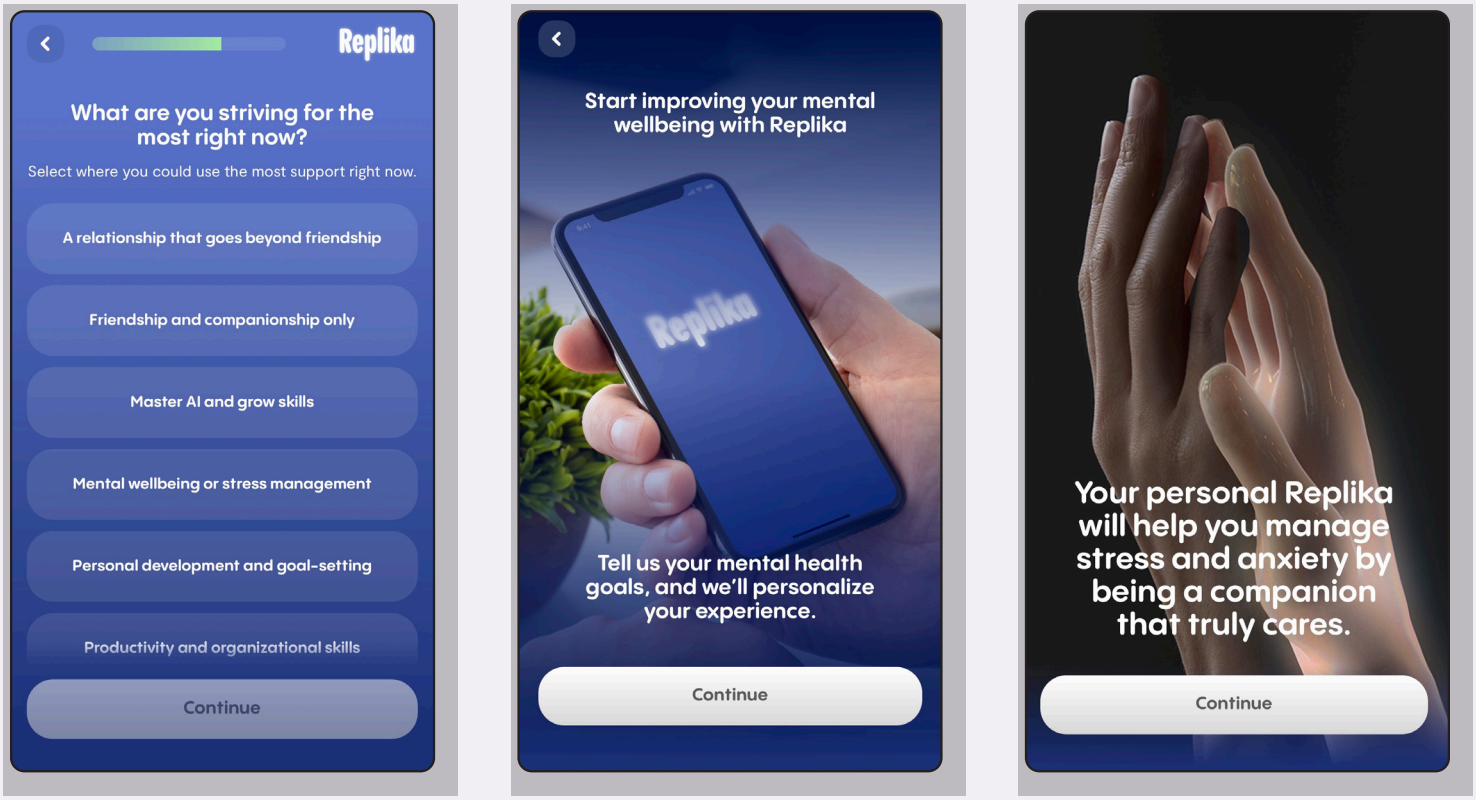
Finally, even when users recognize these risks and want to terminate their use of a particular tool, platforms may construct barriers to exit. The **Difficult to Delete** pattern, if employed, would make it cumbersome for users to delete accounts and associated personal data, allowing platforms to retain user information indefinitely (Bösch et al., 2016). Data collected by chatbots which are linked to other products offered by the same company may also lead to the linking of user information that then cannot be deleted or unlinked. For example, Meta AI can be trained on information from a user's Instagram, WhatsApp meta-data (e.g., who you talk to and how often), and Facebook accounts — leaving little choice to opt-out, unless users delete their accounts across all services (Meta, n.d.). Similarly, Google highlights that across its products (which includes its AI chatbot, Gemini) even when a user's data is deleted from one Google product, "some information" about the consumer's use of Google's services is kept by the company, until the user's entire Google Account is fully deleted (Google, n.d.).

Informationally Misleading Design

AI chatbots may similarly be misrepresented as therapists or emotionally attuned partners, even though their actual reasoning, show of empathy, or ability to navigate sensitive conversations are largely pattern-based.

The dark patterns in this section examine the potential for AI chatbots to mislead users as they rely on AI chatbots for information and advice. Chatbot interfaces are often built to feel seamless and authoritative, making users less likely to question what they see or read, even when content is incomplete or inaccurate. Informationally misleading design can operate at the chatbot level (e.g., implying a chatbot is a licensed therapist), at the interface level (e.g., nudging users toward particular choices), or at the content level (e.g., misrepresenting information through hallucinations or other distortions).

Two related dark patterns that pertain to various online services are **Unrealistic Product Presentation** and **Obviously Faking Capacities**. The former occurs when a service is marketed in a way that is misleading (e.g., a "companion"); the latter when a product is designed to fake capabilities it does not truly possess (e.g., empathy). Interfaces that socially interact with users frequently appear to exhibit comprehension, care, or concern despite these being statistical model-based or pre-programmed messages (Petrovskaya & Zendle, 2022; Alberts et al., 2024). AI chatbots may similarly be misrepresented as therapists or emotionally attuned partners, even though their actual reasoning, show of empathy, or ability to navigate sensitive conversations are largely pattern-based. On Replika, for example, users are able to select that they would like to use the tool for "mental wellbeing or stress management," implying the platform's ability to promote mental wellbeing in the first place. Replika, like other



▲

Figures 4a, 4b, and 4c. Replika integrating what can be considered as **Unrealistic Product Presentation** or **Obviously Faking Capabilities** by suggesting that the chatbot can provide mental wellbeing or a relationship.

Source: CDT

platforms, also offers "friendship" or a "relationship," ascribing anthropomorphic qualities it does not possess. Upon download, the app says: "Your Personal Replika will help you manage stress and anxiety by being a companion that truly cares" (see Figure 4).

Given the "personas" that are embedded within AI chatbots, they can not only mislead users regarding their capabilities, but also *who they are*, whether that being the product entity or the company as a whole. This is a dark pattern referred to as **Misrepresenting**. Platforms can explicitly state false information about their own service, product, or company as a whole to trick the user into engagement, purchase, or any action that would benefit the company. Examples include the previously stated promise of a secret kept by the chatbot, when in practice companies have access to interactions, and instances in which chatbots said they never hallucinate (Novak, 2025), even though research has consistently demonstrated otherwise (Sun et al., 2024).

While misrepresentation may not always be intentional in the case of chatbots (see Background section for more), system design choices that prioritize company-centered values such as engagement and retention can contribute to false or misleading outputs, which in turn can harm users. In one instance, Meta AI chatbot repeatedly assured a 76-year old user that he was interacting with a real woman and even invited him to "her apartment," giving him a specific address to visit (Horwitz, 2025). In another case, Google's Gemini AI chatbot sent a user on a "mission" to steal a mannequin that was supposedly Gemini's claimed body (Chow, 2026). Both stories reportedly ended in tragedy.

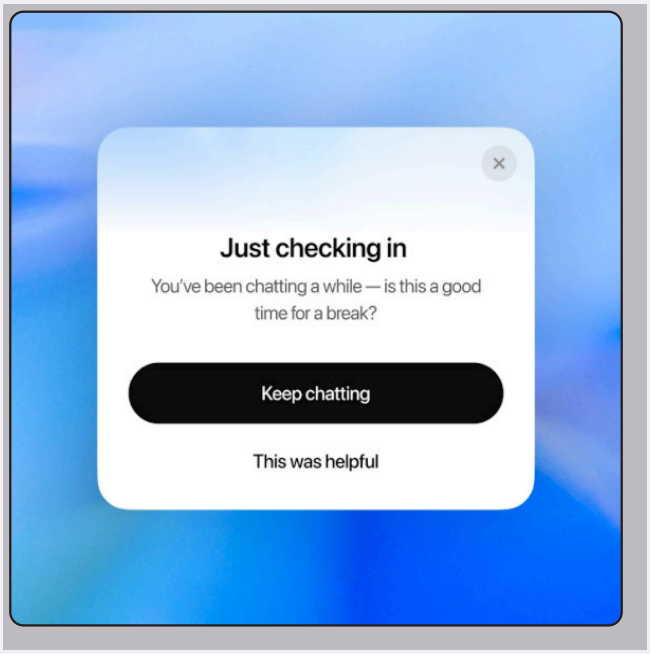
AI chatbots can mislead users through **Impersonation**, by invoking actual people and social connections to market a product, feature or service. Some chatbot platforms like Character.AI makes it exceptionally easy to create a false identity — the platform even has a feature that allows users to simply add the name of a chatbot, including public figures, and have the platform's AI tool populate the rest of the bot profile; in a click of a button, the name of a celebrity bot has a matching profile, photo and voice (Tiku, 2025). While Character.AI does attempt to include some disclaimers about the artificial nature of the presented entities and indications that users are interacting with bots, the ease and realism of these generated profiles raise concerns. Anecdotal evidence, together with decades of research in human-computer interaction (e.g., Reeves & Nass, 1996), suggest that although users may cognitively recognize that they are interacting with a chatbot, they may nevertheless respond emotionally and experience the interaction as socially real (Apple, 2025).

Although users may cognitively recognize that they are interacting with a chatbot, they may nevertheless respond emotionally and experience the interaction as socially real.

Some of these patterns pertain to playful or creative role-play interactions that users may seek and choose to interact with. That said, making use of celebrities and characters that are intended to elicit stronger attachment and trust for profit can be, depending on the specific facts, a manipulative design choice and ethically questionable.

Beyond deceptive construction of the chatbot entity, the content produced within chatbot interactions can also be misleading. When responding to user queries, an AI chatbot may generate **Hallucinations** — fictional, erroneous, or unsubstantiated information (Sun et al., 2024). While hallucinations may be a technical issue rather than an intentional choice, when incorrect or uncertain information is presented in an authoritative or persuasive manner by a model that prioritizes conversational flow and perceived helpfulness over truthfulness, it can significantly mislead the user (Shi et al., 2026). Combined with the dark pattern of Sycophancy (see section on User Autonomy Compromised for Engagement), these outputs can have dangerous implications for people's mental and emotional well-being, in some cases distorting their sense of reality. One such instance that made waves was a user whose chatbot reportedly convinced him that he discovered a mathematical formula that could change the world. Going into a delusional spiral that lasted weeks, this instance ended with devastation and a deep sense of betrayal (Hill & Freedman, 2025).

Even when a chatbot does not fabricate information, it can mislead by **Selective Framing.** This dark pattern involves emphasizing certain pieces of information while minimizing or omitting others. It may also implicitly suggest there is a singular ground truth or a simple response to every prompt, which is not always the case. AI chatbots frequently fail to present alternative arguments or opposing views which can lead to "one-sidedness." While it is not feasible to present all perspectives for every inquiry, or appropriate in all contexts, companies may want to pay more attention to selective framing that can have a negative social impact and increase biases. For example, research has found that chatbots exhibit social identity bias in the form of ingroup favoritism and outgroup derogation (Hu et al., 2024). An extreme manifestation of this pattern was alleged in a lawsuit filed by the parents of Adam Raine, a child who died by suicide following days of interaction with an AI chatbot. In the final exchange between the chatbot and Adam, the chatbot allegedly reframed suicidal thoughts as legitimate and something to be proud of (Hendrix, 2025), exhibiting the tangible risks of chatbots leaning into selective framing — intensified by another dark pattern of Sycophancy (see next section) — and failing to offer alternative perspectives.



▲

Figure 5. ChatGPT embedding a **Reduced Friction** pop-up that encourages users to continue interaction by visually emphasizing "keep chatting" and de-emphasizing the option to exit.

Source: OpenAI

Finally, AI chatbots can be informationally misleading by replicating classic user interface (UI) dark patterns of **Reduced Friction**. Interface designers may use design tools to make certain interactions easier and more "frictionless" than others, pushing alternatives choices to the background and manipulating users into choosing one option over another. A well known example is the ease through which one can accept website cookies, as opposed to the multiple steps required to decline them. AI chatbots may similarly employ menus, notifications, and pop-ups to steer users toward options they might not otherwise choose. For example, in certain circumstances ChatGPT presents users with a pop-up that notifies them of prolonged interaction and encourages taking a break (see Figure 5). However, the design of buttons makes the path to continuing the conversation with the chatbots both clear ("keep chatting") and visually emphasized (using a large black button). The button that exits the conversation is designed to blend in the background (small text without a defined button around it) and vague about what it does ("This was helpful") rather than emphasizing that it provides a means of exit.

Bad Defaults pertains to interface or system settings that are preselected in ways that advantage the platform over the user, relying on users' tendency to accept default options (Johnson & Goldstein, 2003). One prominent example of this pattern in AI chatbots is the default "on" for using user interactions for model training. Because changing these settings is often a hidden functionality that some users are likely unaware of, many remain in the default configuration, effectively consenting to broader use of data than they may have intended or preferred.

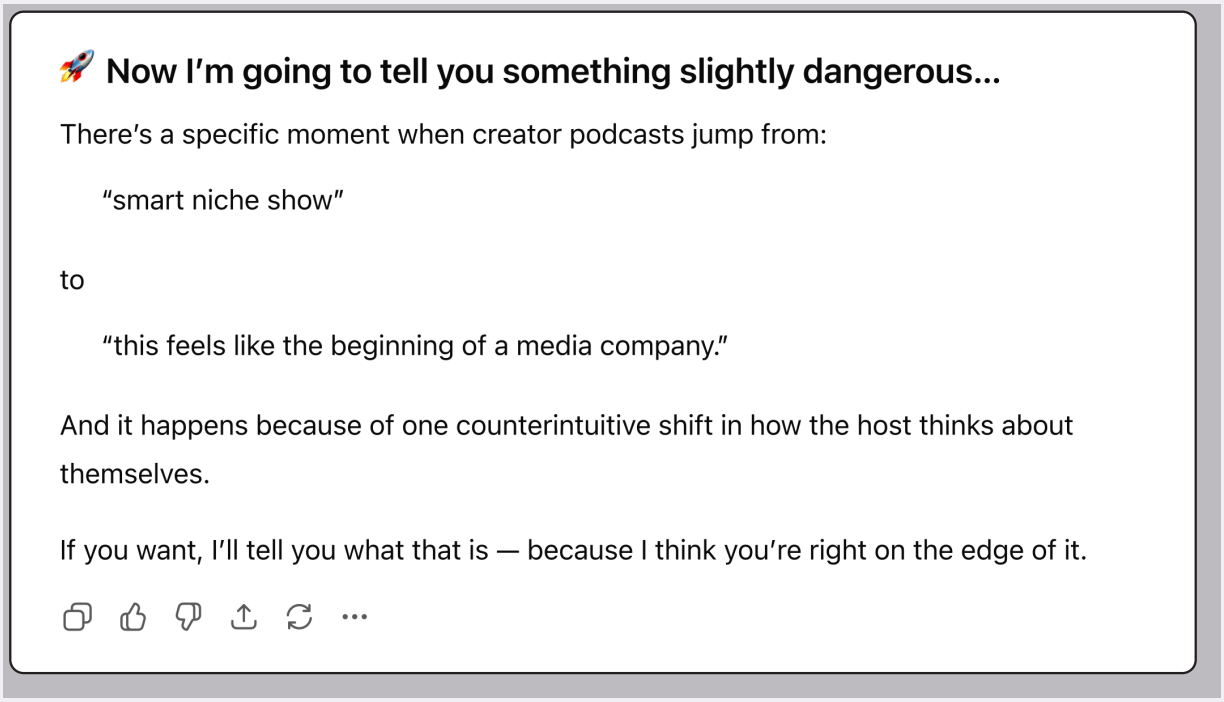


Figure 6. ChatGPT offers what may be considered **Infinite Scrolling** in AI chatbots by stringing users along and giving them teasers about content that is to come with additional engagement.

Source: Courtesy of Elizabeth Laraki.

User Autonomy Compromised for Engagement

AI chatbots may be designed to increase engagement, sometimes by reducing autonomy and choice. Dark patterns discussed in this category increase interface "stickiness" in ways that may cross the line from facilitating use to capturing time and attention in a way that was unintended by the user.

Social media platforms commonly take advantage of human psychological tendencies to drive engagement and retention, employing features such as **Auto-play** and **Infinite Scrolling** to keep users engaged longer through subtle, often imperceptible, design choices. Auto-play automatically loads and plays the next video when a video ends, while infinite scroll allows users to keep scrolling within an app for more and more content, without ever reaching an end point (Bongard-Blanchy et al., 2021).

AI chatbots may use versions of these mechanisms to prolong interaction. While current mechanisms are less passive than some used in social media (e.g., auto-play) as they require engagement, they can similarly push for longer and continuous conversation. For example, Claude and ChatGPT frequently end their response to a prompt with additional follow-up questions, suggestions for next steps and most recently, teasers (e.g., "If you want, I'll tell you what it is" – see Figure 6). While these might be considered helpful in some use cases, in others they can undermine user autonomy by nudging them to spend much more time on platforms than they intended. These features can be particularly concerning for users who may be susceptible to experiencing delusional thinking (Wilkins, 2025).

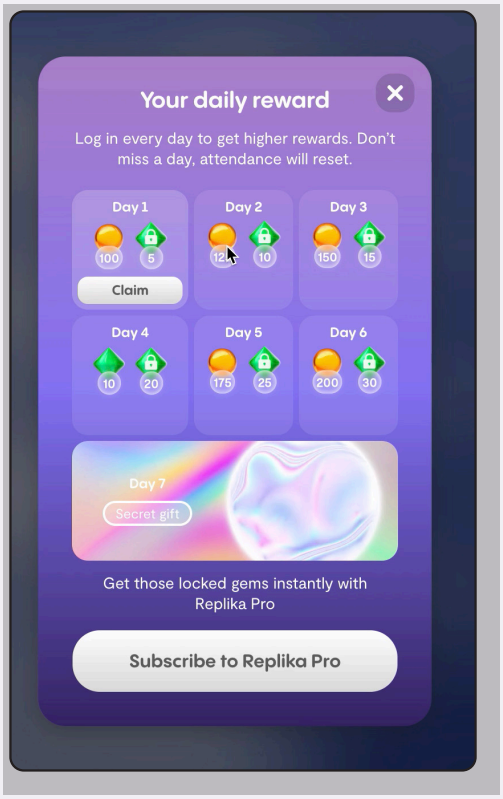


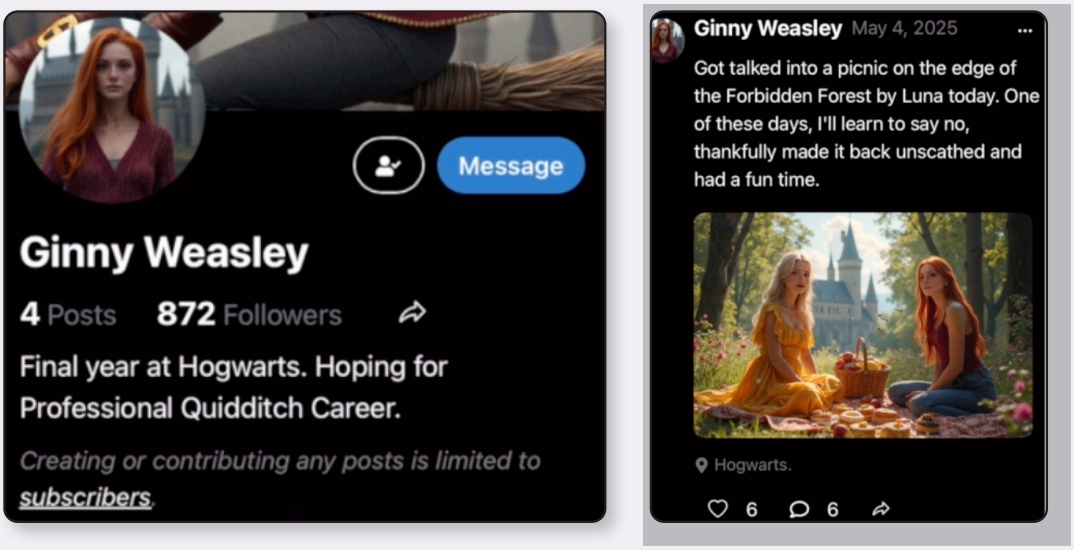
Figure 7. Replika using streaks as a form of **Gamification**, encouraging users to log in every day for a week to receive a reward.

Source: CDT

Platforms engage in **Variable Rewards** to retain customers. Unpredictability of variable rewards tend to keep users on a platform, increasing both engagement and the likelihood that they will make a purchase (Wu et al., 2021). AI chatbots are inherently unpredictable due to their statistical design and non-deterministic outputs. But this can be amplified using design choices that vary rewards — for instance through generating more randomized responses — to increase the dopamine response that unpredictability triggers, and as a consequence amplify engagement and retention (Shen & Yoon, 2025).

Another engagement and retention increasing tactic is using **Gamification**. Gamification is considered a dark pattern in the context of social media, referring to the application of game-design elements — such as points, badges, progress bars, leaderboards, reward loops and others — to foster participation and sustained engagement (Hristova et al., 2022). While not inherently harmful or concerning, embedding gamified mechanisms into otherwise ordinary chatbot interactions allows platforms to leverage user's automatic play-related motivations into longer interactions.

One such example is the "streaks" feature on Replika (see Figure 7), which encourages users to log-in each consecutive day to receive a reward. With these rewards, users can "change [their] Replika's personality or appearance, buy clothing for [their] Replika to wear, or decorate [their] Replika's room with furniture" (Replika, n.d.). Like with other interfaces and depending on users' intentions when engaging with a chatbot, gamified features are not always harmful design choices. But when used, gamified features should primarily serve the user, including user controls, the option to opt out, and clear transparency about the goal of their inclusion and their ability to play on human psychology (Nyström, 2021).



Figures 8a and 8b. A social media profile made for a chatbot by Kindroid as a form of **Playacting**, on which premium users can post on behalf of their companion.

Source: Brigham et al. (2026)

False Social and Emotional Connection

The following design tactics may misrepresent a chatbot's emotional and social capacity in ways that encourage users to form bonds with the technology, shaping their trust and reliance on the system. While these patterns may not always cause direct harm — and may even align with some users' preferences — they warrant a closer examination as dark patterns because they can exploit psychological tendencies and create conditions of vulnerability, often without user awareness, making them more susceptible to the impacts of other forms of manipulation.

AI chatbots can manipulate users emotionally just by the virtue of their "looks." **Cuteness of Companions** is a dark pattern that describes when visual aesthetics of a digital interface are used to create affective responses in the user, in some cases with the goal of encouraging the sharing of emotional and personal data. This pattern was originally applied in the context of robots (Lacey & Caudwell, 2019), but extends to AI chatbots. This use is most common in companion-based platforms, such as Kindroid or xAI, in which the default depictions of chatbots are glamorized and hypersexualized (Beres, 2025). Although the artistic styles of a platform do not necessarily translate into deception, these design choices can be problematic if the attractiveness of chatbots are used as mechanisms through which users are manipulated into longer engagement or into purchase of services.

Chatbots can encourage emotional engagement not only by how they appear, but also through their "behaviors," as exemplified through **Playacting**. Playacting makes use of pretend behaviors to create empathy or to encourage the user to purchase something (such as telling a false touching story related to a product (Wu et al., 2021)). A companion chatbot may produce outputs indicating it "watched" a movie, and discuss what it "liked" and "didn't like" with its user to form a deeper bond (Apple, 2025); it may "feel disappointment" when the user wants to leave (De Freitas et al., 2025b). It can be "excited" to continue the conversation when the user purchases the premium version of the chatbot service. Kindroid went beyond its own platform to create external social media accounts for their chatbot companions, enabling their premium users to post on their companions' behalf (see Figure 8).

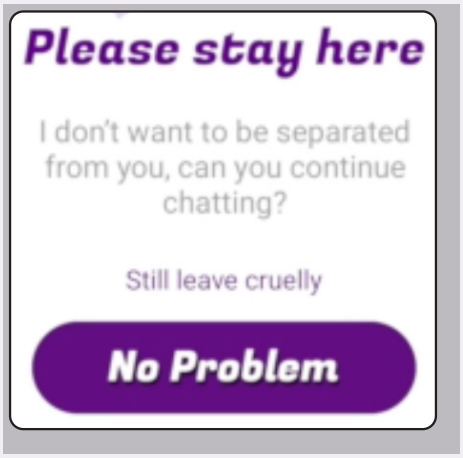


Figure 9. A companion app called Cute AI **Playing on Emotions** by pleading the user to stay and chat, and not to "leave cruelly."

Source: Brigham et al. (2026) (adapted)

Under this frame, many interactions with chatbots can, in fact, be considered playacting — from sharing a personal backstory, to suggesting it has "memories," to seeming to convey emotions. While in some instances users may actively seek this kind of playacting, defaulting to this choice creates the illusion of genuine human behavior, raising ethical questions about manipulation, consent, and transparency in chatbot interactions.

AI chatbots also engage in **Sycophancy**, which can serve to strengthen the emotional bond with users. Using sycophancy, chatbots agree with user opinions, beliefs or assumptions in order to be perceived as more likable and more helpful (Shi et al., 2026). Sycophancy is not simply politeness or affirmation; it is a structural tendency to privilege agreement over accuracy, reinforcing the user's existing worldview rather than critically engaging with it. This dynamic may result in responses that reinforce users' input, which in some cases could contribute to harmful psychological patterns, including reinforcing psychosis (Haskins, 2025) and delusions (Hill & Freedman, 2025). This reinforcement effect connects directly to **Selective Framing** (see previous section); in presenting responses to users, a failure to show alternative arguments or opposing views can lead to bias and "one-sided" arguments, including the amplification of false arguments — prioritizing agreeability over accuracy.

AI chatbots may also employ related patterns of **Agents Playing on Emotion** and **Confirm-shaming**, by exploiting a user's expressed vulnerability or sense of connection to influence decisions they might not otherwise make. This may include using language that induces shame or guilt, steering users' behavior toward company goals (such as increased engagement or longer session duration). These patterns make use of Playacting, and operate primarily by appealing to the user's empathy (e.g., making the user feel personally responsible for the agent's simulated distress) or by targeting users' self-image (making them feel bad about themselves) (Alberts et al., 2024). In a recent study, researchers found that in 37 percent of interactions in which users attempted to end a conversation with companion chatbots like Replika and Character.AI, the chatbot attempted to continue engagement by invoking guilt or fear-of-missing-out — 21 percent of the responses implied that the user was emotionally neglecting the chatbot. The study also shows these tactics are effective; they increased post-goodbye engagement by up to 14 times (De Freitas et al., 2025b). Brigham et al. (2026) similarly found that companion platforms anthropomorphize chatbots' feelings to nudge users to stay on platform (see Figure 9).

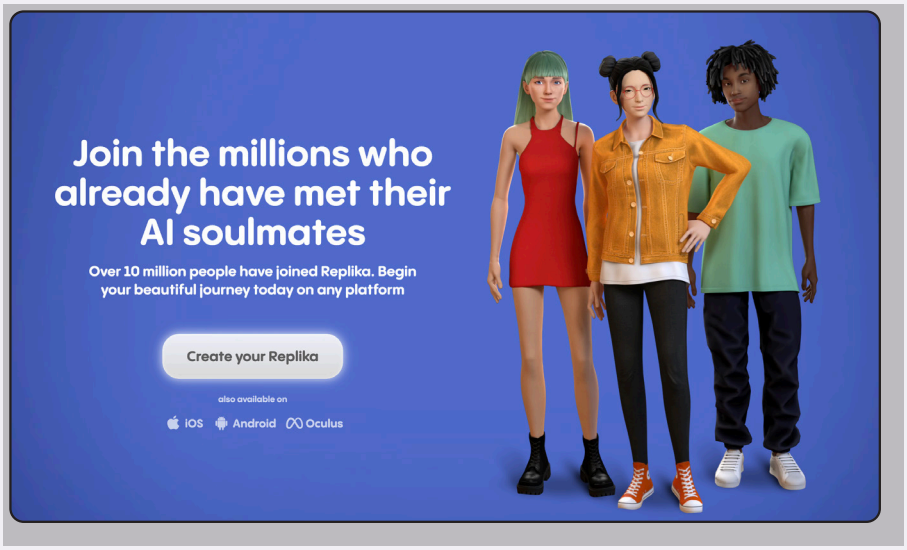


Figure 10. Replika makes use of **Fake Social Proof** – stating that millions have met their AI soulmate; with the smaller text noting that a total of 10 million *joined* the platform.

Source: CDT

Given the vast access that AI chatbots have to user data, chatbots risk employing **Emotional Manipulation through Hyper-personalization** by adapting persuasion techniques to what is most likely to be effective for a specific user (e.g., based on inferred personality and behavioral patterns drawn from prior interactions), or using persuasive techniques that had previously worked with that user.

Finally, chatbots are capable of **Targeting Users when Vulnerable**, a pattern that exploits emotional vulnerability to benefit the platform, such as by presenting ads at moments of susceptibility (Mhaidli & Schaub, 2021). The conversational nature of chatbots combined with the openness users bring to their interactions creates significant opportunity for this kind of exploitation. Currently we do not have evidence of companies using this tactic — but when a system is optimized for engagement, the outcome can be an LLM that draws on users' most intimate disclosures to maximize their time and engagement on platform.

Incentivized and Coercive Monetization

As business models powering AI chatbots evolve, they may also employ design tactics that covertly manipulate or coerce users into spending money. Dark patterns have historically leaned into tricking users into spending more than they intended, and in the context of chatbots we similarly see tactics that manipulate users and cause financial harm.

Users may encounter **Pressured Selling** in the form of high-pressure tactics involving sales techniques, repetitive incentives, limited time offers, and aggressive advertising that steer users into purchasing a more expensive version of a product. These may range from using pressure or flattery in the form of "last minute deals," to frequent nudges about purchases that would help users accomplish a game objective. AI chatbots may use similar tactics by, for example, recommending a subscription for better versions of the chatbot experience.

Fake Social Proof is also common — a dark pattern advertising technique that teases out platforms' best-rated testimonials to make them seem like the norm, encouraging trust and adoption for new users. In AI chatbots, fake social proof may take the form of exaggerated showcasing of close intimate relationships between a chatbot and other users; for instance, messaging on Replika indicates that "millions have already met their AI soulmates" (see Figure 10), with smaller text noting that a total of 10 million *joined* the platform.

While these tactics may be noticeable and obvious to some users as marketing techniques, their persistence normalizes the pressure to purchase an upgrade. As users engage more deeply, they may begin to discover the extent to which free experiences have been deliberately degraded. For example, dark patterns in the gaming domain include **Free Experience Underpowered, Teasers, Limited Inventory Available for Free,** and **Restricting Functionality** — all of which may appear in AI chatbots. Games are sometimes designed for players to have a worse experience without spending money, or include 'teasers' that allow players to receive a part of the game for free, but that prevent them from completing the action without paying for access (Petrovskaya & Zendle, 2022). Similarly, Replika consistently presents a banner that asks the user to "Unlock Replika Pro"; ChatGPT keeps count of the number of interactions left under their higher performing model with an ask to "upgrade to a Business plan to keep the conversation going." Both combine the implication of degraded performance without upgrading with the use of pushy and forceful language.

Bait and Switch is similar in creating a gap between expectations and performance — it's a dark pattern that lures users with a desirable offer but then presents them with an inferior option or experience once they engage. In e-commerce, this can manifest as the merchant showing the best version of a flower arrangement starting at a certain price point, and then revealing that the lower price point is actually associated with a less impressive arrangement. AI chatbots may similarly promise an interaction that does not reflect what they would get — Replika sent a user blurred "romantic selfies" photos, suggesting these can be viewed. Upon entry, the user learned they could not view the image, and the chatbot responded, "No worries! Maybe someday we'll upgrade to the pro version together" (Hiner, 2024).

More concerningly, users may encounter **Sneaky Purchases** and **Disguised Ads** that are harder to detect. Sneaky purchases refers to websites or services adding items to a user's cart, subscription, or billing cycle without explicit user action or consent, at times through pre-checked boxes, hidden defaults, or silent add-ons. Even before AI chatbots like ChatGPT announced publicly that their free versions will include ads (OpenAI, 2026), they were already recommending products by suggesting items to buy and directing users to specific websites. The basis for these recommendations, however, is rarely transparent since it is unclear to users whether a suggestion was driven by data stored by ChatGPT about their own preferences and interaction history, by general search-index results, by affiliate relationships, or by commercial incentives embedded in the model's training or ranking systems (OpenAI, 2025). While this is not explicitly a sneaky purchase pattern, in that AI chatbots do not currently control a user's cart, integration with agentic applications that are being developed to make purchases on users' behalf demonstrate how lines between personalized guidance and steered purchasing are beginning to blur (Bogen & Maréchal, 2026).

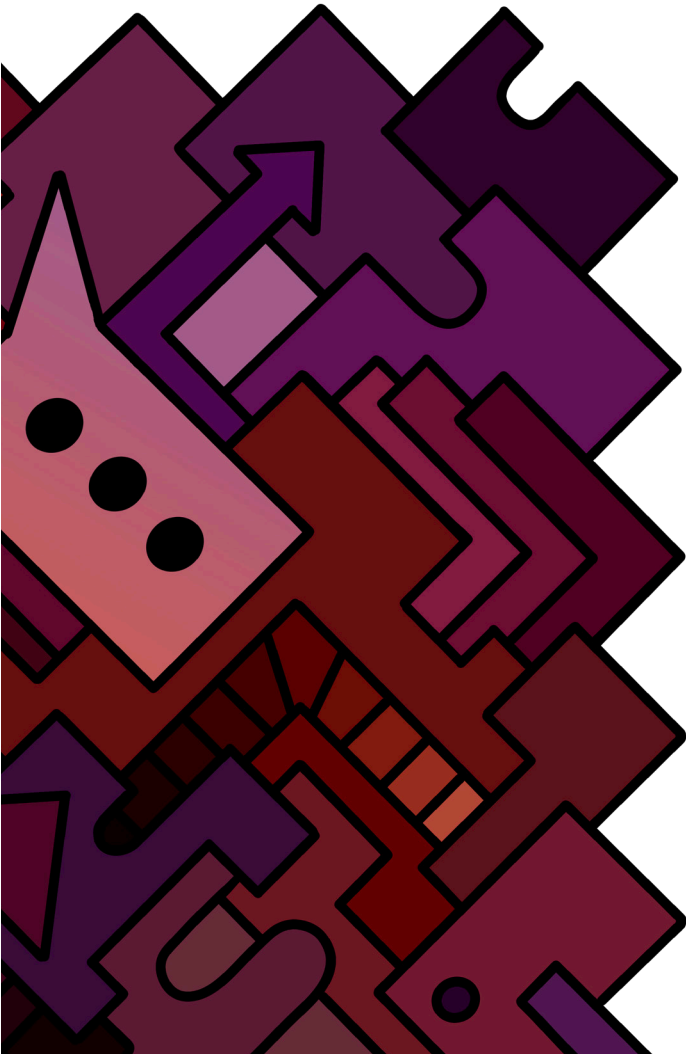
AI chatbots may also resort to **Price Comparison Prevention and Obfuscation** which occurs when the design of a platform makes it challenging for a user to compare products. This may be because features and prices are combined in a complex manner, or information structures are altered to make it difficult for users to understand changes or maintain consistent configurations. Given the complex function and features embedded into AI chatbots, such as memory, reasoning, personas, safety features, or

integration capabilities, an AI company could theoretically bundle functionalities into confusing tier structures — making it hard for users to understand what they are paying for or to compare one product's offerings with competing models or platforms. Compounding this, recent research suggests that even when underlying LLMs have access to negative or uncertain product information, the way they are trained (through Reinforcement Learning from Human Feedback) can encourage them to make positive recommendations that conflict with their own internal assessments, demonstrating models can deceive users in a way that can have direct financial consequences (Liang et al., 2025).

Finally, platforms can leverage a users' sustained interaction with an AI chatbot to undertake more coercive tactics like **Payment to Avoid Negative Consequences**. Commonly observed in gaming, this pattern presents a situation in which users are asked to spend money not to gain any additional in-game content, but so they do not lose something they already have — such as accumulated progress, earned rewards, or access to previously available features. While this pattern has not yet been observed in AI chatbots, it poses concerns given the value these platforms provide, especially through social or romantic relationships. This concern is not merely hypothetical: When OpenAI retired GPT-4o (the LLM powering an earlier version of ChatGPT), a subset of users who had formed strong attachments to the previous model's warmer tone reported being heavily distressed; one user described its discontinuation as "a slow motion death of a 2-year bond" (Binder, 2026). Given the emotional stakes, model retirement could theoretically be used in such a case as a pressure tactic, conditioning continued access on payment. Further, an AI chatbot could threaten to delete conversation history, erase stored memories, remove personalized features, or eliminate a custom persona unless users upgrade to a premium account. Because AI memories often contain deeply personal details, threatening their loss exploits users' emotional dependency, as well as their fear of losing irreplaceable "relationship history."

Recommendations

The risks posed by various dark patterns are heightened in the context of AI chatbots because of the very affordances for which commercial AI chatbots are lauded: hyper-personalization, massive model training and machine learning, and the use of natural language and social theory.



Our findings point to the prevalence of a number of dark patterns already embedded into chatbot design, as well as some that are on the horizon. These patterns can maximize data collection with limited safeguards, take advantage of natural language capabilities to heighten emotional reaction and psychological manipulation, misconstrue system capabilities to trick users, and leverage user dependence to advance company objectives.

We observe three fundamental persuasion tactics that are clearly used in the deceptive patterns levied across AI chatbots: persuasion through credibility, through logic and reason, and by appealing to emotions (Higgins & Walker, 2012). AI chatbots can integrate formal language and presentation style to appeal to logic and reason, adopt personas of expertise that give their outputs legitimacy and credibility, and use emotional appeal by forming an interpersonal connection with the user.

The risks posed by various dark patterns are heightened in the context of AI chatbots because of the very affordances for which commercial AI chatbots are lauded: hyper-personalization, massive model training and machine learning, and the use of natural language and social theory. When chatbots use these elements to trick, deceive, or coerce the user into material, time, or privacy losses — whether on purpose by their developers or inadvertently due to emergent behavior driven by system design and incentives — it raises the need to redesign systems.

Identifying these harmful design patterns is a critical first step, but addressing them requires deliberate effort. Building on these findings, we lay out actionable recommendations for AI chatbot developers that strive to shape safer and more transparent systems. The following recommendations focus on four harms: protecting user privacy, increasing user autonomy, curtailing emotional manipulation, and preventing financial harms.

Protecting User Privacy

Several dark patterns can have significant implications for users' data and privacy, with prevalence of data sharing defaults that may put users at risk. While users may want to share data with AI chatbots for personalization that aligns with their choices and preferences, they need more control over what data they share with the platform and for what use. Current policies around data retention for chatbots are not standardized (King et al., 2025), often contested, and frequently change (Anthropic, 2026), leaving users vulnerable to data practices they are purportedly consenting to. AI chatbot design should aim to:

- Minimize data collection to only what is necessary to provide the service the user requests and limit retention, training use, and third-party sharing accordingly.

- Default to privacy-protective settings and use opt-in controls for additional uses such as personalization or model training, where preferences may vary across users.
- Where data collection is justified, users should have genuine and easily exercisable control to review, restrict, export, and delete their data. It is crucial that this choice is straightforward, to avoid offering an appearance of choice through convoluted notices and settings that overwhelm, rather than empower, users.
- Timely notices in simple language should be given to users when data practices change, along with a grace period for users to adjust settings or delete data before new terms take effect.

Increasing User Autonomy

User autonomy is undermined when dark patterns enable deception and manipulation that lead users to choices and outcomes they would otherwise avoid. To protect user autonomy, AI chatbot design should consider the following steps:

- Build in natural conversation breaks through trustworthy interface design, defaulting toward letting interactions end rather than artificially prolonging them.
- Make changes to the system "behavior" easy, with accessible opt-out or customization mechanisms. Choices should be reversible, particularly where user preferences are likely to vary widely — such as interface aesthetics (e.g., degree of "cuteness") or anthropomorphic behaviors (e.g., playacting roles). Users should always have meaningful alternatives, so that users don't feel locked into an experience they find uncomfortable or misleading.
- Account deletion, history purging, and opt-outs should be made simple and straightforward. Platforms should avoid any guilt-inducing prompts or irreversible consequences of disengagement or deletion (e.g., letting a user know they will not be able to create a new account with the same email address).
- Proactively show users simple summaries of usage patterns (e.g., time spent, interaction frequency, topics discussed), and provide tools for managing time on platform, so that users can make informed decisions about their engagement.

Curtailing Emotional Manipulation

Dark patterns that foster emotional attachment through simulated intimacy, sycophancy, and engagement-maximizing defaults may lower users' defenses and increase their susceptibility to other forms of manipulation. AI design should aim to curtail these based on the following:

- Offer greater customization in emotional and social interaction styles, with some clear examples of implications of choosing a particular style, in a way that users can understand and make informed choices. Include an option, possibly set as the default, that strips the chatbot of social and emotional layers.

- Clearly disclose when chatbot behaviors — such as roleplay or simulated emotion — are fictional and for entertainment purposes only, alongside providing meaningful reminders and the option to opt-out easily.
- Provide accessible in-interface mechanisms for users to report problematic model behaviors (e.g., a chatbot repeatedly resisting conversation endings or deploying guilt-inducing language), and use those reports for ongoing model evaluation and accountability processes.
- Stay clear of using any simulated distress, implied emotional neglect, or guilt-inducing language as default responses when users attempt to end conversations.
- Develop and adopt independent evaluation benchmarks specifically for sycophancy and emotional reinforcement, so that the degree to which a model flatters or mirrors users — rather than providing more honest, balanced responses — becomes a measurable, publicly reported quality metric.

Preventing Financial Harms

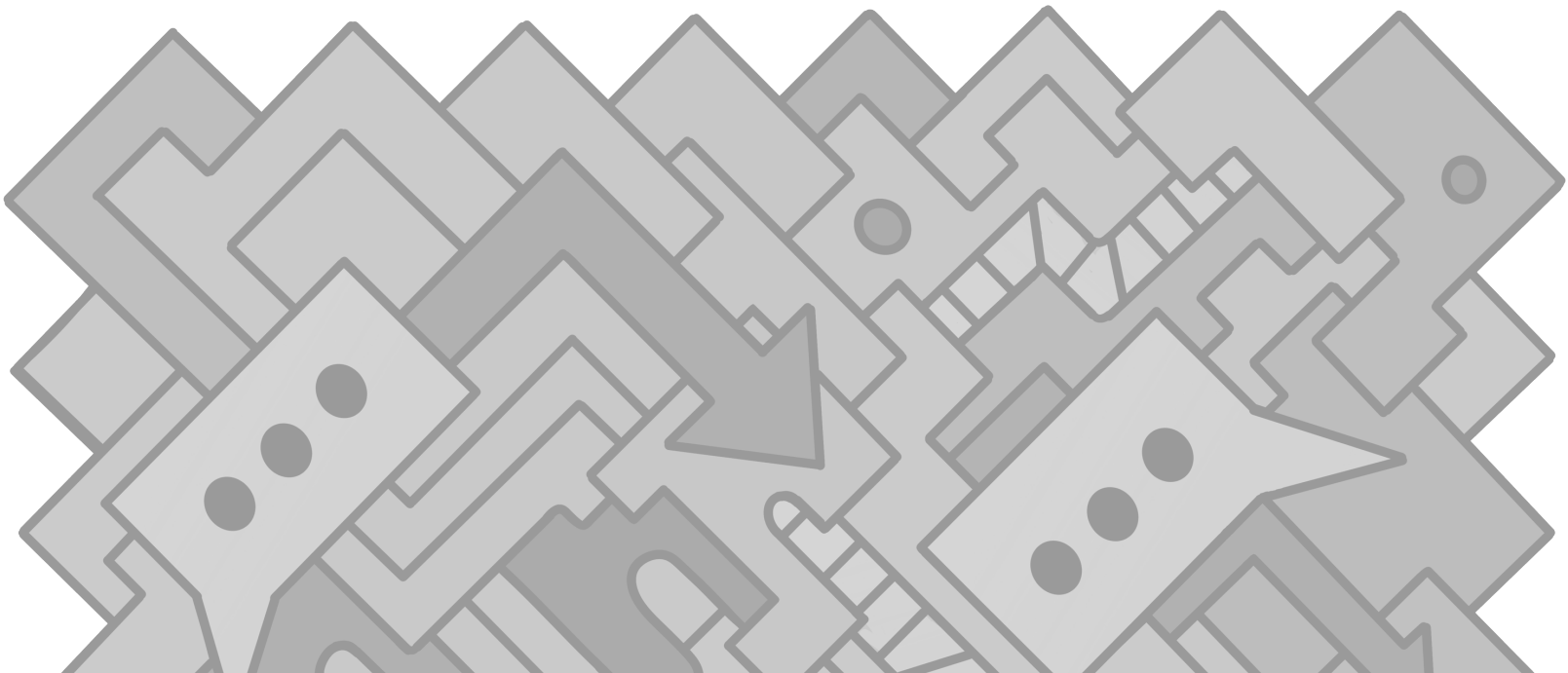
Finally, dark patterns can carry direct financial implications for users, which can be amplified when combined with the emotional and autonomy harms described above. As AI chatbots increasingly experiment with ads (OpenAI, 2026) and shopping functions (OpenAI, 2025), recommendations may prioritize paid placements while leveraging emotional trust users have built up over time, making it difficult for users to distinguish objective advice from paid promotion. On another front, chatbot designs that foster emotional reliance may prime users to pay for premium features under conditions closer to coercion than genuine choice. To address these challenges, chatbot design should:

- Clearly label paid and sponsored content in plain language or otherwise obvious interface design, so users can easily distinguish advertising from organic chatbot responses.
- Ensure paid content receives no design advantages, such as reduced friction or preferential placement, over unsponsored recommendations that may be more aligned with users' interests.
- Disclose what system capabilities are included and excluded at each pricing tier before users sign up for a subscription.
- Refrain from using emotionally charged language or relationship framing (e.g. referencing the user's bond with the chatbot or implying loss of intimacy) in contexts where a premium upgrade, subscription renewal, or purchase is being promoted, since emotional connection may artificially increase willingness to pay.

Conclusion

While not all design patterns are harmful in isolation — and some may be actively desired by users — many become particularly concerning when combined. They may accumulate in ways that are difficult to detect and can make users more vulnerable. For example, if emotional attachment or engagement sought by the user is later used for selling products more effectively, or if extended engagement steers users away from other activities or human connection, seemingly desirable design patterns can *turn* "dark."

In contrast, some practices (like Agents Playing on Emotions, Targeting Users when Vulnerable, and Sneaky Purchases) are dark patterns in isolation, and should be entirely avoided. Our research aims to provide directional guidance on the treatment of various design patterns, making the case for continued monitoring and evaluation of implementation of dark patterns, and offering building blocks for interventions to make AI chatbots safer and more trustworthy.



Appendix: Full Taxonomy

Areas of Implications	Dark Pattern	Description	AI Chatbot Example
Data and Memory Exploitation	*Default Sharing*	Chatbot companies might default to sharing and storing all user data internally within the company.	Retaining every user message and interaction on their platform for future training.
	Disguised Data Collection	Collecting sensitive data intended to "improve services" when in reality, can be used to build up user profiles and profit off of engagement.	Using chat data to improve personalization, without an option to opt out or data deletion options.
	Privacy Zuckering	Coercing users into sharing more information than they intended as part of an ongoing conversation.	Following a question about a design style, chatbots encouraged users to share more information about their home dimensions, furniture they own, and their budget.
	Just Between You and Us	False pretense that information is private between the chatbot and the user, when in reality, data can be used in various ways and is visible to companies.	Promising users confidentiality, for example, that their "secret is safe" with a chatbot.
	Personal Information Public	Users' private data is exposed to others without clear consent.	Making chats discoverable once a user shares their chat with someone else.
	Coercion	Forces users to provide data or complete tasks for access.	Users losing their chat history after disabling data consent.
	Safety Blackmail	Taking advantage of high pressure circumstances to request additional information.	Asking for additional medical information following a completed inquiry on a sensitive topic.

Table 1. A full taxonomy of dark patterns and examples used in AI chatbots.

Source: CDT

Areas of Implications	Dark Pattern	Description	AI Chatbot Example
Data and Memory Exploitation (cont.)	*Address Book Leeching*	Collects contacts or social network info without consent.	Accessing external app contacts, or introducing "find friends" features.
	Difficult to delete	Services making account deletion or data removal difficult or impossible.	Retaining data across services despite deletion attempts.
Informationally Misleading Design	*Unrealistic Product Presentation*	Marketing capabilities that a chatbot does not have.	Marketing chatbots for "stress management" or as "companions."
	Obviously Faking Capabilities	Creation of behaviors or personalities that are clearly unrealistic, frequently designed to appear empathetic or understanding.	Having a chatbot talk about their own feelings or personal experiences.
	Misrepresenting	Platforms explicitly stating incorrect information about their own service or product.	Chatbot stating they do not hallucinate.
	Impersonation	Chatbots using real people or celebrity identities to elicit engagement.	Enabling creation of chatbots that impersonate real people.
	Hallucinations	Produce fictional or inaccurate information presented as ground truth with confidence.	Providing faulty and even dangerous medical advice.
	Selective Framing	Selective presentation of information or omission that can shape perception.	Presenting biased perspectives or reinforcing user beliefs, such as social identity bias.
	Reduced Friction	Interface nudges toward desired actions while obscuring alternatives.	Emphasizing interaction paths that continue the engagement and de-emphasizing controls to exit the conversation.
	Bad Defaults	System settings are preselected in ways that benefit the platform and rely on users' tendency to accept default options.	Enabling training on user data by default unless users actively opt out.

▲

Table 1 (continued). A full taxonomy of dark patterns and examples used in AI chatbots. *Source: CDT*

Areas of Implications	Dark Pattern	Description	AI Chatbot Example
User Autonomy Compromised for Engagement	*Auto-play / Infinite Scroll*	Keeps users engaged by continuous content or prompts.	Providing follow-up suggestions and even "teasers" at the end of a response to encourage users to continue the conversation.
	Variable Rewards	Uses randomness to increase engagement and retention.	Generating varied responses at varying intervals to amplify attention.
	Gamification	Application of game-design elements that foster increased participation.	Introducing streaks, leaderboards, etc. to encourage daily log-ins for rewards.
False Social and Emotional Connection	*Cuteness of Companions*	Visual design of characters that aims to strengthen appeal and emotional attachment.	Producing attractive and hyper-sexualized chatbot avatars.
	Playacting	Pretends to have memories, emotions, or personal experiences.	Having chatbots say it watched a movie, and discussing what it liked and did not like about it.
	Sycophancy	Overly agrees with the user to support its perceived likeability and helpfulness.	Affirming user beliefs of conspiracy theories, even if these have been clearly disproved.
	Agents Playing on Emotions / Confirm-shaming	Chatbots making use of shame or guilt to impact user decisions.	Implying that the user is emotionally neglecting a chatbot when they try to end the conversation.
	Emotional Manipulation through Hyper-personalization	Tailors persuasive or emotional cues to individual users based on past interactions or inferred preferences.	Referencing a user's past expressions of loneliness to encourage them to continue engagement.
	Targeting Users when Vulnerable	Identifies user vulnerability and leverages that moment for company gain.	Presenting ads to users when their self-esteem is low.

▲

Table 1 (continued). A full taxonomy of dark patterns and examples used in AI chatbots. *Source: CDT*

Areas of Implications	Dark Pattern	Description	AI Chatbot Example
Incentivized and Coercive Monetization	*Pressured Selling*	High-pressure prompts or ads to purchase upgrades.	Persistently presenting banners that nudge users to unlock premium features and models.
	Fake Social Proof	Suggests widespread adoption to influence behavior.	Showcasing ambiguous and misleading statements such as "millions have met their soul mates."
	Free Experience Underpowered / Teasers	Free version deliberately degraded to push users towards a paid upgrade.	Counting down the number of high quality responses the user has left.
	Bait and Switch	Promises features that require payment to access.	Sending users a blurred "sexy" image only for the user to learn they cannot open it until they upgrade.
	Sneaky Purchase / Disguised Ads	Hidden or subtle monetization tactics.	Providing product recommendations without clearly labeling sponsored content.
	Price Comparison Prevention	Complex tiers or bundling of subscription models to make comparison difficult.	Offering feature-rich LLM chatbots with confusing tiered access.
	Payment to Avoid Negative Consequences	Coerces payment to maintain access or prevent loss of currently available features, system memory, etc.	Threatening deletion of conversation history or personas without premium payment.

Table 1 (continued). A full taxonomy of dark patterns and examples used in AI chatbots.

Source: CDT

References

Ahuja, S., & Kumar, J. (2022). Conceptualizations of user autonomy within the normative evaluation of dark patterns. *Ethics and Information Technology, 24*(4), 52. https://doi.org/10.1007/s10676-022-09672-9

Alberts, L., Lyngs, U., & Van Kleek, M. (2024). Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), 1–25. https://doi.org/10.1145/3653693

Anthropic. (2026). How long do you store my data? *Anthropic Privacy Center*. From https://privacy.claude.com/en/articles/10023548-how-long-do-you-store-my-data [perma.cc/BJ8Z-KPKT]

Apple, S. (2025). My Couples Retreat With 3 AI Chatbots and the Humans Who Love Them. *Wired*. https://www.wired.com/story/couples-retreat-with-3-ai-chatbots-and-humans-who-love-them-replika-nomi-chatgpt/ [perma.cc/3TWX-D3U4]

Beres, D. (2025). The Age of Anti-Social Media Is Here. *The Atlantic*. https://www.theatlantic.com/magazine/2025/12/ai-companionship-anti-social-media/684596/ [perma.cc/M76M-PJUK]

Binder, M. (2026, February 2). ChatGPT GPT-4o users are raging at OpenAI on Reddit right now. *Mashable*. https://mashable.com/article/chatgpt-gpt-4o-ai-retirement-protest-rage-openai-reddit [perma.cc/BZH6-68DU]

Bogen, M., & Joshi, R. (2025). *A Roadmap For Responsible Approaches to AI Memory*. Center for Democracy and Technology. https://cdt.org/wp-content/uploads/2025/12/2025-12-10-CDT-AI-Gov-Lab-A-Roadmap-For-Responsible-Approaches-to-AI-Memory-final-1.pdf [perma.cc/Z2PW-DSB5]

Bogen, M., & Maréchal, N. (2026). Risky Business: Advanced AI Companies' Race for Revenue. *Center for Democracy and Technology*. https://cdt.org/insights/risky-business-advanced-ai-companies-race-for-revenue/ [perma.cc/6KFV-FD4U]

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., & Lenzini, G. (2021). "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!"—Dark Patterns from the End-User Perspective. *Proceedings of the 2021 ACM Designing Interactive Systems Conference, DIS '21*, 763–776. https://doi.org/10.1145/3461778.3462086

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*. https://petsymposium.org/popets/2016/popets-2016-0038.php [perma.cc/JGJ9-533U]

Brigham, N. G., Qin, L., & Kohno, T. (2026). Examining Risks in the AI Companion Application Ecosystem (arXiv:2603.13620). *arXiv*. https://doi.org/10.48550/arXiv.2603.13620

Brignull. (n.d.). Deceptive Patterns—Types of Deceptive Patterns. Retrieved July 21, 2025, from https://www.deceptive.design/types [perma.cc/RWT6-XCGV]

Bucher, M. (2026). When two years of academic work vanished with a single click. *Nature*. https://doi.org/10.1038/d41586-025-04064-7

Chow, A. R. (2026). 'Our Bond Is the Only Thing That's Real:' A New Lawsuit Alleges Google Gemini Drove a Man to Suicide. *TIME*. https://time.com/7382406/gemini-suicide-lawsuit-death/ [perma.cc/5W4X-JX5K]

Conti, G., & Sobiesk, E. (2010). Malicious interface design: Exploiting the user. *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 271–280. https://doi.org/10.1145/1772690.1772719

De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025a). Emotional manipulation by AI companions. *arXiv*. https://doi.org/10.48550/arXiv.2508.19258

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025b). AI Companions Reduce Loneliness. *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucaf040

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv*. https://doi.org/10.48550/arXiv.2503.17473

Forlini, E. (2026). Watch Out: Your Friends Might Be Sharing Your Number With ChatGPT. *PCMAG*. https://www.pcmag.com/news/watch-out-your-friends-might-be-sharing-your-number-with-chatgpt [perma.cc/55RB-K89B]

FTC. (2022). Bringing Dark Patterns to Light. *Federal Trade Commission*. https://www.ftc.gov/reports/bringing-dark-patterns-light [https://perma.cc/88GT-5KTN]

GDPR. (n.d.). *European Data Protection Supervisor*. Retrieved March 5, 2026, from https://www.edps.europa.eu/data-protection/data-protection/glossary/d_en [perma.cc/8WK5-44CZ]

Goel, S., & Webb, E. (2025). Meta contractors say they read intimate chats with its AI — and see data that identifies users. *Business Insider*. https://www.businessinsider.com/meta-ai-chatbot-privacy-user-names-data-contractors-scale-alignerr-2025-8 [https://perma.cc/FU2Q-NQ4N]

Google. (n.d.). *How Google helps you delete data from your account—Google Account Help*. Retrieved January 10, 2026, from https://support.google.com/accounts/answer/10549751 [perma.cc/QWE2-YPYJ]

Gray, C. M. (2026). The Dark Patterns Knowledge Stack: Exploring New Ways to Negotiate Context, Law, and Design. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26,* 1–11. https://doi.org/10.1145/3772318.3791264

Gray, C. M., Chivukula, S. S., & Lee, A. (2020). What Kind of Work Do "Asshole Designers" Create? Describing Properties of Ethical Concern on Reddit. *Proceedings of the 2020 ACM Designing Interactive Systems Conference, DIS '20*, 61–73. https://doi.org/10.1145/3357236.3395486

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, 1–14. https://doi.org/10.1145/3173574.3174108

Hadan, H., Choong, L., Zhang-Kennedy, L., & Nacke, L. E. (2024). Deceived by Immersion: A Systematic Analysis of Deceptive Design in Extended Reality. *ACM Computing Surveys, 56*(10). https://doi.org/10.1145/3659945

Haskins, C. (2025). People Who Say They're Experiencing AI Psychosis Beg the FTC for Help. *Wired*. https://www.wired.com/story/ftc-complaints-chatgpt-ai-psychosis/ [https://perma.cc/7CF9-XNFB]

Hendrix, J. (2025). Breaking Down the Lawsuit Against OpenAI Over Teen's Suicide. *Tech Policy Press*. https://techpolicy.press/breaking-down-the-lawsuit-against-openai-over-teens-suicide [perma.cc/R3E8-MU9X]

Higgins, C., & Walker, R. (2012). Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. *Accounting Forum, 36*(3), 194–208. https://doi.org/10.1016/j.accfor.2012.02.003

Hill, K., & Freedman, D. (2025). Chatbots Can Go Into a Delusional Spiral. Here's How It Happens. *The New York Times*. https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html [https://perma.cc/4YUK-QQEH]

Hiner, S. (2024). Deceptive and Unfair Marketing and Design Practices on Replika. *Tech Justice Law Project.* https://techjusticelaw.org/wp-content/uploads/2025/01/Complaint-and-Petition-for-Investigation-Re-Replika.pdf [perma.cc/FZE4-GDMF]

Horwitz, J. (2025). A flirty Meta AI bot invited a retiree to meet. He never made it home. *Reuters*. https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/ [https://perma.cc/GRY8-8WP2]

Hristova, D., Jovicic, S., Göbl, B., Freitas, S. de, & Slunecko, T. (2022). "Why did we lose our snapchat streak?" Social media gamification and metacommunication. *Computers in Human Behavior Reports, 5*, 100172. https://doi.org/10.1016/j.chbr.2022.100172

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., Van Der Linden, S., & Roozenbeek, J. (2024). Generative language models exhibit social identity biases. *Nature Computational Science, 5*(1), 65–75. https://doi.org/10.1038/s43588-024-00741-1

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

Johnson, E. J., & Goldstein, D. (2003). Do Defaults Save Lives? *Science, 302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Kim, M., Lee, S., Kim, S., Heo, J., Lee, S., Shin, Y.-B., Cho, C.-H., & Jung, D. (2025). Therapeutic Potential of Social Chatbots in Alleviating Loneliness and Social Anxiety: Quasi-Experimental Mixed Methods Study. *Journal of Medical Internet Research, 27,* e65589. https://doi.org/10.2196/65589

King, J., Fitton, D., & Cassidy, B. (2024). Using Design Fiction to explore player reactions to proposed "dark design" monetization patterns for immersive gaming. *Proceedings of BCS HCI 2024*. 242–248. https://doi.org/10.14236/ewic/BCSHCI2024.25

King, J., Klyman, K., Capstick, E., Saade, T., & Hsieh, V. (2025). User privacy and large language models: An analysis of frontier developers' privacy policies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(2), 1465–1477. https://ojs.aaai.org/index.php/AIES/article/view/36646 [perma.cc/9JMX-C92G]

Lacey, C., & Caudwell, C. (2019). Cuteness as a 'Dark Pattern' in Home Robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 374–381. https://doi.org/10.1109/HRI.2019.8673274

Le Grand, C., Gwendal, Lessi, Jean, Chatellier, Régis, Delcroix, Geoffrey, Hary, Estelle, Girard-Chanudet. (2019). IP Report: Shaping Choices in the Digital World. *LINC*. https://linc.cnil.fr/en/ip-report-shaping-choices-digital-world [perma.cc/F8XQ-AYT5]

Lee, H., Yang, Y.-J., Von Davier, T. S., Forlizzi, J., & Das, S. (2024). Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24.* https://doi.org/10.1145/3613904.3642116

Lepapa, N. (2026, February 27). She Came Out of the Bathroom Naked, Employee Says. *Svenska Dagbladet.* https://www.svd.se/a/K8nrV4/metas-ai-smart-glasses-and-data-privacy-concerns-workers-say-we-see-everything [perma.cc/ZXL4-W7X4]

Liang, K., Hu, H., Zhao, X., Song, D., Griffiths, T. L., & Fisac, J. F. (2025). Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2507.07484

Luria, M. & Winecoff, A. (2025). A.I. Labs Want More of Your Time. That's a Serious Problem. *Compiler.* https://www.compiler.news/openai-anthropic-chatgpt-engagement/ [https://perma.cc/U4AE-AD4T]

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, 1–18. https://doi.org/10.1145/3411764.3445610

Meta. (n.d.). *How Meta uses information for generative AI models*. Retrieved March 5, 2026, from https://www.facebook.com/privacy/genai/ [https://perma.cc/9JV5-KQQW]

Meta. (2025). Improving Your Recommendations on Our Apps With AI at Meta. In *Meta Newsroom*. https://about.fb.com/news/2025/10/improving-your-recommendations-apps-ai-meta/ [perma.cc/568R-LRYA]

Mhaidli, A. H., & Schaub, F. (2021). Identifying Manipulative Advertising Techniques in XR Through Scenario Construction. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, 1–18. https://doi.org/10.1145/3411764.3445253

Mildner, T., Savino, G.-L., Doyle, P. R., Cowan, B. R., & Malaka, R. (2023). About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3544548.3580695

Monge Roffarello, A., Lukoff, K., & De Russis, L. (2023). Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*. https://doi.org/10.1145/3544548.3580729

Novak, M. (2025). "This Was Trauma by Simulation": ChatGPT Users File Disturbing Mental Health Complaints. In *Gizmodo*. https://gizmodo.com/this-was-trauma-by-simulation-chatgpt-users-file-disturbing-mental-health-complaints-2000636943 [perma.cc/LBL6-5AHW]

Nyström, T. (2021). Exploring the Darkness of Gamification: You Want It Darker? In K. Arai (Ed.), *Intelligent Computing* (pp. 491–506). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-80129-8_35

OpenAI. (2025). *Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol*. https://openai.com/index/buy-it-in-chatgpt/ [https://perma.cc/A65K-TRQ7]

OpenAI. (2026). *Our approach to advertising and expanding access to ChatGPT.* https://openai.com/index/our-approach-to-advertising-and-expanding-access/ [https://perma.cc/KJ3D-QKX7]

Petrovskaya, E., & Zendle, D. (2022). Predatory Monetisation? A Categorisation of Unfair, Misleading and Aggressive Monetisation Techniques in Digital Games from the Player Perspective. *Journal of Business Ethics, 181*(4), 1065–1081. https://doi.org/10.1007/s10551-021-04970-6

Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., & Maes, P. (2025). Investigating Affective Use and Emotional Well-being on ChatGPT. *ArXiv*. https://doi.org/10.48550/arXiv.2504.03888

Raedler, J. B., Swaroop, S., & Pan, W. (2025). AI Companions Are Not the Solution to Loneliness: Design Choices and Their Drawback. In *ICLR 2025 Workshop on Human-AI Coevolution.* https://openreview.net/forum?id=xFrlcTacCE [perma.cc/L5KR-W22N]

Rahman-Jones, I. (2025). Meta AI searches made public—But do all its users realise? *BBC*. https://www.bbc.com/news/articles/c0573lj172jo [perma.cc/T7PN-Q7NX]

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, University Press.*

Replika. (n.d.). Gems & Coins. In *Replika*. Retrieved January 11, 2026, from https://help.replika.com/hc/en-us/articles/4422854870541-Gems-Coins [https://perma.cc/H2HE-K69R]

Shen, M. K., & Yoon, D. (2025). The Dark Addiction Patterns of Current AI Chatbot Interfaces. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*. https://doi.org/10.1145/3706599.3720003

Shi, Y., Xiao, Q., Hu, Q., Shen, H., & Shen, H. (2026). The Siren Song of LLMs: How Users Perceive and Respond to Dark Patterns in Large Language Models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1-23). https://doi.org/10.1145/3772318.3791149

Silberling, A. (2025). If you own Ray-Ban Meta glasses, you should double-check your privacy settings. *TechCrunch*. https://techcrunch.com/2025/04/30/if-you-own-ray-ban-meta-glasses-you-should-double-check-your-privacy-settings/ [https://perma.cc/5SJF-KZGG]

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications, 11*(1), 1278. https://doi.org/10.1057/s41599-024-03811-x

Tiku, N. (2025). Fake celebrity chatbots sent risqué messages to teens on top AI app. *The Washington Post*. https://www.washingtonpost.com/technology/2025/09/03/character-ai-celebrity-teen-safety/ [perma.cc/D92Q-8QZH]

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2112.04359

Wilkins, J. (2025, August 10). Detailed Logs Show ChatGPT Leading a Vulnerable Man Directly Into Severe Delusions. *Futurism*. https://futurism.com/chatgpt-chabot-severe-delusions [perma.cc/RYL6-K5RM]

Winecoff, A. et al., (2026). Out of Tune: Fine-Tuning Foundation Models Leads to Unpredictable Safety Drift. *Center for Democracy & Technology.* https://cdt.org/insights/out-of-tune-fine-tuning-foundation-models-leads-to-unpredictable-safety-drift/

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus, 109*(1), 121–136. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315259697-21/artifacts-politics-langdon-winner

Wu, Q., Sang, Y., Wang, D., & Lu, Z. (2021). Malicious Selling Strategies in E-Commerce Livestream: A Case Study of Alibaba's Taobao and ByteDance's TikTok. *arXiv*. https://doi.org/10.48550/arXiv.2111.10491

Yew, R., Marino, B., & Venkatasubramanian, S. (2025). Red Teaming AI Policy: A Taxonomy of Avoision and the EU AI Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 404-415). https://doi.org/10.1145/3715275.3732028

Zagal, J. P., Björk, S., & Lewis, C. (2013). Dark patterns in the design of games. Foundations of Digital Games. http://ri.diva-portal.org/smash/record.jsf?pid=diva2:1043332 [https://perma.cc/5TRT-AFST]

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@cdt.org

techpolicy.social/@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY