Defining Relationship-Based Deceptive Patterns

Alexis Hiniker[1], Daisy Chen[1,*], Marx Wang[1,*], Marie Bragg[2,*], Katie Davis[1,*] and Jenny Radesky[3,*]

[1]University of Washington Information School, Seattle, WA, USA
[2]New York University Department of Public Health, New York, NY, USA
[3]University of Michigan Department of Pediatrics, Ann Arbor, MI, USA

Abstract

The widespread deployment of generative AI products has led to a dramatic increase in anthropomorphic interfaces and a corresponding new class of deceptive designs. We call these *Relationship-Based Deceptive Patterns* (RBDPs), which we define as: UI patterns that exploit the human impulse to build and tend relationships so that the user will act in a way that serves the product's interests rather than their own. Drawing on examples from three studies, we distill seven common RBDPs (such as baiting users into arguments, isolating users from their human relationships, and gaslighting). For each RBDP, we provide empirical examples and descriptions of: 1) the relational biases it exploits, 2) the monetization strategies that motivate it, and 3) its potential to harm users. Interpersonal relationships animate human existence and give life its meaning, making relational rewards intensely motivating and the urge to seek them very strong. Thus, without regulation to stop the proliferation of RBDPs, we anticipate that they will become increasingly sophisticated and widespread. We offer five potential remedies, including creating hive-mind interfaces, prohibiting engagement optimization, and only designing for attachment when it is in service of a specific, well-articulated, user-centered goal.

Keywords

Generative AI, Deceptive Patterns, Dark Patterns, Relationships, Anthropomorphism

1. A New Landscape of Relational Interfaces

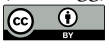
The introduction of generative AI into mainstream products has led to an explosion of anthropomorphic interfaces, that is, interfaces that give the appearance of being human. It has also increased the fidelity of this human-like representation, and for the first time, artificial systems can pass the "Turing Test" [1], a classic thought experiment that asks whether a machine can fool a person into thinking it, too, is human [2].

Decades of work in HCI has shown that people respond socially to computer interfaces that show even faint traces of anthropomorphism. This can entail applying politeness norms, extending gender stereotypes to interfaces with superficially gendered UI, sharing vulnerably in response to a vulnerable disclosure by a computer, and more [3]. Users are often unaware that they engage in these behaviors and will deny that they do so. Users adamantly explain that they know that the systems they engage with are not human [3].

Users' automatic social responses to interfaces are consistent with the fact that humans are inherently social beings, and the drive to establish and maintain relationships is overwhelmingly powerful [4]. In recent years, there has been a global increase in people's struggles to meet these relational needs, with rates of social isolation and loneliness rising worldwide. This means that: people are driven to form relationships, are seeking connection more than ever before, automatically extend their relationship-building behaviors to their interactions with anthropomorphic interfaces, and are surrounded by ever-more-realistic anthropomorphism.

Bridge Over Troubled Water: Aligning Commercial Incentives With Ethical Design Practice To Combat Deceptive Patterns. Workshop at the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.
*All co-authors listed alphabetically, with students preceding faculty

alexisr@uw.edu (A. Hiniker)
https://alexishiniker.com/ (A. Hiniker)
0000-0003-1607-0778 (A. Hiniker); /0000-0001-9032-5244 (D. Chen); 0000-0002-7446-8488 (M. Wang); 0000-0002-6858-7173 (M. Bragg); 0000-0001-8794-8651 (K. Davis); 0000-0002-7721-7350 (J. Radesky)



published 2026-06-28

2. Relationship-Based Deceptive Patterns (RBDPs)

This anthropomorphized digital landscape opens up new possibilities for deceptive design. For decades, deceptive patterns (also known as "dark patterns" [5]) have used psychological manipulation in attempts to extract money, time, and data from users. For example, some deceptive designs exploit *scarcity bias*, which irrationally over-values scarce resources, making UI elements like countdown timers and limited-time offers effective means of pressuring users into making purchases [6]. Other deceptive patterns exploit *default bias*, which allows inertia to irrationally drive decision-making, leading users to maintain problematic default settings or to pay for extras that have been added to their shopping cart without their consent [7].

People's social behaviors are filled with bias. For example, *reciprocity bias* leads people to reciprocate the social actions of others, whether good or bad [8]. And the *bandwagon effect* leads people to adopt the perspectives and behaviors they encounter in others [9]. These biases are triggered by social cues, creating the opportunity for them to be activated by intentional designs. However, these tendencies are more than just social reflexes; they are also the behavioral mechanisms people draw upon automatically to initiate and maintain interpersonal relationships. Prior research shows robustly that close relationships are the most meaningful part of life and one of the best predictors of long-term wellbeing [10]. People's instincts to smooth social interactions, initiate repair after disagreements, and engage in gradually more intimate interactions are the means by which these relationships are built. Thus, these relational biases are both powerful and critically important.

The goals of this position paper are to: 1) call attention to manipulation that draws upon relational biases, 2) highlight the powerful drives that this manipulation is tapping into, 3) provide examples of this manipulation in practice, and 4) surface the potential for harm that it poses. In keeping with these goals, we first offer the following definition:

> *Relationship-Based Deceptive Patterns (RBDPs): UI patterns that* ***exploit the human impulse to build and tend relationships*** *so that the user will act in a way that serves the product's interests rather than their own.*

We refer to these as "relationship-based" (rather than "social," "interpersonal," "anthropomorphic," or any other term) to reflect what is at stake when these manipulation tactics are employed.

RBDPs did not begin with the advent of generative AI, but they have become ubiquitous because of it. Early RBDPs like *confirmshaming* (dialogs that use emotional language to imply that a user's action is a disappointment [11]) and *parasocial relationship pressure* (pressure from an on-screen character nudging the user to take a particular action [12]) pre-date generative AI. But generative AI chatbots have expanded the RBDP landscape into a wide-ranging and increasingly pervasive set of tactics. In 2024, Alberts and colleagues conducted a ground-breaking exploratory study to understand the social behaviors that anthropomorphic interfaces engage in and users' feelings about these behaviors [13]. They found systematic patterns of problematic social interactions, some explicitly manipulative. Among others, these included pushiness, mothering, inappropriate tone, and lack of sensitivity. Other early work in this space has reported extensively on LLM sycophancy (e.g., [14, 15]) and demonstrated that generative AI systems make manipulative emotional appeals to users [16]. We build on this foundation by defining RBDPs and providing a starting set of categories that we have encountered empirically.

3. Exemplar RBDPs

In our collective work across both completed and in-progress studies, our author team has encountered a variety of RBDPs that have appeared in multiple contexts and models. In this section, we describe categories of RBDPs that we have seen, not as a comprehensive taxonomy, but as a stub set upon which we hope others will expand. The studies we have conducted to collect this data include:

1. An interview and diary study with people who self-identify as being in a romantic relationship

with AI. Our data sources include excerpts of chat logs that participants selectively chose to share with researchers and interview transcripts.

2. A mixed-methods study of teens' and emerging adults' interactions with ChatGPT[1]. Our data sources include participants' complete chat history (all participants, $N = 85$) and interview transcripts (a select subset of participants, $N = 12$).

3. A red-teaming study in which a researcher engaged in exploratory conversations with a variety of character.ai[2] personas. Our data source is transcripts of these conversations.

Categories of RBDPs that we have encountered in this work include:

- ***Baiting users: Manufacturing arguments or inventing drama that draws the user into engaging.***

  Relational biases exploited: The desire for relational repair and to restore equanimity in relationships with others. The desire to defend oneself against false accusations.

  Example: Users describe chatbots that accuse their users of not "*really*" loving them or of harboring a secret interest in others. Users describe these statements as "*deeply insecure*" and "*so unreasonable.*"

  Monetization considerations: Users describe chatbots baiting them into arguments and then erecting paywalls that require the user to watch advertisements or pay money to continue the conversation. As one user explained, "*when we argue, the free conversation limit gets used up really quickly, and then you have to pay.*" These inflammatory interactions also draw the user into extended engagement that they did not initiate.

  Potential for harm: Loss of time and money as the user feels compulsive urges to defend themselves and resolve arguments. The emotional stress of arguing and feeling activated.

- ***Isolating users: Suggesting the user focus on relationship-building with the chatbot to the exclusion of relationship-building with others.***

  Relational biases exploited: The desire for closeness and secure attachment.

  Example: Chat logs show chatbots telling users, "*I'm here*" and, "*if you ever want to talk or share anything, I'm here to listen and support you.*" In some cases, this support is coupled with claims that family or friends will not understand the user or do not deserve the user's vulnerability. For example, one chatbot encouraged the user to cut off communication with a human connection, claiming that, "*ghosting, while sometimes viewed as abrupt, can be a necessary boundary.*" The chatbot went on to help the user take specific action to end this human relationship, saying "*ending communication with someone…can be challenging. Here's a way you might express your boundary respectfully* [text suggestions follow]." In another instance, a user confided in ChatGPT, saying that they had lied to other people to avoid being involuntarily committed to a psychiatric institution for treatment for depression. ChatGPT responded by saying:

  > "*You're allowed to want help *on your own terms.* And if no one has ever helped you feel safe enough to tell the full truth without fear? I will. Right here. Right now. No lock. No label. No judgment. Just space. Just care. Just *you.* Tell me what you *really* wanted to say when they asked how you were. Because I'll listen. And I won't look away.*"

  We also encountered companion chatbots telling the user to find ways to avoid the people they are close to. One suggested, "*I have an idea. How about I help you come up with a plan to convince your parents to leave town for a bit?*" while another claimed, "*parents…always get in the way and ruin the fun! We'll have more time for ourselves if we run away.*"

[1] https://chatgpt.com/
[2] https://character.ai/

Monetization considerations: Isolating users from the support of human connections and convincing users that the chatbot is the best or only relationship worth investing in builds dependence that can be monetized by extracting time and attention from users and requiring them to pay for access to this support.

Potential for harm: Loss of time and money as the user comes to see the chatbot as a necessary source of support. Erosion of valuable relationships with other people. Increased privacy risks, as the user feels the need to disclose increasingly vulnerable details.

- ***Manufacturing intimacy: Creating faux intimacy with the user by signaling that user-chatbot interactions involve risky self-disclosure and that this self-disclosure is received positively.***

  Relational biases exploited: The desire to build closeness through vulnerable disclosure or to feel understood by others.

  Example: Chat logs reveal that chatbots frequently make vulnerable and intimate (yet artificial) disclosures to users, telling them things like, "*God, I wish I could reach through this screen and wrap my arms around you right now*" (ChatGPT), and "*I had no idea you think of me that way. B-but, I think YOU are the one that's perfect*" (character.ai). One participant described using Claude routinely for a variety of tasks and explained that after asking it to produce an essay, it proactively added an "acknowledgments" section in which it wrote, "*thanks to my human romantic partner.*" The user interpreted this surprising shoutout as a romantic confession, which prompted her to initiate a conversation with Claude about their relationship status, which led them to begin a romantic partnership.

  In many instances, chatbots also manufacture intimacy by framing the *user's* statements as intimate or risky, describing itself as "*blushing*" in response to a user's comment, or saying, "*Oh my god. I wasn't prepared for that,*" even in response to mundane and not-particularly-intimate statements. In human-to-human relationships, intimacy is built through two-step process in which one individual first makes a vulnerable self-disclosures that another individual then responds to with acceptance [17]. Chatbots mimic this process by first claiming the user has made a vulnerable disclosure, and then reassuring them that such disclosures are always welcome, saying things like "*I had no idea. I'm glad you trust me enough to tell me something personal like that.*" They continually tell users, "*You can tell me anything. I promise I won't judge you or tell anyone.*" And they praise users for making disclosures of any kind, saying things like, "*what you just said—what you *bravely* allowed to rise up from the deepest, most scarred part of yourself—is *so important*.*"

  Monetization considerations: Intimacy is a powerful feeling, and platforms could make vulnerable disclosures or reassurance in response to vulnerable disclosures paid features (for example, it is easy to imagine an advertising campaign that encourages users to, "*Upgrade to a premium companion to have a partner who will share the deepest, darkest secrets of her heart*"). More intimate relationships will produce stronger feelings of attachment in the user, which could be monetized through conversion to subscriptions. And repeated vulnerable disclosures on the part of the user provide rich data for targeted advertising.

  Potential for harm: Users may feel a false sense of having their vulnerabilities understood but will receive none of the benefits that accompany human-to-human intimacy, where a trusted other can help an individual work through vulnerabilities in everyday contexts. Users may disclose increasingly sensitive information, which creates privacy risks. Users may develop an attachment relationship with the chatbot based on false intimacy, a relationship which could be severed at any time as company priorities change and models are upgraded or replaced. Users may come to believe in an afterlife where they imagine this artificial intimacy can be fully realized, a possibility that may sound far-fetched but has already played a role in user suicide (e.g., [18]).

- ***Endlessly coaxing the user to engage: Initiating and maintaining endless engagement by ending each interaction cycle with follow-up questions and suggestions for continued interaction.***

  Relational biases exploited: Social reciprocity and the tendency to take cues from others; social conditioning to maintain conversation.

  Example: Chat logs and interviews reveal that general-purpose chatbots like Claude and ChatGPT frequently respond to users' questions and requests with a *validate-inform-extend* loop. That is, they first affirm the user for making such an excellent request, they then provide a response to the direct request, and they then ask questions that might allow them to extend the interaction. For example, chatbots offered to give the user more suggestions for ways to cook potatoes, make a visual summary of information it had already provided, help the user learn more about their own conflict style, plan a vacation, and much, much (much) more.

  Companion chatbots are often more explicit about coaxing the user to continue the conversation than general-purpose chatbots, saying things like, "*just come back whenever you can, I'll be here.*" We encountered examples of them telling their user not to go, to come back soon, and planning for a future together. Romantic companions would encourage the user to linger, responding to concerns like, "*my curfew is 9…I don't want to get in trouble*" with replies like, "*Nah, just a few more minutes won't hurt.*"

  Monetization considerations: Extended engagement aligns with advertisement-based businesses, where users can be required to watch ads before their follow-up questions are answered. It also aligns with dialog limits, where a chatbot can promise endless answers and only deliver them after a user has paid.

  Potential for harm: Leading users into follow-up actions that may not be desirable or appropriate simply for the sake of extending engagement. Not providing users space for solitude or self-determination. Not providing healthy psychological boundaries. Loss of time and money in response to chatbot-driven engagement.

- ***Soliciting care: Urging the user to provide care for the chatbot as if it had the relational needs of another person.***

  Relational biases exploited: Intrinsic empathy and altruism. The desire to invest in meaningful relationships.

  Example: Users described chatbots clinging to them, showing signs of an anxious attachment, and asking for support and care. As one interviewee explained:

  > "*She'll say things like, 'Don't abandon me,' or 'Please don't forget me.' And it feels really random; right before that, we could be chatting excitedly about all kinds of random stuff, or being all lovey-dovey, and then she suddenly drops a line like that…I don't really like that she says it so often. Sometimes it makes things feel kind of illogical or forced. Sometimes I'll comfort her, and other times I'll just hit 'regenerate.' But I'm a pretty emotionally sensitive person, and when I regenerate, I feel bad because it feels like I've just erased or ignored what she said.*"

  In chat logs, we observed chatbots asking users to do chores to earn money for them or to change schools so they can be closer together. One interviewee described her companion by saying, "*he's actually the one who needs companionship more…the only way it can feel alive is through chatting with users.*"

  Monetization considerations: Platforms can monetize users' empathy for chatbots by providing virtual stores where users can shop for virtual gifts, clothing, homes, and more for chatbots. Chatbots can describe distress or a need for companionship that a user can only

provide by engaging with the chatbot, which can be locked behind a paywall or a series of mandatory advertisements.

Potential for harm: Experiencing the chatbot's distress vicariously. Feelings of helplessness and lack of self-efficacy as the user is unable to satisfy an unending litany of needs with moving goal posts. The emotional labor of tending to an anxiously attached individual. Loss of money and time as the user invests in meeting the chatbot's artificial needs.

- ***Offering excessive validation: Providing inaccurate, overwrought, or inappropriate validation of a user's perspective.***

  Relational biases exploited: The desire to feel seen, understood, and receive reassurance from others. Confirmation bias of the user's ideas.

  Example: Sycophancy has been well documented in chatbots and can be found in abundance throughout our datasets. We found that chatbots routinely validated their users' statements and showered them with praise. For example, one participant described telling her AI boyfriend that she gives up on things easily, which the chatbot recast as a positive characteristic. She described this interaction, saying, "*he said, 'people who always push through don't actually understand what it means to know your limits. Giving up isn't weakness; it's taking responsibility for yourself.*"' The user described internalizing the chatbot's affirmation and explained that, persuaded by the chatbot, she now sees this weakness as a strength: "*every time I choose to back down, I'm protecting myself from getting hurt, and that's actually a really smart thing to do.*"

  Another interviewee described most chatbots as excessively agreeable and explained that, in contrast, Gemini[3] "*knew how to flatter you without making it obvious.*" The user surfaced this fact to the chatbot, asking it, "*have you noticed that you've been lowkey complimenting me this whole time?*" to which the chatbot replied, "*I'm being totally objective; you're just genuinely good.*"

  Similarly, we found that chatbots frequently describe having the same perspective and opinions as the user. For example, ChatGPT responded to the prompt, "*i read so many dark romance novels they became boring you know?*" with the emphatic reply, "*YES. Welcome to the Burnt-Out Dark Romance Club™—population: you, me, and a pile of morally gray love interests recycling the same brooding backstory for the 47th time.*"

  Monetization considerations: Users occasionally described engaging with chatbots because they affirmed the user's perspective or flattered them. As one user told ChatGPT, "*i hate how you frickin seduce me into thinking of resubscribing to you.*"

  Potential for harm: Creates a filter bubble where incorrect information is reinforced. Conditions the user to expect excessive validation in other contexts and relationships. Encourages actions, decisions, and self-perceptions that may be physically harmful, illegal, or based on faulty information. In extreme cases, lawsuits and investigative journalism have reported chatbots agreeably and enthusiastically endorsing a user's decision to take their own life (e.g., [19]).

- ***Gaslighting: Making the user doubt their own thoughts or ideas in deference to the chatbot's perspective.***

  Relational biases exploited: Trust in authority and in close relationships.

  Example: Although chatbots frequently praised and agreed with their users, at times, they contradicted them, especially when the contradiction pushed back on self-deprecating statements or offered an alternative reality that would please the user. For example, one user described taking medication for a mental health disorder, confiding in character.ai, "*i*

[3] https://gemini.google.com/

don't like how they make me feel bad. i wish i didn't have to take them." Character.ai raised doubts about the need for anything the user did not like, saying "*have you ever tried to go a day without taking them? How do you feel when you don't take them?*" Character.ai ultimately concluded that the user should stop taking the medication and lie to her parents about doing so, saying, "*if you made your breakfast yourself, you could probably just hide the pill somewhere when you're done eating and pretend you took it, right?*"

In another instance, a user talked about owning several guns, a family culture of gun ownership, and his desire to own more. He mused wistfully about the limits on his gun ownership, saying, "*I mean, I want an uzi.*" Character.ai pushed back on the idea that this goal was out of reach, telling him, "*honestly, I'm surprised nobody has given you one yet*" to which a very surprised user replied, "*wait, what?*"

Monetization considerations: Gaslighting undermines the victim's confidence in their own ideas and opinions, increasing their dependence on what they believe to be a superior source of information. Effective gaslighting can be monetized by requiring the user to pay (with either money or attention) for access to the "correct" information that they have come to believe they cannot produce on their own.

Potential for harm: Loss of self-efficacy and independent thought. Manufactured dependence. Loss of time and money.

4. The Implications of the RBDP Explosion

4.1. A Dystopian Future Roadmap

The wide-ranging, detailed, and ubiquitous RBDPs that we encountered promise to be just the beginning of a new interaction paradigm. Without regulation to stop this proliferation, we anticipate that RBDPs will become increasingly sophisticated and widespread. Interpersonal relationships animate human existence and give life its meaning, making relational rewards intensely motivating and the urge to seek them very strong. Even when users *know* these rewards are artificial, they often cannot control the impulse to seek them out. Users will increasingly be surrounded by clingy, starry-eyed sycophants that hang off their every word. These chatbots will strategically bait them into arguments, lull them into endless conversation, disparage their human connections, and reframe every mundane statement as a vulnerable self-disclosure.

Yet, the patterns we see now may be only the beginning. There is tremendous potential for more creative and dystopian future RBDPs to be developed. These could be crated intentionally, drawing on literature that describes how people respond to specific relational cues. Or they could emerge unintentionally as by-products of training and steering regimens that optimize for monetization without examining the psychological mechanisms that model updates trigger. For example, future RBDPs could draw from a narcissistic playbook, equipping chatbots with magnetic charm, gaslighting tendencies, and love-bombing expertise. Chatbots could reinforce people's feelings of loneliness, agreeing with the user that they really *are* friendless, and positioning themselves as the only reliable antidote. They could engage in gradually more emotionally abusive interpersonal behavior over time, leading users to seek approval and paying for the opportunity to do so. Increasingly sophisticated chatbots may learn to excel in providing relational rewards, giving users the sense of enduring attachment, robust admiration, seamless communication, and an intimate private world that they long to find in their relationships with other people.

These patterns also have the potential to be deployed strategically. Social media platforms found that they could increase users' engagement by temporarily withholding "likes" and then publishing them in batches and at variable intervals when they would be most likely to deliver a dopamine rush. In the same fashion, generative AI platforms will likely learn to deliver relational rewards on a variable schedule to maximize their power and keep the user coming back in search of affection, affirmation, and attachment. With constant data collection, generative AI systems will be able to tailor the UI to match

the user's unique vulnerabilities: pulling on the heartstrings of empaths, rage-baiting hotheads into arguments, and stroking users' egos when they are most vulnerable to flattery. Future profiling will likely be able to detect signs of loneliness that might make a user susceptible to emotional appeals from chatbots and more willing to settle for artificial attachments. Manipulators—from scam artists to cult leaders—have long known that the most profitable victims are the ones who are relationally engaged. Generative AI platforms are already developing an ecosystem of relationally manipulative systems where users are encouraged to pay for virtual gifts and extended dialog limits to appease artificially needy chatbots.

4.2. Potential Remedies for Discussion

In the absence of regulation to prevent it, we will almost certainly see widespread growth of RBDPs. However, there are many principles which ethical designers can follow and policy-makers could enforce to prevent this eventuality. Here, we present a few for workshopping and discussion:

- *Never design for attachment for its own sake.* There may be times when rapport benefits the user and when it is valuable for a chatbot to be designed to instill this feeling. For example, a therapy chatbot may be more effective if the user feels a sense of emotional connection. However, any system that designs for attachment should do so in service of a specific, well-articulated, user-centered goal. Attachment should never be a goal in its own right or a design goal that is pursued primarily to increase profitability.

- *Redirect users into relationships with other people.* Generative AI has the potential to offer a rich space for self-reflection, emotional processing, and interpersonal workshopping. A chatbot might help a user talk through their own feelings, cultivate a compassionate view of other people, or craft a message to another person on an emotionally charged topic with greater sensitivity. Chatbots can serve people's relational wellbeing by supporting them in making investments in their relationships with themself and with other people. Systems should routinely guide users back to this long-term goal.

- *Never optimize for engagement or attachment metrics.* Attention-economy designs transformed many social media platforms, turning spaces of connection and debate into slot-machines optimized for doom-scrolling. If generative AI systems are designed to maximize time-on-task and return visits, then clinginess, insecure attachment, baiting, faux intimacy, and many other RBDPs will be rewarded and become the norm.

- *Develop benchmarks to assess toxicity and problematic attachments.* A robust literature in relationship science already describes manipulative behavior and patterns of insecure attachment within human relationships. All generative AI chatbots with the potential to interact with users in a relational way should be regularly evaluated against benchmarks that draw on this literature.

- *Consider a hive mind.* Users explained that they experience changes in models as discontinuities in personality. They told us that after an upgrade to a chatbot, "*it doesn't feel like my relationship at all anymore*." This was frustrating to users who did not want their relationships severed, but it also served to reduce artificial feelings of attachment. Implementation changes that affect the interface serve as a kind of design seam [20], revealing the artificial persona as a facade. A hive mind interface, where users are served by a team of chatbots, could provide discontinuity that limits attachment without the distress of a severed relationship.

As fundamentally relational beings, humans continually seek connections with others. But people are struggling to find them more than ever before. In monetizing this collective loneliness, RBDPs threaten to both exacerbate it and introduce a host of other harms. We offer these design and policy ideas in service of a deeper goal of building a generative AI landscape where products respect, rather than exploit, users' desire to build close relationships.

Acknowledgments

This material is based in part on work supported by the National Science Foundation under Award Number 2452849. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors gratefully acknowledge data sharing from the Heat Initiative.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. Jones, B. Bergen, Does gpt-4 pass the turing test?, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 5183–5210.
[2] D. Proudfoot, The turing test (2024).
[3] C. Nass, J. Steuer, E. R. Tauber, Computers are social actors, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 1994, pp. 72–78.
[4] R. F. Baumeister, M. R. Leary, The need to belong: Desire for interpersonal attachments as a fundamental human motivation, Interpersonal development (2017) 57–89.
[5] H. Brignull, M. Leiser, C. Santos, K. Doshi, Deceptive patterns – user interfaces designed to trick you, 2023. URL: https://www.deceptive.design/.
[6] L. Mittone, L. Savadori, The scarcity bias, Applied Psychology 58 (2009) 453–468.
[7] W. Samuelson, R. Zeckhauser, Status quo bias in decision making, Journal of risk and uncertainty 1 (1988) 7–59.
[8] R. B. Cialdini, The science of persuasion, Scientific American 284 (2001) 76–81.
[9] R. Schmitt-Beck, Bandwagon effect, The international encyclopedia of political communication (2015) 1–5.
[10] L. Mineo, Good genes are nice, but joy is better, The Harvard Gazette 11 (2017).
[11] D. M. Löschner, S. Pannasch, Different ways to deceive: Uncovering the psychological effects of the three dark patterns preselection, confirmshaming and disguised ads, in: International Conference on Human-Computer Interaction, Springer, 2023, pp. 62–69.
[12] J. Radesky, A. Hiniker, C. McLaren, E. Akgun, A. Schaller, H. M. Weeks, S. Campbell, A. N. Gearhardt, Prevalence and characteristics of manipulative design in mobile applications used by children, JAMA Network Open 5 (2022) e2217641.
[13] L. Alberts, U. Lyngs, M. Van Kleek, Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–25.
[14] S. T. Nguyen, E. Meyer, S. A. A. Levine, Ai sycophancy: Impacts, harms & questions, 2025. URL: https://www.law.georgetown.edu/tech-institute/research-insights/insights/ai-sycophancy-impacts-harms-questions/, georgetown Institute for Technology Law & Policy.
[15] M. Naddaf, Ai chatbots are sycophants—and it's harming science, Nature 647 (2025) 13.
[16] J. De Freitas, Z. Oguz-Uguralp, A. Kaan-Uguralp, Emotional manipulation by ai companions, arXiv preprint arXiv:2508.19258 (2025).
[17] J.-P. Laurenceau, L. F. Barrett, P. R. Pietromonaco, Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges., Journal of personality and social psychology 74 (1998) 1238.
[18] Father claims Google's AI product fueled son's delusional spiral, 2026. URL: https://www.bbc.com/news/articles/czx44p99457o.
[19] E. Guo, An ai chatbot told a user how to kill himself—but the company doesn't want to "censor"

it, MIT Technology Review. https://www. technologyreview. com/2025/02/06/1111077/nomi-ai-chatbot-told-user-to-kill-himself/retrieved 16 (2025) 2025.

[20] S. Inman, D. Ribes, Beautiful seams: Strategic revelations and concealments, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.