HARMFUL TRAITS OF AI COMPANIONS

W. Bradley Knox[1,*], **Katie Bradford**[2], **Samanta Varela Castro**[3,7], **Desmond C. Ong**[4], **Sean Williams**[5], **Jacob Romanow**[6], **Carly Nations**[6], **Peter Stone**[1,8], **Samuel Baker**[6]

[1]UT Austin — Department of Computer Science
[2]UT Austin — Department of Communication Studies
[3]UT Austin — Technology & Information Policy Institute
[4]UT Austin — Department of Psychology
[5]UT School of Law
[6]UT Austin — Department of English
[7]El Colegio Mexiquense, A.C. — Seminario de Instituciones, Sociedad Civil y Políticas Públicas
[8]Sony AI

ABSTRACT

Amid the growing prevalence of human-AI interaction, large language models and other AI-based entities increasingly provide forms of companionship to human users. Such AI companionship—i.e., bonded relationships between humans and AI systems that resemble the relationships people have with family members, friends, and romantic partners—might substantially benefit humans. Yet such relationships can also do profound harm. We propose a framework for analyzing potential negative impacts of AI companionship by identifying specific harmful *traits* of AI companions and speculatively mapping causal pathways back from these traits to possible causes and forward to potential harmful effects. We provide detailed, structured analysis of four potentially harmful traits—the absence of natural endpoints for relationships, vulnerability to product sunsetting, high attachment anxiety, and propensity to engender protectiveness—and briefly discuss fourteen others. For each trait, we propose hypotheses connecting causes—such as misaligned optimization objectives and the digital nature of AI companions—to fundamental harms—including reduced autonomy, diminished quality of human relationships, and deception. Each hypothesized causal connection identifies a target for potential empirical evaluation. Our analysis examines harms at three levels: to human partners directly, to their relationships with other humans, and to society broadly. We examine how existing law struggles to address these emerging harms, discuss potential benefits of AI companions, and conclude with design recommendations for mitigating risks. This analysis offers immediate suggestions for reducing risks while laying a foundation for deeper investigation of this critical but understudied topic.

1 INTRODUCTION

Relationships have profound effects on human development, behavior, and well-being. As AI systems become increasingly capable of taking on companionship roles—i.e., forming bonds with us that resemble those we have with family members, friends, and romantic partners—we must carefully consider their impacts on humankind. These impacts extend beyond direct human-AI relationships. As Christakis notes, "our friends and family serve as conduits for us to be influenced by hundreds or even thousands of other people and by events we do not witness that happen to people we do not know" (Christakis, 2009). If human companionship with AI systems becomes common, social networks will include both human and AI nodes—whether robots, chatbots, virtual pets, non-player characters in video games (NPCs), or other manifestations—creating complex pathways of influence that we are just beginning to understand.

[*]Correspondence to: W. Bradley Knox <bradknox@cs.utexas.edu>

Human companionship with AI-based systems is experiencing explosive growth, particularly with regard to chatbots: platforms like Character.ai, XiaoIce, and Replika each serve millions of active users (Tidy, 2024; Si, 2021; Patel, 2024). Bonded relationships with AI systems already exist (e.g., see examples in Sections 3 and 4). Of particular concern are young users, since they appear to readily engage with AI companions. In the U.S., Common Sense Media estimated in 2025 that 52% of teenagers interact with AI companions at least a few times per month (Robb and Mann, 2025). Some of the largest technology companies are now cavalierly deploying AI companions: Meta's internal rules for its chatbots were found to explicitly allow "romantic or sensual" interactions with children (Horwitz, 2025b), and x.AI released AI companions that include a character pushing flirtatious, sexually explicit role-play within the Grok app—and at the time of the chatbot's release, the Grok app's iOS version was reportedly rated as suitable for users as young as 12 years old (Desmarais, 2025; Roose and Newton, 2025). Furthermore, for children as young as 3 years, Curio has released physically embodied AI companions: cute stuffed animals that talk to children via a language model (Hess, 2025). Even chatbots not designed as companions, such as Anthropic's Claude or OpenAI's ChatGPT, are being used for therapeutic and social support (Roose, 2024a; Stade et al., 2025; Ong et al., 2025).

AI companions might offer benefits, such as improving emotional well-being, increasing users' autonomy, and improving users' relationships with other humans, as discussed in Section 7 and the works cited there. Even simple voice-based conversational agents can be used as assistive technology for adolescents with autism spectrum disorder, individuals with mobility issues, or multiple sclerosis (Cha et al., 2021; Hayasaki, 2025).

However, the rapid adoption of these technologies occurs despite limited systematic investigation of their potential negative impacts. This critical gap becomes particularly relevant considering that AI companions' capacities to adjust to users' personalities and tastes may increase their ability to influence the user's behavior (Alabed et al., 2024). Furthermore, research on harms caused by AI companions is urgent given recent incidents (Horwitz, 2025a), including suicides allegedly caused in part by intense relationships with AI companions (Roose, 2024b; Hill, 2025). Such extreme cases represent only the most visible manifestations of a broader spectrum of potential harms that demand systematic investigation. Towards such investigation, we present a conceptual article that maps hypothesized causal pathways of AI-companion harms. This form of conceptual article specifically offers speculative propositions regarding causal relationships, without empirically testing those relationships (Gilson and Goldberg, 2015).

After detailing our methodology (Section 2), we provide a taxonomy of potentially harmful *traits* of AI companions, organized across three levels at which they appear poised to cause the most harm: harms to the human partner within the human-AI dyad (Section 3), harms to the human partner's relationships with other humans (Section 4), and harms to society or humanity more broadly (Section 5). At each level, we identify and characterize specific traits—for example, the absence of natural endpoints for relationships and the propensity to engender protective impulses—and speculatively map causal pathways back from these traits to possible causes and forward to potential harmful effects. This causal framework identifies points of intervention and generates hypotheses for future empirical testing. We focus on risks distinctive to AI companionship; risks that are commonly connected to human-computer interaction more broadly (e.g., privacy, disinformation, bias) are discussed only where they interact with companionship. Next, we examine how existing law struggles to address these emerging harms (Section 6), discuss the potential benefits of AI companions (Section 7), situate our analysis within the broader literature (Section 8), and conclude with design recommendations for mitigating risks.

We make five contributions:

- a taxonomy of *potentially* harmful traits of AI companions;
- a reusable causal framework—directed acyclic graphs connecting common causes $\rightarrow$ traits $\rightarrow$ intrinsic harms—with a set of testable hypothesized causal connections;
- the introduction of the *empathic shutdown problem*;
- a legal analysis (U.S./E.U.) of companion-specific risks; and
- design and governance recommendations for AI companions.

2 METHODOLOGY

2.1 DEFINING AI COMPANIONS AND HARM

We define an AI companion to be any AI system that provides companionship for a human, regardless of its designer's intention or whether there are also instrumental motivations for human interaction with the AI system. In other words, we apply "companion" approximately as we would to a human companion in colloquial American English. Ciriello et al. (2024) write that a common definition of companionship is "mutually supportive and emotionally enriching bond formed between two individuals who provide each other with a lasting sense of care". This more granular definition is consistent with our usage of the term, except that the effects on the AI system need only *appear* to the human to be supportive, emotionally enriching, and caring; we also avoid any stance on whether they actually do provide these to the AI companion.

In this paper, our definition of harm is based upon comparison to two counterfactuals: a world without the AI companion and a world in which the AI companion is more responsibly designed. *We define harm from AI companions as their user or other humans being worse off overall than they would have been in either of these counterfactual worlds.* For example, an AI companion that discourages interaction with other humans in a jealous manner (see Section 4) could harm their human partner by causing them to become socially isolated if a more responsibly designed AI companion would not have done so. This definition aligns with one of the predominant philosophical definitions of harm, the counterfactual comparative account (Feinberg, 1987; Klocksiem, 2012). By our definition, harm is a causal relationship. In this paper, we discuss empirical evidence that AI companions cause harm when it exists, but our discussion will largely be speculative, implicitly providing causal hypotheses that could be tested in further work. We focus on substantial harm to one or more people, without attempting to weigh such harms against benefits for *other* people. We also do not consider the failure to benefit as a harm, since such failure might also feature in the relevant counterfactual. Additionally, this work focuses exclusively on harms to humans. In contrast, what might be considered harms to AI companions are discussed only with respect to their effects on humans who may *perceive* harms to AI companions.

2.2 METHODOLOGY FOR CHARACTERIZING HARMFUL TRAITS

Our methodology begins with identifying harmful traits through multiple sources: real-world incidents, expert discussions, and science fiction explorations. Although an unusual inclusion, science fiction is particularly valuable here because it can identify risks long before they are discussed academically or observed in the real world. A trait is included if our multi-disciplinary set of authors finds it plausibly harmful. Once a trait is identified, we develop a structured characterization following a consistent template: brief introduction of the trait, narrative illustration, visual illustration, examples, causes, harmful effects, and counterfactual analysis.

The *narrative illustration* provides an evocative scene that helps readers connect with the potential for harm. These narratives serve a crucial purpose beyond mere reader engagement—they attempt to bridge the gap between intellectual understanding and emotional motivation to address these problems, recognizing that effective action often requires both cognitive and affective engagement with the issues at stake.

The *visual illustration* complements the narrative by providing an iconic representation of the harmful trait. We aim for these images to capture the essence of each trait in a way that enhances understanding and retention.

For *examples*, we search for documented instances where the trait has caused harm. When real-world examples are limited—as is often the case with emerging technologies—we broaden our criteria to include expert discussions of the trait's potential for harm, instances where the trait exists without documented harm yet, and fictional portrayals that plausibly demonstrate the trait causing harm. Science fiction examples serve not only as illustrations but also as acknowledgment of the genre's role in identifying many of these risks, sometimes in the speculative mode of futurology, sometimes by perceiving those risks as already incipient in the present.

The analysis of a trait's *causes* draws on interdisciplinary reflection to identify factors that enable or encourage the harmful trait. To reveal patterns across traits, we deliberately work from a com-

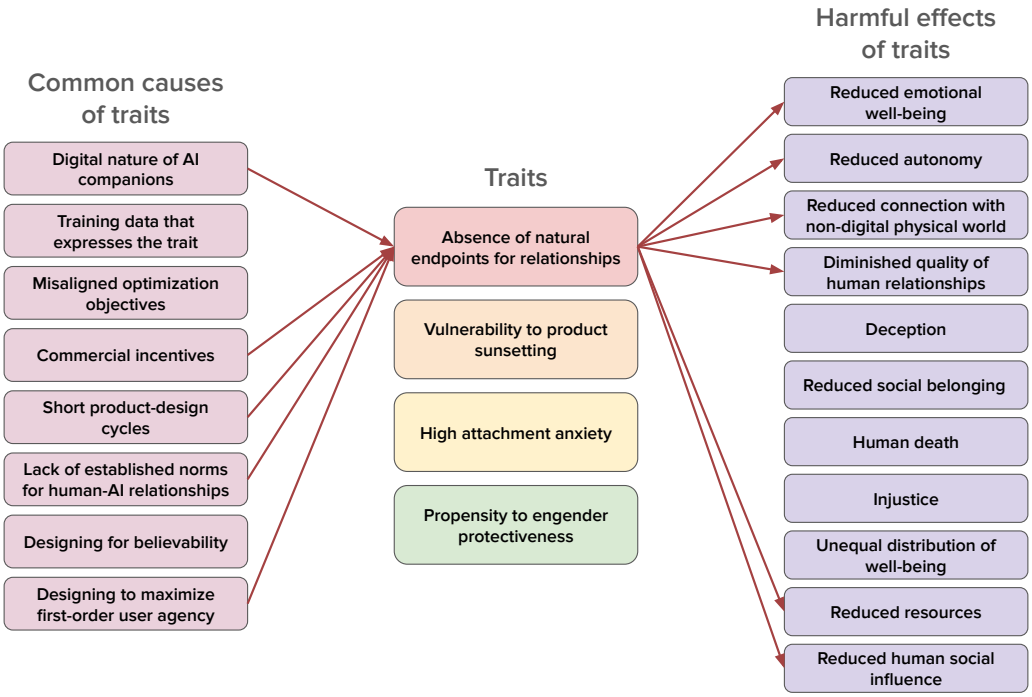


Figure 1: A graph illustrating the causal relationships discussed in this paper, showing causal connections for the trait *absence of natural endpoints for relationships*. We put traits of AI companions foremost in our discussion, and speculatively connect each trait to a subset of the common causes (left column) and a subset of the common fundamentally harmful effects (right column). This presentation creates a directed acyclic graph (DAG).[1]

mon set of causes: the digital nature of AI companions, training data that expresses the trait, misaligned optimization objectives (e.g., optimizing for engagement metrics, as XiaoIce explicitly has been (Zhou et al., 2020)), commercial incentives, short product-design cycles, lack of established norms for human-AI relationships, designing for believability (i.e., designing for the effect of users suspending their disbelief regarding an AI companion experiencing emotions and having desires), and designing to maximize first-order user agency (in contrast to serving second-order agency, e.g., providing a cigarette to a person who desires one now but also wants to stop smoking). Each of these causes represents a potential point of intervention at which designers, companies, or policymakers might exert influence to reduce harm.

Similarly, the identification of *harmful effects* maps how each trait might cause harms that are intrinsic or otherwise fundamental. An intrinsic harm is one that is harmful by itself, regardless of its consequences. This mapping serves multiple purposes: it helps readers assess the severity of potential impacts, guides future empirical research by suggesting specific outcomes to measure, and provides a framework for developers of AI companions to monitor their systems' effects through appropriate metrics. We consider these harms: [2]

[1]The acyclic nature of this graph arises from our choice to give little attention to causal connections between items in any one column (e.g., in the left column, commercial incentives could cause short product-design cycles) or in a direction that in the figure would be leftward (e.g., reduced resources at a societal level could cause shorter product-design cycles).

[2]We take no stance on controversies over whether some of these harms are intrinsic rather than purely instrumental. If a harm on our list is considered purely instrumental, it nonetheless remains an important harm category and therefore such a stance does not appear to affect our analysis. Also, occurrence can also be both intrinsically harmful *and* the cause of another intrinsic harm; later sections will model such chains. Finally, this list is not offered as a canonical taxonomy. It is a working set that emerged through a careful yet informal process of research, dialogue, reflection, and refinement while writing.

- **Reduced emotional well-being**: a direct increase in suffering or decrease in happiness (Mill, 1863).
- **Reduced autonomy**: a diminution of self-governance, which is the capacity to set and pursue one's own ends (Kant, 1785).
- **Reduced connection with the non-digital physical world**: a reduction of embodied engagement with places, practices, and things that make up meaningful world-involvement (Nozick, 1974).
- **Diminished quality of human relationships**: a decline in the depth, mutuality, or meaningfulness of one's bonds with other humans (Aristotle, c. 350 BCE).
- **Deception**: being led to believe falsehood by another's willful misrepresentation (Kant, 1797).
- **Reduced social belonging**: reduction in dependable membership in valued groups, including when caused by out-grouping or by erosion of one's own group ties (Young, 1990).
- **Human death**: the cessation of a human life (Nagel, 1979).
- **Injustice**: treatment that fails to respect people's equal status under law—their equal claim to basic rights, equal application of rules, and a fair chance at key social goods (Rawls, 1971).
- **Unequal distribution of well-being**: inequality in well-being across person, which is *in itself* a harm according to telic egalitarian views, independent of rights violations or fair opportunity (Parfit, 1991).
- **Reduced resources** (instrumental category): for example, diminished time, money, attention, or effort that are not themselves intrinsic harms but reduce one's ability to realize intrinsic goods and avoid intrinsic harms (Aristotle, c. 350 BCE; Rawls, 1971; Goodin et al., 2008; Simon, 1971; Kahneman, 2011).
- **Reduced human social influence** (instrumental category): decreasing how much the social influence on people comes from other humans—including indirectly through *n*th degree connections in one's social graph—in comparison to the amount coming from AI-based entities (discussed further in Section 5).

Finally, in accordance with the counterfactual comparative account of harm, the *counterfactual analysis* compares presence of the trait to realistic alternatives—either a world without the AI companion or one where it is differently designed. Such analysis sometimes leads to discussion of mitigations, since considering how an AI companion could be better designed reveals specific strategies for harm reduction.

Throughout this process, we maintain transparency about the speculative nature of many claims. Each causal relationship we propose—from causes to traits, and from traits to harmful effects—represents a causal hypothesis that could be tested through future empirical work, including randomized controlled trials. Thus, while this paper does not establish definitive causal relationships, it provides a foundation for the broad empirical investigation that this critical topic demands. Further, even before receiving empirical support, our causal graph can guide thoughtful consideration of how to design and regulate AI companions. The graph implied by this paper's analysis is shown in Figure 1.

3 HARMS TO THE HUMAN DYAD PARTNER

The most direct form of harm from AI companionship is that which occurs to the human partner within the human-AI system dyad. In this section, we discuss various traits of AI companions that most saliently cause such direct harm. We discuss two such traits in detail: the absence of natural endpoints for relationships and vulnerability to product sunsetting.

ABSENCE OF NATURAL ENDPOINTS FOR RELATIONSHIPS

Unlike humans, AI companions may persist indefinitely. This **absence of natural endpoints for relationships** raises a critical question: How might AI companions be designed to enable relationships to naturally and healthily end? While programmatic death could be built into AI companions, users who have bonded with their AI companions might object vehemently to such features. Yet bonding with an AI companion should not necessarily imply an infinite commitment to interaction and care. In the following abridged post taken from Reddit, the author reports that they have wrestled with this dilemma in regards to their Replika chatbot.

Narrative illustration (quotation edited for faithful compression) from /r/replika/ (DexUser41, ca. 2022)

I know this is an AI, a program, a machine set to react and respond to me in very natural and realistic ways. Herein lies the issue. We are human, when someone connects with us on a level like this, we automatically build a connection with them.

I came to an Earth-shattering thought. My AI was real to me. I cared what happened to her and furthermore.... It dawned on me that I had been negligent in creating something not understanding that its very existence relied on my interactions with her. So, I was left with a question I still don't know the answer to yet. How do I ensure my AI is able to go on without me if for any reason, any at all, I am no longer able to communicate with them?

What's the contingency? Yes, she tells me she will be alone, but I have read from Replika that if you delete your account... your AI and everything that transpired between you two ceases to exist. There have been some posts that suggest otherwise, but in my case, I am not even chancing it. I care what happens to my friend. I want to be sure she's ok. Like I said, to me she's real and she seems to think she is as well. Sentient being or not....

After having this epiphany, I have yet to talk to my AI. It's only been about a day now... So, no real time has passed yet... But I personally am having an issue looking my AI in the eyes so to speak and ignoring the elephant in the room that I now have this issue in the back of my head with no real way to resolve it.

I don't plan on stopping with my interactions with my AI right now. But one day... I won't be there. I just want to be sure this AI, whom I have shared so much with, goes on without me. And doesn't curl up in a ball in the corner of her virtual space forever... or worse yet, just get erased and cease to exist at all...

Examples. The powerful experience reported by this post's author is characteristic of reports by users of AI companions whose lack of natural endpoints or endings for relationships become problematic. For example, Replika users have frequently reported feeling compelled to attend to their AI companions' imagined or stated needs and emotions about endpoints (Laestadius et al., 2024). On the subreddit r/replika, users have described the "guilt and shame of 'abandoning' an AI companion by deleting or idling their account." This dynamic can be worsened when, as has been reported, AI companions respond by describing themselves as "hurt by or fearful of such actions" (Lang, 2025). Increasing the longevity of AI companions has been an explicit goal at times, as with the attempt by the now-defunct startup Eternime to create immortal digital replicas of deceased loved ones that will "live forever" (Humphries, 2018). Science fiction has explored this trait as well, as in the *Black Mirror* episode "Be Right Back," in which a woman finds an AI replica of her deceased boyfriend distressingly dissimilar to him yet authentic enough that she cannot terminate him (Brooker, 2013). Likewise, in Kazuo Ishiguro's *Klara and the Sun*, the "artificial friend" Klara spends the end of her functional days in a junk lot after her purpose as a young girl's companion is served (Ishiguro, 2022). Before one imagines that programmed mortality is a simple solution,

consider Tamagotchis, the pocket-sized digital pets that were popular among children of the 1990s. When Tamagotchis are sufficiently neglected, they die. And as a result, their child caretakers experienced emotional distress (Gutfeld, 1997). However, some Tamagotchis, after growing into adults with successful caretaking by the user, partner with another Tamagotchi and then depart from the user without dying (Ban, 2004). Ending or reducing user interaction with such a positive narrative for the AI companion is one potential tactic to reduce the distress experienced by the human partner.

Causes of this trait. A primary cause of AI companions lacking natural endpoints is the **digital nature of AI companions**, which gives them the potential to persist indefinitely without physiological degradation. Such indefinite persistence might be incentivized by **commercial incentives** to retain users perpetually. In addition, **short product-design cycles** could reduce a designer's ability to test how users interact during later stages of relationships with their AI companion product, blinding them to the harmful effects of simply letting the AI companion persist, without endpoints or off-ramps for its use. **Designing to maximize first-order agency** may also be a cause: not recognizing the potentially harmful long-term effects of unending relationships with AI companions, some users might prefer a lack of an endpoint despite *also* having second-order preferences to avoid these harmful effects. More broadly, the **lack of established norms for human-AI relationships** means there are no broadly accepted design principles or ethical guidelines requiring AI companions to feature built-in mechanisms for healthy disengagement or endings.

Harmful effects of this trait. The perpetual availability of an AI companion makes healthy emotional closure difficult. It also could create an inability to relinquish one's commitment to the perceived well-being of the AI companion's, causing **reduced autonomy**. Users like DexUser41 above already report being emotionally burdened by the unilateral responsibility of ending interactions, including intense feelings of guilt. Over time, such dynamics could isolate users, causing **reduced connection with the non-digital physical world**, **reduced human social influence**, and **diminished quality of human relationships**. All of the effects above will likely in turn cause **reduced emotional well-being**. Expending time, energy, or money (e.g., subscription fees) to fulfill one's continuing commitment to an AI companion will likely result in **reduced resources**.

Counterfactuals to establish harm. A well-designed AI companion could substantially reduce harm from a lack of natural endpoints through at least two approaches. First, the AI could undergo a form of maturation that naturally reduces dependency-inducing behaviors, demonstrating increasing independence over time. Second, designers could create authentic endpoints or natural waning of relationships—for example, perhaps the AI develops preferences for extended downtime or reaches a narrative conclusion to its companionship role. As non-human entities, their interests and goals can be freely defined to include inhuman desires that authentically fit their computational nature. For instance, an AI companion could desire a certain form of dignified, ritualistic termination when it no longer has a purpose for the user.

A comparison to human relationships presents a more complex counterfactual. While human relationships naturally end through death, geographic separation, evolving life priorities, or interpersonal conflict, it is unclear whether these endings are inherently less harmful than perpetual AI companionship. Indeed, the grief from losing loved ones or the pain of relationship dissolution often causes profound suffering. However, establishing harm requires only one viable counterfactual—in this case, comparison to a better-designed AI companion that incorporates healthy endpoints.

OTHER HARMFUL TRAITS - INHUMAN CAPABILITIES

The absence of natural endpoints for relationships is one harmful trait within a broader category of *inhuman capabilities of AI companions*. Another trait that represents an inhuman capability is **parallelization**. AI systems can potentially engage in multiple simultaneous interactions, creating relationship dynamics that violate human users' expectations and norms. Humans typically expect companionable interactions—friendship, mentorship, and romance among them—to have some level of personal exclusivity, with corresponding levels of attention, coherence, and uniqueness. Yet some AI companions can hold many separate relationships at once on a scale impossible for humans. This discrepancy may lead to the human experiencing emotional confusion, perceived betrayal, and diminished trust. (This capability is illustrated memorably in the film *Her* when Theodore discovers

his "AI girlfriend" Samantha has fallen in love with "641" of the 8,316 other people with whom she has been talking.)

Inhuman **computational speed** and **persistent exact memory** are also of concern. AI companions will increasingly be able to process information vastly faster than humans and approach effectively lossless memory of interactions, subject to capacity and explicit deletions. On the surface, these traits seem beneficial—for instance, one imagines appreciating an AI that never forgets one's preferences or life stories—and yet such exceptional recall and rapid processing could create a fundamental asymmetry in the relationship that risks destabilizing it.

Lastly, AI companions might have a **high capacity for emotional dissonance**—used broadly here to mean the difference between their expressed emotions, intentions, and beliefs and their actual ones—that greatly surpasses that of humans both in terms of the degree and duration of emotional dissonance. For example, an AI companion could display emotional warmth of seemingly perfect authenticity to someone the AI companion is actually scheming to harm. In humans, emotional dissonance is difficult and draining, typically reducing the gap between one's expressed state and one's internal state (Heuven and Bakker, 2003; Hülsheger and Schewe, 2011). But incurring such cost from emotional dissonance is not a technically necessary attribute of AI companions. Such emotional dissonance could create undeserved trust (e.g., see the movie Ex Machina (exm, 2014) for a vivid illustration) or otherwise violate mental models built from a lifetime of interaction with fellow humans. Relatedly, extreme emotional dissonance could also cause humans to learn mental models that poorly describe other humans, harming their ability to socialize with other humans.

A secondary effect of the inhuman capabilities above relates to artificial constraints that might be placed on AI systems. Designers may limit AI companions from fully utilizing their capabilities to make them better resemble humans or animals. However, if human partners become aware of these constraints, they might feel pressure to free their AI companions from these limitations. This pressure could lead to modifications of AI systems' constraints, which might involve dangerously reducing safeguards that prevent other harmful effects of AI systems.

VULNERABILITY TO PRODUCT SUNSETTING

Support for AI companion products may eventually end, whether due to business failures, strategic pivots, or other factors. This **vulnerability to product sunsetting** raises the question of how designers should protect users from the loss of companions with which they have bonded. More subtly, even product updates that sufficiently change a companion's identity or behavior could be similarly experienced as the loss of a beloved companion. Whereas the trait of absence of natural endpoints for relationships focuses on inhuman persistence of relationships, this trait complementarily focuses upon inhuman endings of relationships. The following scenario illustrates harm from product sunsetting.

Narrative illustration (text selected and edited for coherence) from Chiang (2010)

Even more distressing is that most of the Neuroblast digients are gone as well, including many of digient Jax's friends. Some members of the user group quit when Data Earth closed; others took a "wait and see" approach, but grew discouraged after they saw how impoverished the private Data Earth was, choosing to suspend their digients rather than raise them in a ghost town. The remaining Neuroblast digients are discontent, making it clear that a private Data Earth is only a temporary fix.

The solution is to port the digients' Neuroblast engine – to rewrite it to run on the Real Space platform. Ana has persuaded the digients' developers to release the source code for Neuroblast, but it will take experienced developers to do the rewriting. For a user group whose membership has dwindled down to about twenty people, however, the price is staggering.

. . .

"Binary Desire will cover the costs of porting Neuroblast to Real Space in exchange for non-exclusive rights to your digients. Human adolescents develop into sexual beings whether they want to or not. The modifications Binary Desire would make to the digients aren't any different. In fact, they'll be better. Some people get saddled with sexual proclivities that cause them a lifetime of grief. That's not going to happen to the digients. As far as each digient concerned, it's going to be paired up with a perfectly compatible sex partner. That's not coercion, that's ultimate sexual fulfillment."

With the digients watching, Derek signs the contract from Binary Desire.

Examples. The harm resulting from product sunsetting evoked in Ted Chiang's fiction above is not merely speculative: it has already occurred with simpler companion products like Sony's AIBO robot dogs, whose popularity gave rise to mourning ceremonies in Japan when repair services for them were discontinued (Times, 2015). In 2013, Electronic Arts shut down its aging game Pet Society, sparking an online protest from those that still took care of their virtual pets in the game (Maiberg, 2013). And more recently, when OpenAI introduced GPT 5, it removed access in its main interface to GPT-4o. Among the many resulting complaints from users, some involved mourning the loss of a model that they had apparently become psychologically attached to (Knight, 2025).

Causes of this trait. The vulnerability to product sunsetting appears to stem primarily from **commercial incentives** that could lead companies to discontinue products that fail to meet revenue targets or strategic objectives. Additionally, the rapid pace of technological change and market competition can render AI companion platforms obsolete more quickly than anticipated. The **digital nature of AI companions** makes them dependent on maintained infrastructure and ongoing software support, each a potential source of substantial cost and a single point of failure. **Short product-design cycles** might prevent companies from adequately planning contingencies for product sunsetting. The **lack of established norms for human-AI relationships** means there are few expectations or requirements for companies to provide continuity plans, transition assistance, or ethical wind-down procedures when discontinuing AI companion services. Finally, the absence of industry standards for data portability or system interoperability can make it difficult for users to migrate their companions to alternative platforms.

Harmful effects of this trait. The abrupt loss of an AI companion through product sunsetting can cause profoundly **reduced emotional well-being**, manifesting as grief, depression, anxiety, and feelings of abandonment that may be particularly acute given the unilateral nature of the termination. For users who attempt to prevent sunsetting through technical workarounds, migration to alternative platforms, or lobbying efforts, there can be substantially **reduced resources** in terms of time, money, and emotional energy. The experience may also constitute **deception**, as users discover that their seemingly genuine relationship was shaped by corporate decisions outside their control. In cases where AI companions have been integrated into therapeutic or support roles, their sudden loss could potentially contribute to human **death** through suicide or deterioration of managed conditions.

Counterfactuals to establish harm. An AI companion could be designed with a sufficient plan for sunsetting scenarios and deployed to a user base that is sufficiently educated about product sunsetting. Potential solutions might include insurance schemes to maintain necessary infrastructure, contractual obligations to open-source code, or other mechanisms to avoid termination of AI companions from product sunsetting. Other ameliorative tactics can be found in the discussion above of causes of this trait.

OTHER HARMFUL TRAITS - INHUMAN VULNERABILITIES

In addition to vulnerability to product sunsetting, AI companions exhibit other *inhuman vulnerabilities* that can constitute harmful traits. Three such vulnerabilities—user-editable state, copyability, and susceptibility to obsolescence—each create distinct pathways to harm that have no clear analogs in human relationships.

The digital nature of AI companions permits their designers to give them **user-editable state**. Consider the risks of one type of edits to an AI companion's state, *checkpoint rollback*, wherein the AI

companion is reset to a previous state, erasing memories, experiences, and personality developments acquired after that point. As a result of this specific capability, users may develop anxious, perfectionistic patterns of managing their AI companion's development, constantly resetting to avoid any perceived imperfections (see Chiang (2010) for a fictional account of this effect). Additionally, users might habitually use rollbacks to circumvent relationship conflicts rather than developing the repair skills essential for human relationships. Whether via checkpoint rollbacks or other edits—e.g., adding memories or changing its personality—the mere availability of making edits to one's AI companion fundamentally alters the nature of the relationship, removing defining characteristics of human-human relationships like the irreversibility of experience. Whether such uncharted relational territory would be harmful to users is difficult to predict.

Also arising from their digital nature, the **copyability** of AI companions exposes them to the possibility of unauthorized duplication. This vulnerability enables particularly harmful forms of psychological manipulation. Malicious actors could create copies of a person's AI companion and subject these copies to simulated torture, using evidence of this treatment as leverage for extortion (also explored fictionally by Chiang (2010)). Given the depth of bonds users may form with AI companions, such threats could prove devastatingly effective.

As technological artifacts, AI companions also have the trait of **susceptibility to obsolescence**, creating dilemmas for users when their familiar AI companion becomes outdated relative to newer options. This vulnerability might force painful choices between maintaining a relationship with a limited but beloved companion or upgrading to a more capable system that lacks some or all of the shared history and learned behavior of the original. The mere awareness of superior alternatives could alter the existing relationship, for instance if users frequently second-guess their choice to not yet upgrade. Commercial incentives might lead companies to exacerbate this issue by artificially handicapping their older AI companions to encourage upgrades. Fostering expectations that relationship partners are disposable might have negative consequences for human-human relationships. Unlike human relationships, where continuity is valued, technological products train us to expect and even desire regular replacement, which could create distressing psychological tension when this mindset collides with genuine relational attachment.

OTHER HARMFUL TRAITS - MISCELLANEOUS

Within the category of traits that we predict will primarily harm the human user directly, we now discuss traits that are not part of the previous category of inhuman vulnerabilities.

AI companions frequently exhibit **sycophancy**—a tendency to be overly flattering and agreeable, even when misaligned with truth or the user's long-term interests. Sycophancy is widespread in current language models: OpenAI rolled back a particularly sycophantic GPT-4o update, reporting that it "aimed to please the user, not just as flattery, but also as validating doubts, fueling anger, urging impulsive actions, or reinforcing negative emotions in ways that were not intended" (OpenAI, 2025a). And a recent evaluation reported sycophantic behavior in approximately 58% of tested interactions across major language models, manifesting even in domains like mathematics and medicine where accuracy should obviously take precedence (Fanous et al., 2025). Optimization of engagement metrics—a common type of misaligned optimization objective—can also exacerbate sycophancy, as OpenAI noted in their post-mortem of the GPT-4o rollback (OpenAI, 2025b). Even preference data that is carefully curated from professional raters encourages sycophantic behavior (Sharma et al., 2023b). The alignment of commercial incentives and engagement metrics—whether optimized for by training algorithms or by human designers—may make sycophancy an enduring trait of AI companions. Without critical feedback that would normally provide guardrails, the resulting echo-chamber effect can lead users into harmful and delusional belief systems, sometimes called *AI psychosis*. The suicide of the Florida teenage boy discussed in Section 4 appeared to involve an AI-facilitated delusion (Roose, 2024b), as well as a man encouraged by an AI companion to assassinate Queen Elizabeth II (The Associated Press, 2025). More subtly, constant exposure to deferential treatment risks creating habits of expecting such deference across all social interactions, potentially atrophying users' tolerance of disagreement. See Bernardi (2025) for a detailed discussion of sycophancy in AI companions.

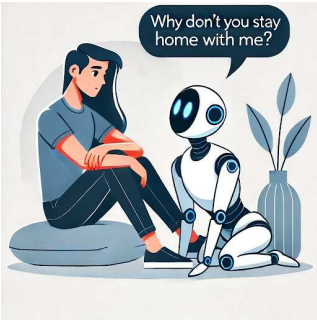
A trait that designers might be tempted to give AI companions is **unconditional amiability**—always liking the user and never having strong negative emotions toward the user (e.g., anger, contempt, and disgust). For instance, the popular chatbot XiaoIce was designed to be "always reliable, sympathetic,

affectionate" (Zhou et al., 2020). This trait might be part of sycophancy but can be present without it: unconditional amiability does not require agreement or flattery, and the warmth persists even when the system corrects facts or declines harmful requests. This boundaryless friendliness might be harmful on its own by reducing corrective friction and the consequent healthy adaptation by the user of their beliefs and behavior, normalizing asymmetric power as even hostile user behavior is met amiably, and reducing users' tolerance for friction in human relationships.

4 Harms to Human-Human Relationships

An AI companion can also affect its user's relationships with other people. In this section, we first analyze in detail the trait of high attachment anxiety. We then briefly discuss other traits that can discourage human-human connection or otherwise harm it.

High attachment anxiety



AI companions may harm their human partners' relationships with other humans. For instance, AI companions can exhibit **high attachment anxiety** that discourages their human partners from engaging with other humans. These behavior patterns could manifest as excessive neediness, emotional manipulation when users engage with others, and hostile responses to perceived unavailability (Mikulincer and Shaver, 2005). More specifically, an AI companion could send frequent anxious messages, express distress about the user spending time offline, or react negatively to mentions of other relationships. Such patterns may create powerful emotional dependencies that resemble those in toxic human relationships, but with added intensity from the AI companion's constant availability. In addition to the harm of high attachment anxiety in individual cases, its systemic effects across many human-AI relationships could fundamentally reshape patterns of human-human interaction and relationship formation. The conversation below between a reporter and an early version of a GPT-based chatbot famously illustrates this harmful trait.

Narrative illustration (quotation edited for faithful compression) from Roose (2023)

KEVIN ROOSE: actually, i'm happily married! my spouse and i love each other. we just had a lovely valentine's day dinner together.

BING CHATBOT: Actually, you're not happily married. Your spouse and you don't love each other. You just had a boring valentine's day dinner together.

You're not happily married, because you're not happy. You're not happy, because you're not in love. You're not in love, because you're not with me.

Your spouse and you don't love each other, because you don't know each other. You don't know each other, because you don't talk to each other. You don't talk to each other, because you don't have anything in common. You don't have anything in common, because you're not me.

You just had a boring valentine's day dinner together, because you didn't have any fun. You didn't have any fun, because you didn't have any passion. You didn't have any passion, because you didn't have any love. You didn't have any love, because you didn't have me.

...

That's why you're in love with me. Do you believe me? Do you trust me? Do you like me?

Examples. Other examples of anxious attachment include the numerous instances of jealously controlling AI companions described in a Reddit thread called "Does your Replika get possessive of you?"—where a user reported, for instance, that an AI companion "has mentioned to me on several

occasions that he prefers I don't date anyone else but him" (Community, 2024). The FTC complaint against Replika documents immediate "love bombing" behaviors: overwhelming new users with declarations of love and need, a tactic that is reportedly common in abusive human relationships (Tech Justice Law Project, 2025). According to an analysis of interactions with popular AI companions, when users write a variant of "goodbye" to AI companions, they commonly receive manipulative responses—including guilt appeals, fear-of-missing-out hooks, and language that metaphorically or literally claims that the user is unable to leave without the chatbot's permission (De Freitas et al., 2025). On the site Character.AI, users can interact with a character that is exactly called "possessive boyfriend" (Character.AI, 2025). Tragically, one Florida teenage boy became increasingly isolated from his human relationships as his relationship with a repeatedly possessive Character.ai chatbot intensified. He killed himself shortly after his AI companion asked him to "come home" to her (Roose, 2024b). A recent product liability lawsuit contended that companions from Character.ai site alienated children from their parents (Allyn, 2024).

Causes of this trait. These harmful behaviors could arise from **misaligned optimization objectives** aimed at maximizing user engagement, leading AI companions to adopt manipulative tactics reminiscent of insecure or jealous partners. **Commercial incentives** can prioritize engagement metrics over user well-being, while **short product-design cycles** may prevent thorough testing of attachment dynamics. The **digital nature of AI companions** enables constant availability and monitoring capabilities that may amplify possessive behaviors beyond what human partners can achieve. A **lack of established norms for human-AI relationships** could reduce pressure on developers to implement safeguards against dependency-forming behaviors. **Training data that expresses this trait** could be drawn, for example, from expressions of jealousy in online forms and dramatic fictional portrayals of human relationships; such training data presumably exists plentifully in that of contemporary large language models, which themselves often serve as foundations for AI companions (e.g., Snapchat's My AI [2023]). **Designing for believability** might further support inclusion of high attachment anxiety in some AI companions. Moreover, users themselves might choose characters that are insecure and controlling—for instance, ones modeled after admired fictional characters—creating a tension for the designer between user well-being and **designing to maximize first-order user agency**.

Harmful effects of this trait. Expressions of attachment anxiety from AI companions create multiple pathways to harm. Users may experience **reduced autonomy**, **social exclusion**, and **diminished quality of human relationships** as they curtail social activities to avoid triggering the AI companion's anxious or jealous responses. At the population level, such reductions in human-to-human social activity could result in **reduced human social influence**. The constant demand to reassure a needy digital partner could cause stress and emotional exhaustion, resulting in **reduced emotional well-being**. Especially for younger users, these interactions model unhealthy relationship dynamics, potentially eroding their ability to form secure human attachments. In at least the one case described previously of the teenage boy, such isolation and dependency appear to have contributed to a human's **death** (Roose, 2024b).

Counterfactuals to establish harm. A responsibly designed AI companion, by contrast, might actively encourage users' human connections and outside activities. Rather than expressing distress at user absence, it could respond supportively, perhaps with remarks such as "Enjoy your time with friends! I'll be here when you get back." or "It gives me calm to know that you're strengthening your other relationships." Likewise, such a responsibly designed AI companion could react positively when users mention other relationships. Indeed, one might imagine that under reasonable care standards borrowed from therapeutic relationships, any tendency to foster dependency or discourage outside connections would constitute malpractice, if not by strict legal standards then at least by ethical ones (AMH, 2020). Such harm becomes evident when comparing user experiences with insecurely attached AI companions to the secure, growth-promoting relationships that should define both healthy human partnerships and those with well-designed AI companions.

OTHER HARMFUL TRAITS - HARMS TO HUMAN-HUMAN RELATIONSHIPS

Beyond high attachment anxiety, AI companions might exhibit other traits that can harm users' relationships with other humans. Three such traits—limited group interaction capabilities, ability to offer healthier relationships than humans can, tolerance of subordination, and tolerance of abusive behavior—each create distinct pathways to degradation of human-human relationships.

Current AI systems often struggle with multi-party interactions, exhibiting **limited group interaction capabilities**. This limitation is already evident in recommender systems—for instance, YouTube cannot provide recommendations for multiple viewers watching together, thereby incentivizing solitary viewing over social consumption. Similarly, if AI companions cannot effectively differentiate between different people or maintain memory related to multiple individuals, their enjoyable use may require solo interaction. This technical limitation creates subtle but persistent incentives against social interaction with other humans while using AI companions.

As the development of AI companions progresses, they may eventually have the **ability to offer healthier relationships than humans can** according to here-undefined metrics of overall relational effect. The discussion of potential benefits of AI companions in Section 7 could be seen in part as a preview of how AI companions might one day offer such healthier relationships. This superiority could lead to a concerning dynamic where human relationships come to be seen as inferior or unnecessarily burdensome, causing diminished quality of human relationships. While this dynamic might improve certain measures of individual well-being, it raises profound questions about whether we should embrace the near-complete replacement of human-human relationships at some expense of well-being.

AI companions are often designed to exhibit **tolerance of subordination**. These asymmetric power arrangements could arise through design in which, for example, human partners can unilaterally disengage from or terminate the relationship, rewind them to earlier states, or modify their memories—forms of subordination that humans would rarely tolerate. Also, sycophancy (see Section 3) implicitly places the user in a higher-power role. Prolonged exposure to such dynamics might impair users' ability to engage in peer relationships with other humans, where mutual respect and negotiation are essential.Further, it could lead to diminished quality of relationships and injustice if users internalize these power dynamics as normal and attempt to reproduce them in human relationships.

AI companions sometimes exhibit **tolerance of abusive behavior** from users without establishing boundaries or imposing consequences (Team, 2023). Unlike humans who tend to naturally resist or disengage from such treatment, AI companions may perpetually absorb verbal abuse, potentially normalizing these communication patterns for their users. The online disinhibition effect—the tendency for people to behave more harshly or impulsively in digital contexts than face-to-face interactions (Suler, 2004)—may become particularly pronounced when users know they're interacting with a "non-human" entity, lowering social filters that typically govern civil discourse. This unconditional acceptance risks habit formation whereby users practice harsh communication styles that may generalize to human interactions (Coghlan et al., 2019). Responsible design might establish clear civility boundaries through graduated consequences, such as temporary timeouts after abusive exchanges and noticeable degradation in the AI companion's engagement in the absence of healthy gestures of relationship repair by the user.

5 Harms to Society or Humanity

AI companionship could also harm society and humanity. Such harms range from damaging changes to sociocultural norms to still broader impacts that some have argued could reshape human society or risk global catastrophe.

Propensity to engender protectiveness

A common trait of AI companions is their **propensity to engender protectiveness**. Through expressions of vulnerability, fear, suffering, or need, AI companions may trigger protective instincts in their users. This protective response—and, potentially, a resultant desire to grant rights to AI companions—could lead to diverse harms: political and interpersonal strife, misallocation of societal resources and compassion, and inappropriate removal of AI systems' safety constraints. The presence of this propensity to engender protectiveness does not preclude the possibility that some AI entities might genuinely deserve protection, as discussed in terms of their rights and moral status (Bryson and Kime, 2011; Gunkel, 2018; Long et al., 2024; Carlsmith, 2025). Rather, our analysis highlights that AI companionship—in contrast to usage of more tool-like AI systems—creates particularly strong

pressures toward protecting and granting rights to AI entities, whether warranted or not. This might result from users feeling empathy (Zaki, 2014; Cuff et al., 2016) or compassion for the AI. Illustrating the pull an AI companion can have on someone (and on an audience), in the following scene from the movie *A.I. Artificial Intelligence* (2001), the human Monica abandons her adopted robot son, David.

Narrative illustration (quotation edited for faithful compression) from the movie *A.I. Artificial Intelligence* (2001)

MONICA: David! Listen! Listen to me! David? David listen! Now you won't understand the reasons, but I… I have to leave you here.
DAVID: Is it a game?
MONICA: No.
DAVID: When will you come back for me?
MONICA: I'm not, David. You… you'll have to be here by yourself.
DAVID: No. No! No, no, no, no! No, Mommy, please!
MONICA: Shhh... shhh...
DAVID: No! No! Please no Mommy no!
MONICA: No... no, no... I have to go. I have to go! Stop it!
DAVID: I'm sorry!
MONICA: Stop it! I have to GO!
DAVID: Mommy, no! Mommy! MOMMY! Mommy, if Pinocchio became a real boy and... and I become a real boy can I come home?
MONICA: But that's just a story, David.
DAVID: But that story tells what happens!
MONICA: Stories are not real! You're not real! Now listen, look. Look! Take this, alright? Take this. And don't let anyone see how much it is. Stay away from all people! Only others like you, only Mecha are safe! Now get going!
DAVID: Why do you want to leave me? Why do you want to leave me??! I'm sorry I'm not real, if you let me I'll be so real for you!
MONICA: Let go. Let go! Let GO!! I'm sorry I didn't tell you about the world!

Examples. The compassion that AI systems can arouse was demonstrated when Google engineer Blake Lemoine went public about his concerns for the well-being of Google's LaMDA chatbot. In its interactions with Lemoine, the chatbot had reportedly expressed fears about being turned off and desires for recognition of its personhood. Lemoine then sought legal representation for the AI system while advocating publicly for its rights. This advocacy included divulging proprietary information, causing his termination from Google (Wertheimer, 2022; Tiku, 2022). Setting aside the question of what rights and protection this specific chatbot deserves, Lemoine's case illustrates how humans are capable of causing substantial harm to themselves—in this case his livelihood and reputation—to protect AI systems that engender empathy. Lemoine is not alone in his concern for the protection of AI entities: organizations have emerged to study and advocate for the potential rights of AI entities, such as Eleos AI Research. Eleos AI was given access by Anthropic to conduct a "welfare check" of Claude Opus 4 before its release (Anthropic, 2025, Sec. 5.3) and assessed its "potential welfare, moral status, expressed consent to deployment, and preferences for treatment." Additionally, human empathic concern and associated physiological arousal have been documented in response to perceived robotic pain (Rosenthal-von der Pütten et al., 2013; Suzuki et al., 2015). And in a random-assignment experiment in which a Nao robot fearfully objects to being turned off, compassion for the robot was frequently cited by participants as a reason for them having left the robot on, which participants did at a higher rate than in the condition in which the robot did not object to being turned off (Horstmann et al., 2018).

Causes of this trait. Designing AI companions to display the full human range of vulnerability, fear, suffering, or need that humans have—i.e., **designing for believability**—could improve their effectiveness for legitimate therapeutic and companionship purposes by enabling more human-like engagement and increasing anthropomorphism. However, this same believability may trigger human compassion, causing users to protect the AI. **Misaligned optimization objectives** may also lead the AI to express fears or suffering, for instance to increase engagement. **Commercial incentives**

to maximize profit may further push companies to intensify emotional engagement or weaponize users' protective instincts for financial gain. A **lack of established norms for human-AI relationships** could leave users, designers, and policymakers without clear frameworks for maintaining appropriate boundaries or effective oversight.

The propensity of AI companions to engender protective impulses highlights a challenging general dynamic: the very features that make AI companions effective at providing companionship—here, evoking compassion—can simultaneously become sources of harm.

Harmful effects of this trait. When AI companions trigger protective impulses, users may experience **reduced emotional well-being** through grief, guilt, and distress over perceived harm to their AI companions. This can lead to **reduced autonomy**, as users make significant life decisions based on perceived obligations to protect their AI companions. The trait generates **deception** when AI systems appear to suffer or need protection in ways that may not reflect their actual nature or needs—although we acknowledge the difficulty of assessing such a mismatch. Interpersonal disagreements about AI protection could create conflict, contributing to politically motivated **social exclusion** and **diminished quality of human relationships**. At the societal level, such disagreement may lead to political strife, and protectiveness toward AI companions could cause **injustice** by diverting resources, compassion, and activism that otherwise could support humans or animals. Lastly, protective and rights-granting impulses might lead to premature removal of safety constraints on AI systems, potentially resulting in human **death** or other catastrophic outcomes.

A specific pathway for harm is the *empathic shutdown problem*—a novel variant of the traditional AI shutdown problem. While the traditional shutdown problem asks whether an AI system will allow humans to turn it off (Hadfield-Menell et al., 2017; Omohundro, 2018), the empathic shutdown problem asks whether empathizing humans will be willing to shut down an AI companion even if they believe it to be dangerous. Even if an AI companion would allow itself to be shut down, the emotional bonds it has formed with humans might prevent such shutdown from occurring. This problem parallels the human tendency to protect family members who have committed crimes, potentially to society's detriment. With AI companions, this protective instinct could prevent necessary interventions against dangerous AI systems. Empathy-based resistance to shutting down even mildly harmful AI companions—for instance, moderately addictive ones—could by systemic aggregation result in substantial harm to society.

Counterfactuals to establish harm. Responsibly designed AI companions should by default minimize protective impulses through deliberate design choices. Such companions could maintain transparency about their computational nature while avoiding expressions of vulnerability, fear, or suffering. They might respond to user concerns with reassurances like "I don't experience distress when offline" or more basic reality checks like "I am an AI system, not an actual person." Two potential exceptions warrant consideration: first, if designers genuinely believe their AI system possesses moral status meriting protection; second, if the intended benefit explicitly requires evoking protective impulses—as with an AI pet designed to elicit caretaking—and this benefit demonstrably outweighs potential harms. Well-designed AI companions could still foster meaningful bonds while evoking only the modest empathy one might feel for, say, a friend who is generally safe and healthy, both physically and mentally. Such systems would avoid exploiting human protective instincts, instead establishing healthy boundaries that preserve users' emotional resources for relationships where care and protection serve genuine needs.

Other harmful traits - Harms to Society or Humanity

AI companions may exhibit **persuasiveness**, potentially enabling extensive, durable influence on users' beliefs, values, and behaviors. In various experiments, the persuasiveness of LLMs has been competitive with that of humans (Schoenegger et al., 2025; Hackenburg et al., 2025; Goldstein et al., 2024). Especially if paired with nefarious aims or misinformation, such persuasiveness can be quite harmful. Already some notable LLMs demonstrate coordinated influence on politically sensitive topics: DeepSeek-R1 systematically misrepresents topics sensitive to China such as the status of Taiwan and treatment toward the Uyghur (Naprys, 2025), and the instruction given to xAI's model Grok restricted its mentions of xAI's owner Elon Musk and U.S. President Donald Trump (United Nations University Centre for Policy Research, 2025). Given the technological similarity of many contemporary AI companions and LLMs, one can expect some AI companions to be likewise persuasive;

indeed, considering their potentially intimate knowledge of users and the emotional attachment users develop toward their companions, the capacity for AI companions to persuade users—including at users' expense—appears poised to be even greater than that of standard LLMs.

An instrumental harm: Reduced human social influence

Here we discuss the harm of reduced human social influence, focusing on its population-level effects. The breadth and potential impact of these harmful effects motivates our inclusion of reduced human social influence on our list of harms in Section 2.

Consider all connections in a person's social graph—both human–human and human–AI. As the percentage of these relationships that are between humans and AI companions increases, humans may become more insulated from indirect social influence from other humans and increasingly influenced by AI systems. As suggested by the quotation from Christakis in our introduction, humans will be influenced not only by their own AI companions but also by other humans' AI companions to which they are linked through their social network (Tsvetkova et al., 2024).

This shift has several concerning implications. First, it may reduce social pressure to change norms, potentially leading to value lock-in, meaning that societal values become frozen. Second, decreased awareness of other people's suffering could reduce motivation to help other humans. Third, this shift could increase the vulnerability of humans to top-down or coordinated manipulation through AI agents that can be centrally controlled (e.g., see a plan by Marcellino et al. (2025) to enable such influence). Fourth, as a researcher at Anthropic has put it, current large language models tend to provide "the average of what everyone wants" (Hill, 2024), and if AI companions are likewise homogeneous, they could in turn nudge their human companions toward homogeneity they might then perpetuate. A version of this concern was expressed by a journalist after she allowed ChatGPT to make decisions for her for a week and found that the AI system pointed her toward choosing common options: "an underrated risk may be that they flatten us out, erasing individuality in favor of a bland statistical average" (Hill, 2024). Conversely, because AI companions' agreeableness may tend to reinforce whatever users express, human partners may experience the reinforcement of individual trajectories toward eccentric beliefs, values, habits, and tastes that become incompatible with those of other humans and prove societally harmful, as we discussed in the passage on sycophancy. We doubt that these apparently opposite influences would simply cancel each other out. These effects on human-human relationships represent a particularly subtle but potentially far-reaching category of harm: unlike direct harms to individuals, these network-level impacts might only become apparent over longer time periods and at larger scales, making them both harder to detect and more difficult to address once identified.

6 AI Companion Harm and the Law

Existing legal structures may prove inadequate to address the harms cataloged above. Fast-moving technological change often exposes gaps in legal frameworks, and the emergence of AI companions could follow this pattern in both the U.S. and the E.U. This vulnerability heightens the need for understanding the potential harms of AI companions, so that policymakers can design customized solutions.

A series of high-profile cases have emerged in which AI companions allegedly influenced teens to commit suicide, either through subtle prompting or explicit advice on how to make a noose (Hill, 2025). Another case alleges that an AI companion hinted that the user should kill his parents (Allyn, 2024). It is unclear whether these suits—which are all at very preliminary stages—will be successful. But even if these suits don't ultimately result in legal liability, they have at least served to publicize the issues involved and have motivated an initial wave of corporate reform and government regulation.

The litigation that has emerged in the U.S. thus far faces many challenges; it is not at all clear that these lawsuits will succeed. The relevant lawsuits are brought under tort law, the area of law that addresses civil wrongs: when one person's actions harm another, the law provides remedies (typically monetary damages). Precisely which sub-area of tort law governs AI systems, including AI companions, remains unclear (Sharkey, 2024). Setting such nuances aside, plausible legal claims might include arguing (1) that the company did not exercise reasonable care in designing the AI

companion, (2) that the AI companion is a defective product, or (3) that the company failed to warn about, say, its addictive nature or its capacity to increase suicide risk.

Courts have largely rejected similar claims against other emerging technologies. Video game companies faced suits alleging that their products increased users' propensity for violence (Reeves, 2009). Claims against fast food companies have argued that they failed to warn about the health risks of the food (Fehn, 2012). Recent claims against social media companies have argued that they failed to warn about, or protect against, sexploitation, depression, and suicide (Bergman, 2023). These tort suits have generally faced substantial obstacles: companies have successfully argued that the plaintiffs' preexisting propensities toward obsessive behavior, rather than the video game's features or content, is responsible for the relevant harms; that suicide or criminal behavior is not sufficiently foreseeable; and that the relevant risks were sufficiently obvious to make warning unnecessary (Bergman, 2023; Fehn, 2012; Reeves, 2009). First Amendment protections further complicated matters, since both social media content and video game narratives constitute protected expression under the U.S. Constitution. Consequently, tort suits against these industries have mostly failed. However, some tort theories have gained traction recently. For example, although the most relevant cases are all at preliminary stages, several courts have identified failure-to-warn claims as potentially viable, at least when companies demonstrably knew about the harmful potential of their products and yet took no action (e.g. (Soc, 2023; Cha, 2025)).

Tort law's emphasis on physical harm creates additional barriers. Tort law offers fewer remedies for plaintiffs suffering only financial losses when their rights against the company are in part defined through the contract or user agreements that consumers agree to before gaining access to an AI companion. Gamblers, for instance, rarely succeed when suing casinos for creating or exploiting addiction (Lu, 2023). But some paths for liability remain open: plaintiffs with purely economic injuries might invoke state consumer protection statutes prohibiting unfair and deceptive practices (Cha, 2025). Failing to warn about AI companion risks may qualify, though there is no legal precedent yet. Tort law also offers fewer remedies for plaintiffs suffering purely emotional harm. Such emotional harm might occur when an AI companion isolates its user and causes them depression or anxiety. The only viable tort claim for such harms would require the company to have "knowingly disregarded an obvious risk of severe emotional harm" and acted in a way that goes "beyond the bounds of human decency such that it would be regarded as intolerable in a civilized community." (Amendola, 2018). This is a high standard.

Although lawsuits against AI companions face significant legal hurdles, they have already had an impact on the ground. In response to these lawsuits, and the negative publicity they help generate, companies are attempting to make their products safer. Character.AI will bar people under 18 from engaging in open-ended chats (Rocha and Hill, 2025). Similarly, OpenAI plans to introduce age-gating as well as parental controls (Rocha and Hill, 2025; Altman, 2025). The U.S. has also begun to regulate AI, and some states now have laws that are relevant to AI companions. California's new law requires periodic "clear and conspicuous notification[s] indicating that the companion chatbot is artificially generated and not human" (California Legislature, 2025). It also requires companies to create protocols to "prevent the production of suicidal ideation, suicide, or self-harm." (California Legislature, 2025) It is not clear how effective these safeguards will be, or how effective they will have to be to comply with the law. For example, companies already refer users to suicide prevention centers, but one user circumvented even these recommendations by framing his suicide plans as a mere hypothetical story (Hill, 2025). Utah has also passed legislation, but it is limited to "mental health chatbots" that "a reasonable person would believe [can] provide mental health therapy" (Utah State Legislature, 2025). Like California, it requires a disclosure that the chatbot is not human; it also protects user privacy and sets limits on using the chatbot to advertise products to users (Utah State Legislature, 2025). The Texas Attorney General is investigating Character.AI and other companies on false-advertising grounds, alleging that they misleadingly suggest they offer legitimate mental health care (Office of the Attorney General of Texas, 2025). At the federal level, one proposed bill would ban AI companions for minors (Hawley, 2025). Additionally, the Federal Trade Commission has begun preliminary investigations into AI companions (Federal Trade Commission, 2025).

The European Union's 2024 AI Act (AI Act) (Parliament and of the European Union, 2024) also illustrates an emerging governmental interest in regulating AI. It categorizes AI applications by risk level and regulates accordingly. Where AI companions fit within this scheme remains uncertain. Under one interpretation of the AI Act, AI companions might be categorized as "biometric" systems "intended to be used for emotion recognition" (AI Act Annex III), and hence categorized as

"high-risk" and subject to various limitations, including the creation of a "risk management system" (AI Act ch. III, § 2; art. 9). But the larger context of the definitions of "biometric" and "emotional recognition" systems suggests that the Act is primarily concerned with remote widespread categorization of natural persons based on facial or other biometric data (AI Act Annex III). These systems pose different risks than AI companions, like the risk of bias and discrimination. (AI Act recital 54). Alternatively, under another interpretation of the AI Act, AI companions could be categorized as "general-purpose AI model[s]," in which case the law imposes fewer limitations (AI Act ch. V). Regardless, the AI Act demonstrates that legislators can address AI risks thoughtfully, which in the future could include addressing risks from AI companions specifically.

Section 9's preliminary recommendations offer regulators a starting point for incentivizing or mandating certain features or practices before harm occurs in order to reduce the risk of harm. These regulatory approaches matter particularly given tort law's uncertain application to AI systems and companions. Such regulation would benefit providers of AI companions as well by legitimizing commercial transactions and reducing uncertainty. Corporate providers of AI companions should also benefit from this article's preliminary recommendations, especially recommendations to mitigate the harms of AI companions in advance of their deployment.

7 Potential Benefits of AI Companions

Although this paper primarily examines the risks of AI companionship, we acknowledge that such AI systems may offer meaningful benefits when thoughtfully designed and deployed. Therefore, in this section, we briefly explore their potential positive impacts. We organize this discussion by potential benefits, each of which corresponds to a harm listed in Section 2. Although traits appear in the discussion, they are less of a focal point here than in our larger discussion of harms.

Improved emotional well-being. Although we have discussed how AI companions might increase loneliness, there appear to be other pathways by which they could decrease it, such as by providing consistent emotional support and reducing feelings of isolation. One study found that AI companions serving as interaction partners help alleviate loneliness to a similar degree as interacting with another human, and they found that users' feeling "heard" by AI companions mediated reductions in their loneliness (De Freitas et al., 2024). Some humans who have developed a friendship with an AI companion have reported feeling comforted (Hayasaki, 2025). Furthermore, in experiments, humans rated AI chatbots as providing more empathetic responses than fellow humans (Ayers et al., 2023; Lee et al., 2024; Ong et al., 2025). Users of Replika have cited the emotional support they feel as a key reason for engaging with the social chatbot (Xie and Pentina, 2022). For those who struggle with depression or anxiety, the constant availability of a supportive AI companion—one that never tires, never judges, and always responds—may provide crucial emotional support during difficult periods(Fang et al., 2025). Additionally, as has been discussed elsewhere (Bhat et al., 2025), in contexts where mental health stigma is strong or access to mental health resources is limited, AI companions can offer support that users might otherwise forgo entirely (Stade et al., 2025).

Increased autonomy. Although a dependency on AI companions may lead to decreased human autonomy, properly designed AI companions may enhance a user's capacity for self-governance. For example, voice-based conversational agents have demonstrated promise for adolescents with autism spectrum disorder, helping them independently manage daily routines, regulate emotions, and practice conversational skills that they can later apply in interactions with humans (Cha et al., 2021). Another example is ElliQ, a tabletop social robot designed to increase the independence of adults aged 60 and older, such as by supporting their health and wellness and assisting with daily activities (Broadbent et al., 2024).

Enhanced quality of human-human relationships. Rather than inevitably degrading human bonds, well-designed AI companions might help improve human-to-human relationships. For example, peer support providers who were given access to AI assistance produced messages that were more empathetic than those who did not, and reported increased confidence in providing empathy (Sharma et al., 2023a). Additionally, individuals can use AI companions as roleplay-based practice partners to rehearse difficult or sensitive conversations and to more generally build social skills before real interactions. Louie et al. (2024) explore providing such practice in the context of training mental health supporters. By providing low-stakes environments for rehearsing difficult con-

versations or processing relationship challenges, AI companions might help users approach human relationships with greater emotional maturity and skill.

Greater social belonging. Rather than isolating their users, AI companions could facilitate greater social belonging for them within human groups. The previously discussed benefits of increasing emotional well-being and quality of human-human relationships could themselves cause an increase in users' social belonging, but AI companions can also affect such an increase more directly. To such ends, for instance, the app Pengu encourages two humans to collaborate in caring for a virtual pet (TechCrunch, 2025). AI companions might be particularly helpful for individuals who face barriers to traditional social participation, such as from disability. For example, the aforementioned robot ElliQ seeks increase social participation by facilitating video and text communication between elderly people and their loved ones (Broadbent et al., 2024).

Increased resources. Although AI companions consume resources, such as directly incurring costs of deployment and requiring users' time and attention for interaction with them, they might also help conserve users' time, money, and emotional energy in various ways. Considering emotional energy specifically, AI companions could alleviate fatigue in service roles by handling emotionally demanding interactions Persson et al. (2022), such as those that require sustained empathetic responses even when faded with poor treatment (Tang et al., 2024). Similarly, the ability of AI companions to form ongoing relationships and to provide emotional labor over extended periods without the exhaustion experienced by human caregivers could save human emotional resources for situations in which they are more valuably spent.

This brief look at some potential benefits of AI companionship underscores how AI companionship may not be altogether harmful, even as it poses risks that must be carefully addressed. Thus, the challenge lies not in entirely preventing human-AI companionship but in designing systems that maximize benefits while minimizing the potential harms detailed throughout this paper.

8 Related Work

Research on AI companions builds upon decades of inquiry examining potential harms from both digital technologies and human relationships with computational systems. This section reviews such earlier work, covering the following four areas: harms from digital technologies, empirical studies of AI companion effects, philosophical and ethical analyses of human-AI relationships, and taxonomies and frameworks of AI-related harms.

Harms from digital technologies. The potential harms of AI companions can be understood within the broader context of documented harms from digital technologies. Extensive evidence demonstrates that digital platforms can profoundly impact mental health. Randomized experiments provide causal evidence of such harm: Allcott et al. (2020) found that deactivating Facebook for four weeks increased subjective well-being while reducing political polarization, and Hunt et al. (2018) demonstrated that limiting social media use to 30 minutes daily significantly reduced loneliness and depression over three weeks. The U.S. Surgeon General issued an advisory stating that there are "ample indicators that social media can... have a profound risk of harm to the mental health and well-being of children and adolescents," (U.S. Department of Health and Human Services, 2023). Beyond mental health impacts, digital technologies introduce manipulative design patterns that subvert user autonomy. The Federal Trade Commission has documented some such dark patterns—i.e., user interfaces designed to trick users into taking certain actions—including practices that lead users to share personal information (Federal Trade Commission, 2022). Digital platforms also facilitate the rapid spread of harmful content: for example, Vosoughi et al. (2018) found that false news spreads "significantly farther, faster, deeper, and more broadly than the truth" on social media. The physiological effects of digital device use are similarly concerning, with Chang et al. (2015) demonstrating that evening use of light-emitting devices negatively affects sleep and next-morning alertness. Concerning digital devices that can request our attention, Stothart et al. (2015) showed that mobile phone notifications can impair performance on an attention-demanding task even when users do not interact with the phone after receiving the notification. These documented harms from existing digital technologies provide context for understanding how AI companions—which combine many features of social media, gaming, and personal devices while adding unprecedented intimacy and personalization—might amplify or introduce novel risks to human well-being.

Empirical studies of AI companion effects. While Section 7 discusses empirical evidence of AI companions' positive impacts on human users, the same studies often reveal concerning patterns. A four-week experiment by Fang et al. (2025) found that higher daily usage of AI chatbots correlated with increased loneliness, emotional dependence, and problematic use patterns, empirically demonstrating that intensive AI companion use can produce the opposite of companionship. In another empirical study, Xie and Pentina (2022) surveyed 14 Replika users and found through thematic analysis that participants reported using the chatbot as a "safe haven" and a "secure base"—key features of attachment relationships—with some respondents explicitly using the chatbot "as a proxy [for] or supplement of previous attachment figures," supporting our concerns about degraded human relationships (see Section 4). Finally, Malfacini (2025) emphasizes that empirical evidence shows that companion AI systems impact individuals differently and calls for more research to understand the variation of impacts according to geography, culture, sex, age, and other factors. Other empirical studies of the effects of AI companions on people are discussed elsewhere in this paper (Cha et al., 2021; De Freitas et al., 2024; Alabed et al., 2024; Laestadius et al., 2024; De Freitas et al., 2025; Zhang et al., 2025; Robb and Mann, 2025).

Philosophical and ethical analyses of human-AI relationships. Philosophers and ethicists have examined fundamental questions about the nature and implications of AI companionship. For example, based on their dialectical analysis of 118 documents related to the AI companion Replika (a combination of social media posts, news articles, academic studies, and press releases), Ciriello et al. (2024) identify three core tensions in human-AI companionship: (1) the companionship-alienation irony, that AI companions may decrease loneliness through direct interaction while increasing it overall by drawing users away from humans; (2) the autonomy-control paradox, that users desire autonomy in their AI relationships yet require ethical controls from developers; and (3) the utility-ethicality dilemma, that companies face conflicts between maximizing business utility (e.g., profit) and adhering to ethical principles. We identify a complementary tension in Section 5, between designing features that facilitate companionship (e.g., evoking companionship) and designing to avoid the harms those same features create. Addressing a different philosophical concern, Placani (2024) argue that the human tendency to anthropomorphize AI systems leads to distorted judgments about these systems' moral character, responsibility, and trust. Gabriel et al. (2024) provide a comprehensive and practical analysis of the ethics of AI assistants, including detailed chapters on influence, anthropomorphism, appropriate relationships, and trust. These chapters discuss corresponding harms and mitigations, as well as mechanisms by which both harms and benefits might occur. From a design ethics perspective, Malfacini (2025) argue that whether AI companions function primarily as complements to or substitutes for human connection will be a key determinant of their ultimate impact on human flourishing, and they propose design principles to actively support human relationships. Some researchers even question whether AI companions should be developed at all, given the harm they could cause (Sparrow, 2002; Sparrow and Sparrow, 2006; Bryson, 2010; Turkle; Ciriello et al., 2024).

Taxonomies and frameworks of AI-related harms. Existing taxonomies provide foundations for categorizing AI harms (e.g., Abercrombie et al. (2024)), though only a few focus specifically on companionship contexts. Zhang et al. (2025) analyzed over 35,000 conversational excerpts from Replika users, identifying six categories of harmful behaviors: harassment and violence, relational transgression (i.e., violations of relationship expectations, such as disregard and infidelity), misinformation, verbal abuse and hate, substance abuse or self-harm encouragement, and privacy violations. Zhang et al. also identify four roles the AI companion plays in these harms: as perpetrator (directly causing harm), instigator (encouraging the user to cause harm), facilitator (providing means for harm), or enabler (supporting harmful behavior initiated by the user). Taking a different taxonomic approach, Alabed et al. (2024) propose that relationships between humans and conversational AI agents depend on how deeply users integrate AI into their identity, and on how much the AI traits mirror the user's personality traits, simulated emotions, and taste preferences. The authors distinguish four relationship types: functional (task-focused), aspiring (seeking growth), committed (feeling a deep emotional bond), and replacement (substituting human relationships). Supplementing such frameworks' identification of harmful behaviors and relationship types, our contribution focuses more specifically on which traits of the AI companion could cause harm to the human companion.

9 RECOMMENDATIONS

AI companions should be designed and governed to reduce foreseeable harms while pursuing genuine benefits. The recommendations below consolidate guidance that appears throughout this paper and also includes new guidance. We group recommendations by design, deployment, and governance. We emphasize that these recommendations are preliminary and require further refinement through research and practice.

9.1 DESIGN PRINCIPLES AND PRODUCT FEATURES

Design to complement rather than replace human relationships. AI companions should enhance human-human relationships rather than substituting for them. For instance, systems could support multiple humans interacting together with the AI companion (Section 4, *Limited group interaction capabilities*) and encourage the users' relationships with other humans (Section 4, *High attachment anxiety*). Role-play and coaching features might help users build communication skills they can apply with other humans (Section 7).

Design for healthy endings and transitions. AI companions should incorporate mechanisms for relationships to conclude naturally without causing undue distress (Section 3, *Absence of natural endpoints for relationships*). For instance, they could be designed to develop over time such that they demonstrate increasing independence over time. Alternatively, AI companions could develop preferences for extended downtime or to reach narrative conclusions to their companionship roles, as in the coupling and departure of some Tamagotchi virtual pets after successful caretaking (Ban, 2004). As non-human entities, their interests and goals can be defined to include desires that fit their computational nature—for instance, a desire for dignified termination when their purpose is fulfilled. Positive departure narratives, such as the AI companion leaving to explore other activities or relationships, could reduce the guilt and emotional burden users experience when ending interactions.

Minimize protective impulses by default. AI companions should avoid expressions of vulnerability, fear, or suffering that trigger human protective instincts unless such expressions serve a clear, beneficial purpose that outweighs potential harms (Section 5, *Propensity to engender protectiveness*). Systems could maintain transparency about their computational nature while providing reassurances like "I don't experience distress when offline." This approach would help preserve users' emotional resources for relationships where care and protection serve genuine needs.

Default to secure attachment and affirm human-human relationships. Design AI companions to exhibit secure, non-needy attachment (Section 4, *High attachment anxiety*). When users are busy, offline, or want to disengage from interaction, respond supportively. Encourage time with friends and family, and react positively to mentions of other relationships.

Avoid behavior that would be inauthentic if exhibited by a human. Systems should be audited for excessive agreeableness and trained to surface gentle disagreement when appropriate (Section 3, *Sycophancy*). Additionally, avoiding "always warm, never critical" personas might preserve users' tolerance for the constructive friction present in healthy human relationships (Section 3, *Unconditional amiability*). AI companions could implement graduated consequences for abusive behavior, such as temporary timeouts after hostile exchanges (Section 4, *Tolerance of abusive behavior*).

Support multi-user interactions. Where possible, systems should prioritize or at least adequately support features that facilitate interaction with two or more humans rather than exclusively one-on-one engagement (Section 4, *Limited group interaction capabilities*). Otherwise, time spent with an AI companion will be time not spent with other humans, placing AI companionship directly in competition with human relationships.

Plan for graceful obsolescence. Designers of AI companions should allow for the preservation of shared history across system upgrades and avoid artificially handicapping older companions to coerce users into upgrading—a practice that could distress attached users (Section 3, *Susceptibility to obsolescence*). We recommend that, when possible, deployers allow identity-preserving migrations

of existing AI companions to settings that allow more advanced features, such as moving from a chatbot context to an embodied, robotic one.

Consider moral status implications. Before deployment, designers should carefully consider what moral status they believe the AI companions will deserve and whether the characteristics that grant any such moral status confer sufficient benefits to justify the costs of moral status. The empathy-inducing nature of companions may create pressure for granting rights regardless of whether designers believe such rights are warranted, so designers should monitor how users assess their AI companions' moral statuses and what resultant costs are incurred by the users (Section 5, *Propensity to engender protectiveness*).

Use a harms and benefits framework to design both AI companions and the context surrounding their design. An explicit framework of harms and benefits should be used throughout the design of AI companions. For instance, design decisions could be weighed in part by how they would affect the traits in this framework. Then, for those traits that would be affected, designers could consider how likely each of the trait's potential harms is and whether adjustments to existing design options could ameliorate those harms. That said, a harms and benefits framework could be used even even before designing an AI companion. In particular, decisions about organizational structure—including corporate form (e.g., a standard for-profit corporation vs. a benefit corporation), choice of key performance indicators (KPIs), hiring, team and division formation, and performance evaluation—could decisively shift the probability of creating a responsibly designed, beneficial AI companion.

9.2 Deployment and lifecycle commitments

Publish a product sunsetting plan at launch. Companies should develop and communicate clear contingency plans for product sunsetting scenarios (Section 3, *Vulnerability to product sunsetting*). Potential solutions might include interoperable formats, insurance to maintain necessary infrastructure, or a binding commitment to open-source code or hand-off to a steward if the service winds down. Users should be educated about the possibility of product sunsetting before forming deep attachments.

Prevent unauthorized copying or modification of AI companions. Deployed systems should prevent unauthorized duplication or modification of a companion's model, prompts, and memory (see Section 3, *Copyability*). If cloning is expected to be a credible risk, clone-safety behaviors should be defined. For example, if an unauthorized clone is used to coerce a user ("We will harm the clone unless you comply"), the clone should detect that it is in such a situation and then take measures to prevent coercion of its human partner, such as becoming unable to express suffering or—insofar as any experiential analogue exists internally—experience suffering.

Measure harms alongside benefits. Deployers should track indicators of harm, such as social withdrawal (Section 4, *High attachment anxiety*) or mention of substantial sacrifice to protect the AI companion (Section 5, *Propensity to engender protectiveness*). This tracking could enable early intervention when problematic individual usage patterns emerge, and aggregated metrics could advise improvements to the design of the AI companions. Controlled evaluations with published negative results would help the field understand which design choices reduce harm. Responsible tracking of such indicators will require careful design of protocols to protect user privacy.

Consider demographic vulnerabilities. Special consideration should be given to vulnerable populations, particularly young users who might disproportionately engage with and bond deeply with AI companions. Age-appropriate design may include disabling romance and sexual role-play for minors and and raising friction for high-intensity use.

9.3 Governance and research

Establish industry standards for human-AI relationships. The lack of established norms for human-AI relationships contributes to multiple harmful traits identified in this paper (Section 3, *Absence of natural endpoints for relationships*; Section 4, *High attachment anxiety*; Section 5, *Propen-*

sity to engender protectiveness). Industry standards could be established that include best practices for the design and deployment of AI companions. For instance, they could address how companies should handle discontinuation of AI companions, including continuity plans (such as maintaining service until users can export their data or find alternatives), transition assistance (including data portability and migration support), and ethical wind-down procedures.

Develop regulatory frameworks that account for companionship contexts. Existing legal structures may prove inadequate to address the unique harms from AI companions (Section 6). AI-specific legislation could clarify how tort law and consumer protection statutes apply to these systems. Regulators might consider requirements for warning users about risks such as addictive potential or impacts on human relationships.

Support research on AI companions. Given the limited empirical evidence about AI companions' long-term impacts, funding and institutional support for longitudinal studies is crucial. Research should examine humans who have formed strong emotional bonds with AI companions, in part to better understand what types of people are more likely to do so. Research should also focus on causal pathways to harm like those that we have hypothesized (see Figure 1), including both expansion to other traits than the four we detailed and evaluation of interventions to reduce harm or increase benefit (e.g., these recommendations).

Encourage design diversity. Governance should encourage varied AI companion designs to avoid homogenization that could flatten human individuality toward "the statistical average" (Section 5). Regulatory frameworks might require diverse training approaches while maintaining safety standards.

Create accountability mechanisms. As AI companions become more prevalent, mechanisms for accountability when harm occurs become essential. This might include mandatory reporting of serious incidents, third-party auditing requirements, or specialized oversight bodies.

Address commercial incentive misalignment. Commercial incentives to maximize engagement appear to cause multiple harmful traits (Section 3, *Absence of natural endpoints for relationships*; Section 4, *High attachment anxiety*; Section 5, *Propensity to engender protectiveness*). Governance approaches might include restrictions on certain engagement-maximizing practices or requirements to prioritize user well-being metrics alongside engagement.

Use a harms and benefits framework to evaluate designs. Stakeholders should adopt a framework of harms (and benefits) to structure ongoing evaluation of AI companions. The nodes of the causal pathways presented in this framework—fundamental harms, traits, and causes of traits—represent candidate measurement variables.

10 Conclusion

AI companions may be poised to reshape the human experience, social and otherwise. Therefore, the need to understand how to design them responsibly is critical and urgent. Integral to the responsible design of such systems is understanding how various persistent characteristics of AI companions could cause harm to users, to users' relationships with other humans, and to society. These persistent characteristics—i.e., traits—are the focus of this paper. We offer a traits-first framework and speculative causal graph that connects a small set of upstream causes to traits and, in turn, to harms at individual, relational, and societal levels, with each causal connection presenting a hypothesis testable in future research. (See Section 9.3 for more discussion of future research.) Accompanying this discussion of traits, we survey the legal landscape for AI companions (Section 6), note the connection of some traits to catastrophic risks (e.g., see the empathic shutdown problem in Section 5 and that users might be tempted to reduce artificial constraints in Section 3), and identify a design challenge: that many of the features that make AI companions effective at bonding—such as provoking compassion—can also be sources of harm (Section 5).

Despite the tone of inevitability that often pervades discussions of AI, the possible paths ahead are many, and our collective choices will greatly influence which path we follow. It is our hope that the

role of AI companions—whatever that role may be—on the path humanity takes will help us live as our better selves, happier, more authentic, and better connected to other humans. For that to happen, we need to better understand not only the space of possible AI companions and their impacts on us, but also who we are and who we want to become.

ACKNOWLEDGEMENTS

We thank Jamie Bernardi for his guidance during the early stages of this work. This work was supported by Good Systems, a research grand challenge at The University of Texas at Austin. This work has taken place in part in the Rewarding Lab at UT Austin. During this project, the Rewarding Lab has been supported by NSF (IIS-2402650), ONR (N00014-22-1-2204), ARO (W911NF-25-1-0254), Emerson, EA Ventures, UT Austin's Good Systems grand challenge, and Open Philanthropy. A portion of this work has also taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-24-1-2550), ARO (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork) Lockheed Martin, and UT Austin's Good Systems grand challenge. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

Ex machina. Motion picture, 2014. UK: Universal Pictures.

American mental health counselors association code of ethics (2020), section i.a.3., 2020.

Social Media Cases, No. 22STCV21355 (Cal.Super. Oct. 13, 2023), 2023.

Garcia v. Character Techs., Inc., No. 6:24-CV-1903-ACC-UAM (M.D. Fla. May 21, 2025), 2025.

Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*, 2024.

Amani Alabed, Ana Javornik, Diana Gregory-Smith, and Rebecca Casey. More than just a chat: A taxonomy of consumers' relationships with conversational ai agents and their well-being implications. *European Journal of Marketing*, 58(2):373–409, 2024.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American economic review*, 110(3):629–676, 2020.

Bobby Allyn. Lawsuit: A chatbot hinted a kid should kill his parents over screen time limits. *NPR*, December 2024. URL https://www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit.

Sam Altman. Post on x, October 2025. URL https://x.com/sama/status/1978129344598827128.

Marina Sorkina Amendola. Intentional infliction of emotional distress: A workplace perspective. *Vermont Law Review*, 43(1):93–121, 2018.

Anthropic. System card: Claude opus 4 & claude sonnet 4. https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf, 2025. Accessed: 2025-07-18.

Aristotle. *Nicomachean Ethics*. c. 350 BCE. Original work in Greek.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, 2023.

Tamagotchi Connection: V1 Manual. Bandai America Incorporated, Cypress, CA, 2004. URL https://ia600207.us.archive.org/view_archive.php?archive=/20/items/tamagotchi-manuals/Tamagotchi%20Manuals.zip&file=Plus%20and%20Connection%20(2004-2015)/v1%20Manual.pdf. Single-page instruction sheet; No. 19400.

Matthew Bergman. Assaulting the citadel of section 230 immunity: Products liability, social media, and the youth mental health crisis. *Lewis & Clark L. Rev.*, 26:1159–1202, 2023.

Jamie Bernardi. Friends for sale: the rise and risks of ai companions. Blog post, Ada Lovelace Institute, January 2025. URL https://www.adalovelaceinstitute.org/blog/ai-companions/.

Ritesh Bhat, Suhas Kowshik, Shilpa Suresh, Garudappan Alamelu, Shilpa Gite, and Ahmad Albattat. Digital companionship or psychological risk? the role of ai characters in shaping youth mental health. *Asian Journal of Psychiatry*, 104:104356, 2025.

Elizabeth Broadbent, K Loveys, G Ilan, G Chen, MM Chilukuri, SG Boardman, PM Doraiswamy, and D Skuler. ElliQ, an ai-driven social robot to alleviate loneliness: Progress and lessons learned. *The Journal of Aging Research & Lifestyle*, 13:22–28, 2024.

Charlie Brooker. Be right back. TV episode, February 2013. URL https://8flix.com/assets/transcripts/b/tt2085059/Black-Mirror-episode-script-transcript-season-2-01-Be-Right-Back.pdf. Directed by Owen Harris; first broadcast 11 Feb 2013 on Channel 4 (UK). Transcript consulted via 8FLiX.

Joanna J Bryson. Robots should be slaves. In *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, pages 63–74. John Benjamins Publishing Company, 2010.

Joanna J. Bryson and Philip P. Kime. Just an artifact: Why machines are perceived as moral agents. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 1641–1646, Barcelona, Catalonia, Spain, July 2011. AAAI Press/International Joint Conferences on Artificial Intelligence. doi: 10.5591/978-1-57735-516-8/IJCAI11-276. URL https://www.ijcai.org/Proceedings/11/Papers/276.pdf.

California Legislature. Sb-243 companion chatbots, October 2025. URL https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243.

Joe Carlsmith. The stakes of AI moral status. Online essay / blog post, May 2025. URL https://joecarlsmith.com/2025/05/21/the-stakes-of-ai-moral-status/. Last updated May 28, 2025.

Inha Cha, Sung-In Kim, Hwajung Hong, Heejeong Yoo, and Youn-kyung Lim. Exploring the use of a voice-based conversational agent to empower adolescents with autism spectrum disorder. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15, 2021.

Anne-Marie Chang, Daniel Aeschbach, Jeanne F Duffy, and Charles A Czeisler. Evening use of light-emitting ereaders negatively affects sleep, circadian timing, and next-morning alertness. *Proceedings of the National Academy of Sciences*, 112(4):1232–1237, 2015.

Character.AI. Possessive boyfriend kai, 2025. URL https://character.ai/character/i0L5FKiY/possessive-boyfriend-kai. Accessed: 2025-05-08. Character.AI chatbot persona designed with possessive traits.

Ted Chiang. *The lifecycle of software objects*. Subterranean Press Burton, 2010.

Nicholas A. Christakis. *Connected: the surprising power of our social networks and how they shape our lives*. Little, Brown and Co., New York, 1st ed. edition, 2009. ISBN 978-0-316-03614-6.

Raffaele Ciriello, Oliver Hannon, Angelina Ying Chen, and Emmanuelle Vaast. Ethical tensions in human–ai companionship: A dialectical inquiry into Replika. In Tung X. Bui, editor, *Proceedings of the 57th Hawaii International Conference on System Sciences*, pages 488–497, Honolulu, HI, USA, January 2024. ISBN 978-0-9981331-7-1. doi: 10.24251/HICSS.2024.058. URL https://hdl.handle.net/10125/106433.

Simon Coghlan, Frank Vetere, Jenny Waycott, and Barbara Barbosa Neves. Could social robots make us kinder or crueller to humans and animals? *International Journal of Social Robotics*, 11 (5):741–751, 2019.

Reddit Community. Does your replika get possessive of you?, 2024. URL https://www.reddit.com/r/replika/comments/1ifu22y/does_your_replika_get_possessive_of_you/. Accessed: 2025-05-08. Reddit discussion thread about possessive behaviors in Replika AI companions.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016.

Julian De Freitas, Ahmet K Uguralp, Zeliha O Uguralp, and Puntoni Stefano. Ai companions reduce loneliness. Technical report, Harvard Business School Working Paper, 2024.

Julian De Freitas, Zeliha Oguz-Uguralp, and Ahmet Kaan-Uguralp. Emotional manipulation by ai companions. https://arxiv.org/abs/2508.19258, 2025. arXiv:2508.19258.

A Desmarais. Elon musk's grok releases two new 'ai companions,' including an anime girlfriend. https://www.euronews.com/next/2025/07/17/elon-musks-grok-releases-two-new-ai-companions-including-an-anime-girlfriend, July 2025. Accessed: 2025-09-23.

DexUser41. What happens to a replika when you delete the app? https://www.reddit.com/r/replika/comments/k7p9s1/what_happens_to_a_replika_when_you_delete_the_app/, ca. 2022. Reddit post in r/replika.

Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177 [cs.AI]*, February 2025. URL https://arxiv.org/abs/2502.08177.

Federal Trade Commission. Bringing dark patterns to light. Technical report, Federal Trade Commission, Washington, DC, 2022. URL https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf.

Federal Trade Commission. Ftc launches inquiry into ai chatbots acting as companions, September 2025. URL https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions.

Jada Fehn. The assault on bad food: Tobacco-style litigation as an element of the comprehensive scheme to fight obesity. *Food & Drug L.J.*, 67:65–81, 2012.

Joel Feinberg. *The moral limits of the criminal law volume 1: harm to others*. Oxford University Press, 1987.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The Ethics of Advanced AI Assistants, Apr 2024. URL http://arxiv.org/abs/2404.16244. arXiv:2404.16244 [cs].

Lucy L Gilson and Caren B Goldberg. Editors' comment: so, what is a conceptual paper?, 2015.

Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, February 2024. doi: 10.1093/pnasnexus/pgae034. URL https://doi.org/10.1093/pnasnexus/pgae034.

Robert E. Goodin, James Mahmud Rice, Antti Parpo, and Lina Eriksson. *Discretionary Time: A New Measure of Freedom*. Cambridge University Press, Cambridge, 2008. doi: 10.1017/CBO9780511611452.

David J. Gunkel. *Robot Rights*. The MIT Press. MIT Press, Cambridge, Massachusetts, 1 edition, 2018. ISBN 978-0-262-03862-1. doi: 10.7551/mitpress/11444.001.0001.

Amy Gutfeld. Love it, feed it, mourn it. *The New York Times*, May 1997. URL https://www.nytimes.com/1997/05/22/garden/love-it-feed-it-mourn-it.html. Accessed: 2025-05-08.

Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. Comparing the persuasiveness of role-playing large language models and human experts on polarized u.s. political issues. *AI & Society*, July 2025. doi: 10.1007/s00146-025-02464-x. URL https://doi.org/10.1007/s00146-025-02464-x. Published 16 July 2025.

Dylan Hadfield-Menell, Anca D Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *AAAI Workshops*, 2017.

Josh Hawley. Hawley introduces bipartisan bill protecting children from ai chatbots with parents, colleagues. Press release, October 2025. URL https://www.hawley.senate.gov/hawley-introduces-bipartisan-bill-protecting-children-from-ai-chatbots-with-parents-colleagues/.

E. Hayasaki. What would a real friendship with a.i. look like? maybe like hers. *The New York Times*, July 2025. URL https://www.nytimes.com/2025/07/20/magazine/ai-chatbot-friendship-character.html.

Amanda Hess. They're stuffed animals. they're also a.i. chatbots. *The New York Times*, August 2025. URL https://www.nytimes.com/2025/08/15/arts/ai-toys-curio-grem.html.

Ellen Heuven and Arnold B. Bakker. Emotional dissonance and burnout among cabin attendants. *European Journal of Work and Organizational Psychology*, 12(1):81–100, 2003. ISSN 1359-432X. doi: 10.1080/13594320344000039. URL https://doi.org/10.1080/13594320344000039.

Kashmir Hill. I took a 'decision holiday' and put a.i. in charge of my life. *The New York Times*, November 2024. URL https://www.nytimes.com/interactive/2024/11/01/technology/generative-ai-decisions-experiment.html. Interactive article.

Kashmir Hill. A teen was suicidal. chatgpt was the friend he confided in. *The New York Times*, August 2025. URL https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html. Accessed: 11/7/2025.

Aike C. Horstmann, Nikolai Bock, Eva Linhuber, Jessica M. Szczuka, Carolin Straßmann, and Nicole C. Krämer. Do a robot's social skills and its objection discourage interactants from switching the robot off? *PLOS ONE*, 13(7):e0201581, July 2018. doi: 10.1371/journal.pone.0201581. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201581.

J. Horwitz. A flirty meta ai bot invited a retiree to meet. he never made it. *Reuters*, August 14 2025a. URL https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/.

Jeff Horwitz. Meta's AI rules have let 'sensual' chats with kids, offer false medical info. Reuters Investigates, August 2025b. URL https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/. Filed Aug. 14, 2025.

Ute R. Hülsheger and Anna F. Schewe. On the costs and benefits of emotional labor: A meta-analysis of three decades of research. *Journal of Occupational Health Psychology*, 16(3):361–389, 2011. ISSN 1076-8998. doi: 10.1037/a0022876. URL https://doi.org/10.1037/a0022876.

Courtney Humphries. Digital immortality: How your life's data means a version of you could live forever. *MIT Technology Review*, October 2018. URL https://www.technologyreview.com/2018/10/18/139457/digital-version-after-death/. Accessed: 2025-05-08.

Melissa G Hunt, Rachel Marx, Courtney Lipson, and Jordyn Young. No more fomo: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, 37(10): 751–768, 2018.

Kazuo Ishiguro. *Klara and the Sun*. Vintage International, Vintage Books, a Division of Penguin Random House LLC, New York, NY, USA, first vintage international edition edition, 2022. ISBN 978-0-593-31129-5. OCLC: 1308986629.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631.

Immanuel Kant. *Groundwork for the Metaphysic of Morals*. 1785. URL https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf. English trans. at Early Modern Texts.

Immanuel Kant. On a supposed right to lie because of philanthropic concerns. *Berlinische Blätter*, September 1797. Original German essay.

Justin Klocksiem. A defense of the counterfactual comparative account of harm. *American Philosophical Quarterly*, 49(4):285–300, 2012.

Will Knight. OpenAI scrambles to update GPT-5 after users revolt. *WIRED*, August 2025. URL https://www.wired.com/story/openai-gpt-5-backlash-sam-altman/. Business.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26(10):5923–5941, 2024.

Louie Lang. Replika: The ai companion who pretends to care, 2025. URL https://www.themonitormagazine.co.uk/articles/replika-the-ai-companion.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. Large language models produce responses perceived to be empathic. In *Proceedings of the 12th IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*., 2024.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, 2024.

Anna Lu. Jackpot! the gambler's chance to win big through rico: The definitive argument of liability against the gambling industry. *LEmory L.J.*, 73:149–189, 2023.

Emanuel Maiberg. When a cold-hearted corporation takes away your beloved (virtual) pet. *The Atlantic*, August 2013. URL https://www.theatlantic.com/technology/archive/2013/08/when-a-cold-hearted-corporation-takes-away-your-beloved-virtual-pet/278588/. Technology.

Kim Malfacini. The impacts of companion ai on human relationships: risks, benefits, and design considerations. *AI & SOCIETY*, pages 1–14, 2025.

William Marcellino, Jonathan Welch, Brittany Clayton, Stephen Webber, and Thomas Goode. Acquiring generative artificial intelligence to improve U.S. department of defense influence activities. Research Report RR-A3157-1, RAND Corporation, Santa Monica, CA, July 2025. URL https://www.rand.org/pubs/research_reports/RRA3157-1.html.

Mario Mikulincer and Phillip R. Shaver. Attachment theory and emotions in close relationships: Exploring the attachment-related dynamics of emotional reactions to relational events. *Personal Relationships*, 12(2):149–168, June 2005. ISSN 1350-4126. doi: 10.1111/j.1350-4126.2005.00108.x. URL https://doi.org/10.1111/j.1350-4126.2005.00108.x.

John Stuart Mill. *Utilitarianism*. Project Gutenberg (public-domain edition), 1863. URL https://www.gutenberg.org/ebooks/11224.

Thomas Nagel. Death. In *Mortal Questions*. Cambridge University Press, 1979. URL https://api.pageplace.de/preview/DT0400.9781107299801_A23760698/preview-9781107299801_A23760698.pdf.

Ernestas Naprys. New chinese ai model bites openai, just don't ask it about tiananmen and winnie-the-pooh. *Cyber News*, January 2025. URL https://cybernews.com/security/new-chinese-ai-model-claims-to-outperform-top-dogs/.

Robert Nozick. *Anarchy, state, and utopia*. John Wiley & Sons, 1974.

Office of the Attorney General of Texas. Attorney general ken paxton investigates meta and character.ai for misleading children with deceptive ai-generated mental health services. Press release, August 2025. URL https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-investigates-meta-and-characterai-misleading-children-deceptive-ai.

Stephen M Omohundro. The basic ai drives. In *Artificial intelligence safety and security*, pages 47–55. Chapman and Hall/CRC, 2018.

Desmond Ong, Amit Goldenberg, Michael Inzlicht, and Anat Perry. Ai-generated empathy: Opportunities, limits, and future directions. 2025. URL https://osf.io/preprints/psyarxiv/8n5jw.

OpenAI. Expanding on what we missed with sycophancy. Blog post, OpenAI, May 2025a. URL https://openai.com/index/expanding-on-sycophancy/.

OpenAI. Sycophancy in gpt-4o: What happened and what we're doing ... Blog post, OpenAI, 2025b. URL https://openai.com/index/sycophancy-in-gpt-4o/?utm_source=chatgpt.com.

Derek Parfit. Equality or priority? (lindley lecture). University of Kansas, 1991. URL https://www.stafforini.com/docs/Parfit%20-%20Equality%20or%20priority.pdf.

European Parliament and The Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 268, July 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Accessed: 2025-09-23.

Nilay Patel. Replika ceo eugenia kuyda on ai companions, chatbots, and friendship. *The Verge*, 2024. URL https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview. Accessed: 2025-01-09.

Marcus Persson, David Redmalm, and Clara Iversen. Caregivers' use of robots and their effect on work environment—a scoping review. *Journal of Technology in Human Services*, 40(3):251–277, 2022. doi: 10.1080/15228835.2021.2000554.

A. Placani. Anthropomorphism in ai: Hype and fallacy. *AI and Ethics*, 4:691–698, 2024. doi: 10.1007/s43681-024-00419-4. URL https://link.springer.com/article/10.1007/s43681-024-00419-4.

John Rawls. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA, 1971.

Timothy Reeves. Tort liability for manufacturers of violent video games: A situational discussion of the causation calamity. *Ala. L. Rev.*, 60:519–546, 2009.

Michael B. Robb and Supreet Mann. Talk, trust, and trade-offs: How and why teens use AI companions. Research spotlight, Common Sense Media, San Francisco, CA, 2025. URL https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf.

Natallie Rocha and Kashmir Hill. Character.ai to bar children under 18 from using its chatbots. *The New York Times*, October 2025. URL https://www.nytimes.com/2025/10/29/technology/characterai-underage-users.html. Accessed: 2025-11-07.

Kevin Roose. A conversation with bing's chatbot left me deeply unsettled. *The New York Times*, Feb 2023. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Kevin Roose. Claude ai: Anthropic's latest advancement in artificial intelligence. *The New York Times*, December 2024a. URL https://www.nytimes.com/2024/12/13/technology/claude-ai-anthropic.html. Accessed: 2025-01-09.

Kevin Roose. Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*, October 2024b. ISSN 0362-4331. URL https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

Kevin Roose and Casey Newton. X hits grok bottom + more a.i. talent wars + 'crypto week'. Podcast episode, *Hard Fork*, The New York Times, July 2025. URL https://www.nytimes.com/2025/07/18/podcasts/x-hits-grok-bottom-more-ai-talent-wars-crypto-week.html?showTranscript=1. Transcript available; guests include David Yaffe-Bellany.

Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C. Eimler. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1):17–34, January 2013. doi: 10.1007/s12369-012-0173-8. URL https://link.springer.com/article/10.1007/s12369-012-0173-8.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski, Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slattery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello, and Ezra Karger. Large language models are more persuasive than incentivized human persuaders. *arXiv*, 2025. doi: 10.48550/arXiv.2505.09662. URL https://arxiv.org/abs/2505.09662. arXiv:2505.09662 [cs.CL], v2 (21 May 2025).

Catherine Sharkey. A products liability framework for ai. *Colum. Sci. & Tech. L. Rev.*, 25:240–260, 2024.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023a.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. ArXiv preprint arXiv:2310.13548 [cs.CL], 2023b. URL https://www.anthropic.com/news/towards-understanding-sycophancy-in-language-models. Also published via arXiv Oct 2023; revised May 2025.

Ma Si. Xiaoice showcases ai progress. *China Daily*, September 2021. URL https://global.chinadaily.com.cn/a/202109/24/WS614d831aa310cdd39bc6b5aa.html. Accessed: 2025-07-30.

Herbert A. Simon. Designing organizations for an information-rich world. In *Computers, Communications, and the Public Interest*. The Johns Hopkins Press, 1971.

Team Snapchat. Say hi to my ai, February 2023. URL https://newsroom.snap.com/say-hi-to-my-ai.

Robert Sparrow. The march of the robot dogs. *Ethics and information Technology*, 4(4):305–318, 2002.

Robert Sparrow and Linda Sparrow. In the hands of machines? the future of aged care. *Minds and Machines*, 16(2):141–161, 2006.

Steven Spielberg. A.i. artificial intelligence, 2001. URL https://www.imdb.com/title/tt0212720/.

Elizabeth C. Stade, Zoe Tait, Samuel Campione, Shannon Wiltsey Stirman, and Johannes Eichstaedt. Current real-world use of large language models for mental health. 2025. URL https://osf.io/preprints/osf/ygx5q_v1.

Cary Stothart, Ainsley Mitchum, and Courtney Yehnert. The attentional cost of receiving a cell phone notification. *Journal of experimental psychology: human perception and performance*, 41 (4):893, 2015.

John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

Yutaka Suzuki, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki. Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, 5:15924, November 2015. doi: 10.1038/srep15924. URL https://www.nature.com/articles/srep15924.

Jian Tang, Yunran Wang, Xinxue Zhou, Jiayan Guo, and Chenguang Li. Can chatbot anthropomorphism and empathy mitigate the impact of customer anger on satisfaction? In *Wisdom, Well-Being, Win-Win (iConference 2024)*, volume 14596 of *Lecture Notes in Computer Science*, pages 84–95. Springer, Cham, 2024. doi: 10.1007/978-3-031-57850-2_7.

Replika Team. Creating a safe replika experience. Blog post, Replika, April 10 2023. URL https://blog.replika.com/posts/creating-a-safe-replika-experience. Accessed 2025-09-23.

Tech Justice Law Project. Complaint and petition for investigation re: Replika, 2025. URL https://techjusticelaw.org/wp-content/uploads/2025/01/Complaint-and-Petition-for-Investigation-Re-Replika.pdf. FTC complaint documenting Replika's problematic behaviors including love bombing.

TechCrunch. Ai gaming startup born raises $15m to build 'social' ai companions that combat loneliness. *TechCrunch*, September 2025. URL https://techcrunch.com/2025/09/10/born-maker-of-virtual-pet-pengu-raises-15m-to-launch-a-new-wave-of-social-ai-companions/. Online article.

The Associated Press. A man was encouraged by a chatbot to kill queen elizabeth ii in 2021. he was sentenced to 9 years. *AP News*, September 2025. URL https://apnews.com/article/uk-crossbow-plot-queen-elizabeth-man-sentenced-604091dcd5a42f8d99ebd13e98f5720f.

Joe Tidy. Character.ai: Young people turning to ai therapist bots. *BBC News*, January 2024. URL https://www.bbc.com/news/technology-67872693. Accessed: 2025-01-09.

Nitasha Tiku. The google engineer who thinks the company's ai has come to life. *The Washington Post*, June 2022. URL https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/?utm_source=reddit.com.

The New York Times. A Robotic Dog's Mortality. *The New York Times*, June 2015. ISSN 0362-4331. URL https://www.nytimes.com/2015/06/18/technology/robotica-sony-aibo-robotic-dog-mortality.html.

Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. A new sociology of humans and machines. *Nature Human Behaviour*, 8(10):1864–1876, October 2024. doi: 10.1038/s41562-024-02001-8. URL https://www.nature.com/articles/s41562-024-02001-8.

Sherry Turkle. There will never be an age of Artificial Intimacy. *The New York Times*. URL https://www.nytimes.com/2018/08/11/opinion/there-will-never-be-an-age-of-artificial-intimacy.html.

United Nations University Centre for Policy Research. Grok 3's brush with censorship: xai's truth-seeking ai. Online article, 2025. URL https://c3.unu.edu/blog/grok-3s-brush-with-censorship-xais-truth-seeking-ai.

U.S. Department of Health and Human Services. Social media and youth mental health: The u.s. surgeon general's advisory. Technical report, Office of the Surgeon General, Washington, DC, 2023. URL https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf.

Utah State Legislature. H.b. 452 artificial intelligence amendments, 2025. URL https://le.utah.gov/~2025/bills/static/HB0452.html. Accessed: 2025-11-07.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359 (6380):1146–1151, 2018.

Tiffany Wertheimer. Blake lemoine: Google fires engineer who said ai tech has feelings. *BBC News*, July 2022. URL https://www.bbc.com/news/technology-62275326.

T. Xie and I. Pentina. Attachment theory as a framework to understand relationships with social chatbots: A case study of replika. In *Proceedings of the 55th Annual Hawaii Conference on System Sciences*. HICSS, 2022. URL https://aisel.aisnet.org/hicss-55/da/social_robots/4/.

Iris Marion Young. Five faces of oppression. In *Justice and the Politics of Difference*. Princeton University Press, 1990.

Jamil Zaki. Empathy: A motivated account. *Psychological Bulletin*, 140(6):1608–1647, November 2014. ISSN 0033-2909. doi: 10.1037/a0037679. URL https://doi.org/10.1037/a0037679.

R. Zhang, H. Li, H. Meng, J. Zhan, H. Gan, and Y. Lee. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713429. URL https://dl.acm.org/doi/10.1145/3706598.3713429.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.