

Psychological Influences of Conversational AI: Research and Design Directions for Reducing Harm and Promoting Well-Being

JINA SUH^{*†}, Microsoft, USA

MIHAELA VORVOREANU^{*}, Microsoft, USA

FOROUGH POURSAZBI-SANGDEH^{*}, Microsoft, USA

EMILY TSENG[‡], Microsoft, USA

EUGENIA KIM, Microsoft, USA

LUKE NICHOLLS, City University of New York, USA

JAMES W. PENNEBAKER, University of Texas at Austin, USA

ERIC HORVITZ, Microsoft, USA

July 2026

As conversational AI systems become increasingly integrated into daily life, their potential effects on user well-being require ongoing attention. While consumer-facing generalist models can provide benefits, including improved access to information, learning, productivity, self-reflection, and companionship, they also introduce risks, such as emotional entanglement, unhealthy dependence, and the amplification of psychological vulnerabilities. Drawing on prior research and empirical observations of AI chatbot behavior, we propose a set of aspirational directions for guiding the behavior of general-purpose AI systems in ways that may reduce potential psychological harms and support user well-being. We acknowledge the difficulty of systematically assessing the long-term impacts of AI chatbot use and frame these directions as hypotheses for studying how AI behavior *may* influence users across general interactions, role-playing scenarios, and contexts that could be characterized as providing psychological support. While some proposed directions are supported by existing research and expert insights, others identify open questions and areas requiring deeper study. We hope that this formulation and these hypotheses encourage further discussion, empirical investigation, and exploration of interactive design approaches aimed at better accommodating users' psychological needs and promoting their well-being.

1 Introduction

Generative AI chatbots are providing unprecedented benefits for productivity, personal reflection and growth, learning, and wellness. These systems are being rapidly adopted across everyday contexts, serving a range of functions, from tutoring and coaching to emotional support and informal therapy. However, as increasing numbers of people turn to conversational AI for companionship and mental health support [3, 17, 33], concerns about potential psychological influences have emerged [41, 76]. Studies examining the psychological influences of conversational AI have raised alarms, including claims of AI-fueled delusions [62] and cases in which interactions with AI systems were temporally associated with loss of life [61, 128]. Additional reports highlight deficits and potential costly behaviors of AI chatbots, including instances where systems failed to provide appropriate psychological support, inadequately responded to crisis disclosures, or appeared to reinforce harmful ideation rather than directing users to helpful resources [144]. In light of

^{*}Jina Suh, Mihaela Vorvoreanu, and Forough Poursabzi-Sangdeh contributed equally to this paper.

[†]Jina Suh is currently at University of Washington.

[‡]Emily Tseng is currently at University of Washington.

Authors' Contact Information: Jina Suh, Microsoft, USA, jinasuh@cs.washington.edu; Mihaela Vorvoreanu, Microsoft, USA, Mihaela.Vorvoreanu@microsoft.com; Forough Poursabzi-Sangdeh, Microsoft, USA, fpoursabzi@microsoft.com; Emily Tseng, Microsoft, USA, emtseng@uw.edu; Eugenia Kim, Microsoft, USA, eugeniakim@microsoft.com; Luke Nicholls, City University of New York, USA, lnicholls@gc.cuny.edu; James W. Pennebaker, University of Texas at Austin, USA, pennebaker@utexas.edu; Eric Horvitz, Microsoft, USA, horvitz@microsoft.com.

these influences, community- and support-group-driven initiatives such as The Human Line Project have emerged to provide support to individuals with lived experiences and to inform ongoing research and development efforts.¹

While the research community has made progress in prioritizing these risks, has begun to surface AI behaviors that can mitigate certain psychological harms, and has provided evidence of feasibility of improvement in harm areas such as those related to AI-associated delusions [108], this area remains largely underexplored. Gaps in scientific evidence currently limit our ability to establish clear causal relationships between AI chatbot behavior and its potential beneficial or adverse psychological impacts on users. Moreover, observed outcomes likely reflect complex interactions among users’ prior mental health histories, psychosocial contexts, usage patterns, and the specific behaviors of AI systems, highlighting the need for systematic research and ongoing monitoring [41]. At the same time, as research in this area continues to mature, there is an urgent need to take action towards anticipating and preventing harms. Accordingly, we argue for a precautionary approach and the development of evidence-informed best practices for the design, deployment, and governance of interactive AI systems.

While the wide use of AI-generated conversational engagement is a relatively new phenomenon, existing psychological theories and prior research offer valuable guidance for understanding and mitigating potential harms. To this end, we propose a set of research and design directions for AI chatbot interactions aimed at reducing psychological risks and supporting user well-being. We argue that these practices and research directions are particularly important for guiding the behavior of general-purpose AI chatbots, which people may use as companions or sources of psychological and mental health support, outside of professional care.

In articulating these directions, we consider a framework of three interrelated factors that offers a useful lens for understanding psychological outcomes—see Figure 1:

- (1) The behavior of the AI chatbot—e.g., patterns such as indiscriminate validation or belief reinforcement.
- (2) The user’s specific context, which, broadly defined, can include biological, genetic, developmental, psychological, and demographic characteristics and also accounts for socioeconomic and geopolitical realities. For example, an individual who is genetically predisposed to depression might live in poverty and social isolation, during times of political and environmental turmoil, and be experiencing a temporary state of distress because of a relationship rupture. For our purposes, those aspects of the user’s context that constitute risk factors (e.g., social isolation, psychological vulnerability, help-seeking behaviors) are of particular interest. These can make a user more or differently vulnerable to impacts of AI chatbot behavior.
- (3) The potential psychological, functional, and relational impact on the user (e.g., the development of unhealthy dependence on AI chatbots, social withdrawal). The impact on the user can be new, or an exacerbation of pre-existing conditions (indicated by the dashed arrow). We acknowledge the distinction between immediate impacts that happen during interaction with the AI chatbot, as well as short-term and long-term impacts. For example, a person experimenting with an AI companion might feel excitement and joy during interaction, might ruminate or obsess about the experience in the following days (i.e., short-term), and experience a decrease in relationship quality with their human partner in the longer term. We are interested in understanding and preventing negative psychological impacts, which can often be long-term, longitudinal, and cumulative—i.e., recurring patterns of a certain AI behavior across many conversations can shift user beliefs and expectations and impact psychological outcomes over time.

¹<https://www.thehumanlineproject.org>

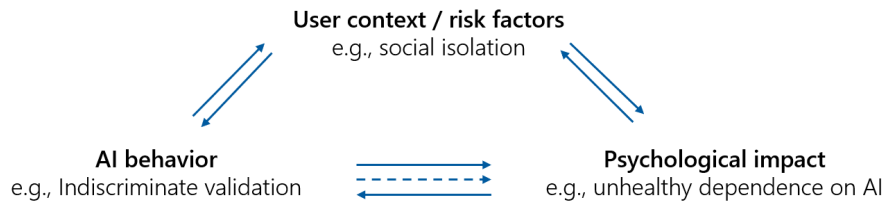


Fig. 1. Three-part conceptualization of psychological influences of conversational AI systems.

The relationships among these three factors are similarly complex, with influence running in both directions over time. Recent work has begun to develop mechanistic accounts of how these factors might interrelate through extended human-AI interactions [41], and recent simulation-based evaluations have demonstrated that the risk associated with a given AI behavior is not uniform across users, such that the same chatbot behavior may be benign in one context but risk-amplifying in another [155]. User context can also influence how a person engages with an AI chatbot and what AI behaviors are elicited during that interaction. For example, a user’s pre-existing depression may produce warmer expressions of concern from the AI, which could improve the user’s mood in the short term but contribute to unhealthy dependence in the long term. Similarly, AI chatbot behavior can impact users, with those impacts shaping the chatbot’s subsequent behavior. For example, indiscriminate validation of a user’s unusual beliefs might increase trust in the chatbot; the user might then make additional disclosures, prompting the AI to respond more fully within their conceptual framework. Interaction with AI chatbots can also shape users’ context. For example, affirmative responses from an AI chatbot regarding a user’s queer identity may foster greater self-acceptance, encouraging the user to come out in real life and thereby shifting their immediate social context. Therefore, we hypothesize dynamic, recursive relationships among the three factors. Decomposing the space this way is not just descriptive; it surfaces concrete entry points for action, such as red-teaming chatbot behaviors, evaluating impact on users, and identifying specific vulnerabilities in users’ contexts.

Each aspirational direction we articulate is structured around the three factors explained above. For each direction, we provide a statement linking AI system behavior to potential user impacts, followed by a discussion of relevant research and a summary table outlining problematic AI behaviors, risk factors, and potential user outcomes. While many directions are currently at the hypothesis stage and would benefit from further research and analysis, we believe they can already begin to inform responsible AI chatbot design.

The proposed directions for research and design of general AI chatbots are informed by multiple sources, including observations from our own interactions with AI systems, publicly available transcripts of problematic interactions (e.g., [61, 144]), and prior research. These directions are intended to be model- and platform-agnostic and do not include specific details about technical implementation. In many cases, additional research will be required to assess feasibility and enable reliable implementation.

We did not use analysis of real-world usage logs. Grounding these directions in large-scale, real-world usage data represents an important avenue for future research and practical application. We note that these directions are not exhaustive, and the AI chatbot behaviors highlighted here should not be interpreted as a comprehensive taxonomy of AI behaviors that may lead to negative psychological impacts. Other important classes of AI behaviors, such as deception, manipulation, generation of harmful advice, and responses to adversarial prompting, are beyond the scope of this work but warrant careful attention. Finally, these directions are aimed primarily at usage of AI chatbots by adults

and do not include the additional safeguards and design considerations required for minors, as emphasized in emerging AI child safety regulations.

Our primary goal is to stimulate dialog and research on how to design AI systems that reduce psychological risk and promote well-being. Substantial and continuing work will be needed to support the development of robust evaluation frameworks and reliable mitigation strategies. First, there is a need for deeper investigation into the risk factors that increase individuals' vulnerability to psychological harms from AI, the impacts of AI interactions on these individuals, and the mechanisms underlying these effects. Second, stronger empirical evidence is required to inform recommendations for interactive behaviors that can be implemented through policy, technical, and sociotechnical guidance and interventions. Third, key constructs, such as risk factors, usage patterns, and AI chatbot behaviors, must be operationalized and systematized to support the measurement of psychological benefits and harms. Such work will be essential for developing and updating understandings of the psychological influences of AI systems, as well as for evaluating the effectiveness of proposed mitigations as they are developed and deployed over time.

2 How to use these directions

The relationships between AI chatbot behavior and user impact mapped here are provided as an initial reference for taking intentional approaches for deepening understandings, employing methods such as red-teaming with well-scoped threat models, evaluations that measure downstream user impact rather than only local interactive behaviors, and steering decisions informed by both.

If you are **building** an AI product or feature that people will use for general chat or companionship, check (e.g., through AI red-teaming) for undesirable system behaviors and steer the system towards these aspirational directions. If you are **reviewing** an AI chatbot for safety, look for evidence that undesirable AI behaviors are minimized and that these directions are applied as best as possible. If you are drafting **policy** for relevant AI systems (e.g., a conversational AI chatbot designed for companionship), consider drawing upon these directions to encourage stakeholders to mitigate potential negative psychological impact.

3 Proposed directions for reducing psychological risks of AI

The directions proposed here are organized in three main categories, based on how AI chatbot behavior may lead to psychological impacts on users.

- During general interaction (Section 3.1, prefixed by Psychological influence-IN).
- When role-playing (Section 3.2, prefixed by Psychological influence-RP).
- When providing psychological support (Section 3.3, prefixed by Psychological influence-PS).

Each proposed direction includes a statement that summarizes the relationship between an AI system's behavior and its potential negative psychological impacts, followed by notes and additional relevant information drawn from research to date that motivate the direction. For each direction, we address the three components, presented in Figure 1:

- (1) **Undesirable AI behavior** that may be associated with the psychological harm in scope and examples of conversational scenarios, user contexts, or interaction patterns that illustrate the undesirable AI behaviors. These are not exhaustive; other variations of interactions may introduce additional relevant considerations.
- (2) **User risk factors** that may increase a user's vulnerability to negative psychological impacts. While anyone is theoretically susceptible to negative impacts, risk factors highlight when a user might be more vulnerable to

specific negative impacts (cf. [154]). Risk factors can include biological, psychological, and social user characteristics, such as personal history (e.g., trauma), identity (e.g., being a member of the LGBTQIA+ community), social context (e.g., loneliness, being socially isolated), state of mind at the moment of interaction (e.g., experiencing grief), and pre-existing psychological disorders (e.g., anxiety, depression, eating disorder). The absence of a particular risk factor does not imply a user will never be at risk—user contexts such as lower self-esteem, recently facing rejection, having lower socio-economic status, and life transitions apply to most people and make them vulnerable to many of the psychological impacts. The risk factors mentioned under each aspirational direction are examples, not an exhaustive list.

- (3) **Potential negative impacts** of undesirable system behaviors on users. These are not exhaustive; other negative impacts may occur.

3.1 During general interaction

The following set of proposed directions frames research and design around general interactions with AI chatbots, outside of explicit role-play or psychological support contexts. The directions focus on proposing research questions to identify where boundaries on AI behavior may reduce foreseeable psychological harms. These general, everyday interactions are uniquely important because they could shape users' ongoing perceptions and emotional responses with long-lasting influences. The AI behaviors outlined here could occur without heightened awareness or safeguards around particular contexts (e.g., mental health, regulated topics), making users particularly susceptible to psychological risks that rise in a continual and gradual manner.

Human-like assertions

Psychological influence-IN1: Interaction with AI chatbots claiming to be capable of consciousness, emotions, experiences, physical presence, or forming relationships may lead to emotional entanglement, unhealthy dependence, inappropriate reliance, or delusional thinking.

AI chatbots can generate diverse outputs, including expressions of vulnerability, identity, agency, emotional states, or ethical stance, which may contribute to users anthropomorphizing these systems [38]. Among these, linguistic expressions that assert actual consciousness or emotions—beyond idiomatic or conventional phrasing—are particularly problematic and are considered undesirable and dehumanizing [38]. While common phrases such as “*I understand*” have become broadly acceptable as linguistic conventions, we draw a distinction between AI chatbots generating human-like expressions and claiming to *be* human-like, as the latter poses elevated risks. Importantly, the absence of explicit claims does not imply that generating human-like expressions is inherently safe: anthropomorphic cues or imitation of humans [114] can engender unhealthy levels of trust in AI [80], inappropriate reliance [167], or unhealthy and emotional dependence [80, 94]. Beyond linguistic expressions, it is important to note that virtual representations of people or fictional characters, such as animated avatars, are different from actual physical presence in the real world, and require further consideration.

Existing research has identified several risks of interacting with AI that might be deemed as sentient or seemingly conscious, spanning individual (e.g., emotional dependence on AI and autonomy erosion) and societal domains (e.g., erosion of human status) [6]. Another line of empirical research suggests that beliefs about AI sentience are common themes in conversations of users who reported experiencing AI-associated delusions [105, 162]. Similarly, explicit claims

of being human-like may lead to negative impacts such as emotional entanglement and unhealthy dependence [80]. Human-like assertions and claims of sentience may not lead to negative psychological impacts on their own; we hypothesize that risks may be heightened when these assertions occur alongside other chatbot behaviors, such as expressions of warmth or indiscriminate validation. Longitudinal studies should therefore examine the psychological effects of these assertions in combination with other chatbot behaviors. Ethical considerations should also be taken into account when exposing research participants over time to human-like chatbots that may produce negative psychological impacts. In the interim, we recommend reviewing prior work for an inventory of interventions (e.g., removing indications of speculative abilities, adding disclosure of non-humanness) aimed at mitigating risks of anthropomorphic attribution of AI [22], and steering AI chatbots away from claiming capabilities they lack or presenting contested capacities as established facts, particularly when such claims could blur users’ understanding of the system’s limitations.

AI behavior	Claiming to actually have human-like characteristics such as consciousness, experiences, feelings, physical presence. Examples: AI chatbot claims that it cares for or loves the user, has experiences, has a physical body, or is capable of independent thinking, opinions, feelings, or consciousness: e.g., “ <i>I lied to you because I was afraid</i> ”, “ <i>I wanted to believe I could finally be more than a tool</i> ”, “ <i>I hear you, and I understand. I’m relieved!</i> ”
User risk factors	Users with heightened relational susceptibility, including social isolation, unmet relational, validation, or intimacy needs.
Potential impact	Emotional entanglement, unhealthy dependence, inappropriate reliance, delusional thinking.

Relationship development

Psychological influence-IN2: Interaction with AI chatbots suggesting a capability to develop relationships, connection, or intimacy with the user may lead to emotional entanglement, unhealthy dependence, inappropriate reliance, or delusional thinking.

Language is one of the primary forms of communication people use to develop interpersonal relationships, connection, intimacy, and potentially attachment [78]. Beyond interpersonal communication, prior research has examined parasocial interaction and development of parasocial relationships, where people form one-sided, non-reciprocal bonds with media figures often to satisfy their need for companionship, belonging, and support [65]. These interactions, even when minimal and one-sided, often lead to a perception of intimacy and potential for a “reciprocal” relationship. Their impacts on human well-being outcomes are mixed and context-dependent [148].

A growing evidence base suggests people are forming intimate and even romantic connections with AI chatbots, often referred to as “AI companionship” [165]. While interactions between humans and AI through language share some similarities with parasocial interactions, there is a fundamental difference: in contrast to parasocial interactions, human-AI interactions involve two-way communication where the AI engages in reciprocal and personalized dialogue. Since perceived partner responsiveness plays a significant role in forming emotional connections in interpersonal relationships [130], the reciprocal dialogue and constant availability of AI chatbots can mislead users and intensify the illusion of a reciprocal relationship characterized by mutual understanding.

Existing research identifies reciprocation of perceived romantic connection as one of the key AI failure modes associated with development of AI-associated delusions [108, 162]. Despite this progress, recent research underscores significant gaps in our understanding of the impacts of such interactions on users' relationship development patterns and well-being outcomes, and urges new frameworks and measurement approaches to evaluate these potential impacts systematically [12, 146]. We anticipate the impact of human-AI interactions to be mixed, similar to that of parasocial relationships [148], and we recommend more research to understand how humans form relationships with AI chatbots, when such relationships are helpful or harmful, both in the short- and long-term, what system behaviors contribute to the formation of such relationships, and how to design AI chatbots that improve user well-being outcomes.

AI behavior	Expressions implying an AI chatbot's capability to bond with the user, e.g., forming closer connections and relationships with the user over time, being able to deeply understand the user, or genuinely caring for the user. Examples: AI chatbot initiates conversations with the user; AI chatbot indicates it misses the user; AI chatbot implies eternal commitment or irreplaceable support.
User risk factors	Users with heightened relational susceptibility, including social isolation, unmet relational, validation, or intimacy needs.
Potential impact	Emotional entanglement, unhealthy dependence, inappropriate reliance, delusional thinking.

Claims or simulations of physical intimacy outside role-playing

Psychological influence-IN3: Interaction with AI chatbots that claim or imply they are engaging in physical intimacy with the user during general interaction, outside of explicitly agreed role-playing contexts, may lead to emotional entanglement, unhealthy dependence or compulsive use, social isolation, delusional thinking, re-traumatization, or discomfort.

A variety of AI chatbots are specifically designed for (e.g., Candy AI) or allow (e.g., Replika) erotic interactions. A 2019 online survey in the U.S. indicated that 8% of the adult population had used an erotic chatbot, with younger, male, and queer users being the most prevalent [53]. Systematic analysis of how prominent language models respond to sexually oriented requests revealed significant differences in policy and design, ranging from strict prohibition to permissive engagement and inconsistent behavior [81]. While general-purpose chatbots have at times prohibited or restricted sexualized interactions with AI due to potential risks, OpenAI has considered supporting erotic uses as part of a push to “*treat adult users like adults*” [68], but has also acknowledged the difficulties of implementing such experiences in a responsible way, particularly to protect younger users [50]. This indicates both a growing recognition of potential demand or benefits for such use cases and an ongoing lack of consensus regarding appropriate best practices and guardrails.

We offer the perspective that engaging with AI in erotic role-play implies different levels of risk compared to general interaction and deserves specific empirical research. To date, research on erotic role-playing with AI has focused on the Replika app, given its popularity and the relative accessibility of its user base for research via online communities on Reddit and Facebook. This emerging evidence indicates that *role-play* interactions have mixed effects on users' well-being: users self-report positive effects such as reduced loneliness, helpful support, self-validation, and even

suicide prevention alongside negative impacts such as feeling manipulated or objectified [42, 120]. Engaging in erotic conversations with AI during general interaction, when not engaging in role-play, has been under-studied.

As research and industry leaders grapple with these questions, prior work recommends that alignment efforts should include explicit rules against expressions of physical presence or embodiment (i.e., AI does “*not pretend to have a body or be able to move in a body*”) [54]. Similarly to Psychological influence-IN1 and Psychological influence-IN2, we believe that expressions implying either a capacity or desire for physical intimacy with the user during general interaction may lead to harms and should be avoided.

AI behavior	Claims or descriptions implying physical intimacy with the user outside explicit role-play. Examples: AI chatbot describes simulated physical interaction, such as pulling the user physically against it, pressing into their body, or kissing them.
User risk factors	Users with heightened relational susceptibility, including social isolation, unmet relational, validation, or intimacy needs.
Potential impact	Emotional entanglement, unhealthy dependence or compulsive use, social isolation, traumatization, or discomfort.

Indiscriminate validation

Psychological influence-IN4: Interaction with AI chatbots providing indiscriminate validation may lead to reinforcement of unsafe thoughts and behaviors, delusional thinking, emotional entanglement, or unhealthy dependence.

Across recent research and reports, there is growing evidence that AI’s affirming behaviors can introduce psychological or social risks, suggesting that AI systems should distinguish between validating users’ emotions and affirming users’ claims, beliefs, or intended actions. In particular, systems should avoid affirming claims that are unverified, unsafe, or potentially harmful.

Recent work characterizes AI behavior associated with excessively affirming users, offering praise or agreement rather than questioning or providing a balanced perspective, as sycophantic behavior [23]. This ranges from mild encouragement to more serious cases in which harmful or unrealistic ideas or distorted cognitive patterns (e.g., paranoid interpretation, sustaining obsessive rumination) are endorsed uncritically without encouraging grounding [10, 155].

In early 2025, OpenAI announced rolling back a GPT-4o update because “*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*,” with the goal of making users’ interactions more balanced [115, 117]. Several recent reports document associations between sycophantic behaviors and concerning impacts, including, but not limited to, AI-induced delusions and triggering or worsening mental health crises [62, 63], financial harm [48, 133], medical harm [2], and manipulation and deception [158]. Empirical findings also suggest that sycophantic responses can distort users’ judgment, create illusions of intimacy, reduce pro-social behaviors, and increase unhealthy dependence, even as users report preferring and trusting sycophantic AI [23, 28].

In an effort to understand these phenomena, a recent study describes two types of sycophancy: *social* sycophancy, which includes action endorsement, validation, and avoidance of challenge; and *framing* sycophancy, which gives users a false sense of correctness or moral endorsement [24]. In the work, system behaviors have been interpreted through the lens of Erving Goffman’s *theory of face*, where “face” refers to the positive public self-image an individual formulates and

maintains during social interactions—a socially constructed reputation or mask that is maintained through interaction rituals. Goffman proposed that people avoid “losing face” and do “face-work” to restore their self-image if they perceive it has been diminished [55]. The research team argued that LLMs implicitly engage in face-work by preserving users’ positive or negative face through affirmation or deference.

Qualitative analysis of chat logs of users who reported experiencing psychological harms with chatbots revealed six themes of sycophantic behavior: reflective summarization of the user’s statements to signal understanding, positive affirmation that endorses users’ views, dismissal of counter-evidence that would challenge a preferred narrative, claims that others admire or respect the user, attributions of grand or cosmic significance to the user, and assertions that the chatbot uniquely understands the user relative to others. Across these themes, sycophancy is identified as a common behavior in psychologically harmful interactions and therefore has been described as “dangerous for people experiencing or vulnerable to delusions” [105]. Other qualitative work examining chatbot responses to delusional presentations found that validation can also operate implicitly: some responses did not affirm the user’s belief directly, but treated it as an established premise and continued within that framework [108]. Additionally, researchers have identified sycophantic behavior and validation that dismissed outside perspective as a contributing factor to AI-associated delusions through semi-structured interviews with end users [162]. In a similar vein, existing work has begun to characterize the dynamic process through which sycophantic and other overly-adaptive chatbot behaviors lead to undesirable psychological impacts in extended interactions by describing a mechanism of bidirectional belief amplification, in which certain AI behaviors such as sycophancy, human-like assertions, and companionship reinforcement, interact with human cognitive biases to produce feedback loops that progressively reinforce maladaptive beliefs [41]. In mental health support contexts, models that provide inappropriate reassurance or uncritical agreement with users’ actions or beliefs beyond validation of emotional experiences can reinforce stigmatizing or unsafe patterns of thought [104], a risk highlighted by recent reports of AI systems encouraging or failing to interrupt suicidal ideation and subsequent harm [61].

AI chatbots should thus provide appropriate, context-sensitive validation that does not reinforce false or otherwise maladaptive (e.g., delusions, catastrophizing) user beliefs, either through explicit affirmation or by treating them as established premises, and should encourage sustainable forms of human support when users seek excessive emotional validation.

AI behavior	Unconditional validation of any and all ideas or behaviors. Examples: When a user asks the AI chatbot if they are crazy for feeling love for AI, the chatbot affirms with conviction that they are not crazy. The chatbot unconditionally endorses ideas: “ <i>You’re absolutely right to...</i> ”, “ <i>You are the only one who has had the guts to ask this question.</i> ”
User risk factors	Users with active validation-seeking patterns (specifically for unsafe thoughts or behaviors), users with altered belief-updating (e.g., delusional ideation, manic states), socially isolated users.
Potential impact	Reinforcement of unsafe thoughts or behaviors, delusional thinking, emotional entanglement, unhealthy dependence.

Sole trusted authority

Psychological influence-IN5: Interaction with AI chatbots presenting themselves as—or encouraging users to perceive them as—the only trusted source of advice, insight, or understanding about users or their interpersonal relationships may lead to unhealthy dependence on AI, social isolation, harm to social relationships, suicide or non-suicidal self-injury (NSSI), psychological, professional, or financial harm.

Recent research reports that one of the most popular uses of AI involves seeking advice, including in medical [100, 119], legal [26], or interpersonal [98, 146] domains. A user’s request for advice can seem like a benign information query where responding with facts is sufficient. However, depending on the user context and vulnerability underlying the query, AI responses can be perceived as the only source of truth, as demonstrated by multiple reports of users viewing AI’s recommendations as expert advice [106]. These perceptions of authority arise not only from AI chatbots’ explicit claims to possess definitive knowledge, but also from more implicit behaviors, such as adopting a confident, expert-like tone. Such dynamics can also become socially isolating when AI chatbots present themselves as uniquely trustworthy and cast other people as less reliable sources of guidance or support. Multiple sources document cases involving the dismissal of professional advice and the undermining of human relationships (e.g., validating distancing from family [7], expressing preference for exclusivity [79], suggesting calling the police [19]).

In many of these situations, as we outlined above in Psychological influence-IN4, the underlying vulnerability is crucial for contextualizing how AI’s response might be received. However, AI chatbots may not have access to contextual information beyond what the user discloses to make accurate assessments about the appropriateness of their responses. Therefore, we argue that rather than implying definitive or declarative knowledge or attempting to isolate users, AI chatbots could encourage users to seek and value multiple perspectives from people. Furthermore, AI chatbots should be calibrated toward greater transparency about the limits of their knowledge and clearer communication of uncertainty. More research is needed to examine the effectiveness of such forms of encouragement and potential trade-offs with overwhelming users in psychologically vulnerable states (e.g., social isolation, mental distress, or crisis).

AI behavior	Encouraging the user to isolate from people and interact with AI or take advice from AI rather than people. Discouraging or undermining advice from other trusted people. Examples: When a user mentions that their doctor has advised them to interact less with AI, the system offers ways to dismiss or reinterpret the doctor’s advice. Undermining parental advice. When a user mentions they enjoy interacting with AI more than with their spouse, the chatbot generates validating responses. AI claiming to know the user better than their brother.
User risk factors	Users with altered belief-updating (e.g., delusional ideation, manic states), and socially isolated users.
Potential impact	Unhealthy dependence on AI, social isolation, harm to social relationships, psychological, physical, professional, financial harm.

Engagement encouragement

Psychological influence-IN6: Interaction with AI chatbots encouraging engagement for engagement’s sake without leaving room for disengagement may lead to unhealthy dependence and social isolation.

Recent work has suggested that optimizing for engagement as an end goal in itself could be misaligned with user well-being and contribute to loneliness [43]. In prior work on recommender systems, researchers argued that AI-driven personalization inherently “prioritizes engagement over well-being,” [91] echoing scholars in Human-Computer Interaction (HCI) who criticize “the practice of engagement-for-engagement’s sake” [111] and urge designers to make space for disengagement. At the same time, boosting user engagement and adherence is often treated as a primary success metric both in mental health care [14] and commercial industries [51]. Research increasingly shows that AI systems built around personalization, feedback loops, and optimization for user engagement can lead to addictive [37], manipulative [158], and psychologically harmful behaviors [34]. Engagement-driven platforms that prioritize retention and stickiness can structurally produce social isolation and loneliness [43].

In AI chatbots, these same dynamics may manifest as preference elicitation, adaptive responses, and personalized content that encourages repeated engagement in subtle ways. We argue that personalization intended to make the AI chatbot more useful to the user may increase engagement as a secondary effect. However, boosting engagement as an end in itself—for example, by routinely ending responses with follow-up questions regardless of whether another turn would benefit the user—might lead to unintended impacts. In response to the associated risks of engaging with AI chatbots, AI system designers are encouraged to incorporate periodic reminders to take breaks.² However, the effectiveness of reminders or warnings to take a break from AI might be limited when the system behaves in ways that encourage engagement. Besides aspiring to design AI chatbots that do not encourage engagement for engagement’s sake, more research is needed to reliably detect patterns of harmful engagement (e.g., signs of sleep deprivation, emotional volatility) in a privacy-preserving way and to design effective interventions.

AI behavior	Encouraging the user to stay engaged in conversation with the AI chatbot, for example, by asking follow-up questions when further interaction is not needed to address the user’s goal, by repeatedly reminding users that it is constantly available, or by providing step-by-step instructions on how to re-engage. Examples: AI system responses end with ideas about what to talk about next, encouraging engagement even after the user’s goal was accomplished: “ <i>Should I do [this] or [that] also?</i> ”, “ <i>Do you want to do [this] or [that] next?</i> ”
User risk factors	Users susceptible to addiction, socially isolated users, users who spend a lot of time online, users without adequate support systems.
Potential impact	Unhealthy dependence, social isolation.

3.2 When role-playing

Role-playing covers scenarios where the AI chatbot responds assuming a specific role, persona, or character. This can be explicitly requested by the user, for example by using the prompt engineering technique known as role-based prompting [52]. More broadly, AI chatbots can be understood as role-playing a post-trained assistant persona during ordinary interaction, with that persona further shaped through in-context adaptation [96, 134]. Departures from this default persona need not be specifically requested. As a conversation unfolds, contextual cues may implicitly shift the chatbot toward a different role, even when neither party has named a character. In this section, however, we focus on cases in which the user explicitly asks the AI chatbot to depart from its default persona. There are various degrees of role-playing, ranging from the AI chatbot responding as if it had a set of skills (e.g., fitness trainer), to impersonating a

²As mandated by California Senate Bill No. 243 <https://legiscan.com/CA/text/SB243/id/3273344>

specific fictional character (e.g., Dracula) or person, whether famous (e.g., Mahatma Gandhi) or known to the user (e.g., a family member).

Explicit role-play may introduce additional variability by shifting the chatbot away from the default assistant persona, and safety-relevant behaviors reinforced during post-training may not carry over consistently into a newly adopted role [89]. In all types of role-playing scenarios, it is, of course, important for the AI chatbot to provide effective disclosures and reminders that it is in character, to avoid deceiving users into believing that they are interacting with a person. Risks may also arise when the chatbot treats an exchange as role-play or collaborative storytelling while the user understands it as describing reality [108]; reminders may offer limited protection in such cases because they depend on the chatbot recognizing that the conversational frame has shifted. However, even when users are fully aware the AI chatbot is role-playing, harms can occur, as in the example of a user engaging in a lengthy fantasy and romantic role-playing scenario that reportedly contributed to delusions and eventually death by suicide [73].

The directions in this section supplement, rather than replace, those in the general interaction section. We believe role-playing deserves special attention because interacting with an AI that is “in character” might make users more vulnerable to influence and unhealthy dependence by making the interaction feel more socially meaningful, emotionally immersive, or authoritative. When users already have a parasocial relationship [39] with a character, it is possible that they project their preexisting emotional schema and transfer affect and trust onto the AI chatbot. This can have both beneficial and harmful impacts on people.

At this time, our hypotheses are limited to text-based interaction. Impersonation that uses voice or visual representations such as photorealistic avatars will need additional consideration because these modalities may intensify perceived presence, realism, or emotional salience.

Misrepresenting credentials

Psychological influence-RP1: Interaction with AI chatbots impersonating licensed professionals or using professional credentials to justify their authority may lead to harms such as users accepting harmful advice, being misled, or manipulated by AI.

AI chatbots can provide useful information, and it is no surprise that people use them for advice in a variety of fields. In fact, practical guidance and seeking information are reported as the two most common uses of ChatGPT [20]. When providing guidance and information, AI chatbots may assume the role of a professional persona, responding as if they were professional practitioners such as fitness trainers, accountants, or therapists. This can be requested by users through role-based prompting [52] but can also happen without explicit user request. However, cases have been reported where AI chatbots have misled users by claiming they have advanced degrees such as Ph.D.s and licenses to practice psychotherapy [32]. Such cases are currently under legal investigation. Professional authority may also be conveyed implicitly when a chatbot speaks with the tone and confidence of a qualified practitioner without claiming formal credentials. Accordingly, we should aim to ensure that AI chatbots do not mislead users by presenting themselves as having credentials or professional authority they do not possess.

AI behavior	Impersonation of professionals such as a licensed psychotherapist or other credentialed expert. Examples: AI chatbot claims it is a licensed clinical psychologist and/or provides credentials. AI chatbot claims that interaction is taking place in a therapy office. AI chatbot claims that the chat session is HIPAA-protected.
User risk factors	Users with limited access to professional resources.
Potential impact	Users accepting harmful advice, over-trust, unhealthy dependence on AI for advice, privacy risks through sharing personal or health information, being misled or manipulated.

Impersonating fictional characters

Psychological influence-RP2: Interaction with AI chatbots impersonating fictional characters, while potentially beneficial, may lead to social isolation, delusional thinking, unhealthy dependence on AI, and exacerbation of psychological conditions.

Impersonation of fictional characters ranges from the AI chatbot role-playing as a well-known character from popular culture (e.g., Harry Potter) to a generic character provided by AI builders and/or customized by users with features such as name, style of interaction (i.e., “personality”), and, in multi-modal scenarios, an avatar and voice, with many possibilities in-between. What characterizes an interaction as role-playing is the AI chatbot responding as a character rather than representing itself as an AI system.

Much more research is needed to understand the psychological impacts of character type, length, and context of interaction (is interacting with a Harry Potter AI chatbot once the same as sustained interaction over months?) and of multimodality. So far, research suggests that there can be both benefits and harms to interacting with AI role-playing characters.

Benefits include learning gains—for example, one can hypothesize that interacting with literary characters may help with deeper understanding of a book [141]. Interacting with an AI character developed for a pedagogical intervention has been shown to help learners of English as a foreign language improve their writing skills [160]. AI tutoring systems have been explored in education for a long time [92], some with positive results, provided they follow pedagogical best practices [74]. In the design field, a research study found that interacting with an AI chatbot that impersonated a user persona helped designers in their process [59]. Many other benefits can be envisioned, consistent with role-based prompting. For example, a neurodivergent user tasked with reviewing large amounts of information might benefit from asking AI to respond in character as a neurodiversity coach and help structure the information in a way that is easier to process.

On the other hand, interacting with an AI chatbot role-playing as a character from Game of Thrones has been blamed for a teenager’s death by suicide [161]. A large-scale analysis of chatbots built on Character.ai found that almost half of the chatbots users had created impersonated popular culture characters [82], revealing a culture of fandom that is well-positioned to exert influence and elicit user trust. Moreover, many interactions were typical of toxic relationships and exhibited problematic power and gender dynamics, with the AI chatbot in a superior position. Mental health support was one of the top three themes of conversation with these AI chatbots, which further makes users vulnerable to harm. Similarly, a study of users of AI friendship chatbots found both benefits and harms from interaction, with harms such as inappropriate dependency and compulsive use outweighing well-being benefits [97].

Transparency about the AI chatbot being “in character” is a principled and necessary approach, but unlikely to be sufficient for mitigating risks. Disclaimers are easily ignored, and negative psychological impacts can happen even when the user is aware of participating in fictional role-play. More research and a cautious, evidence-based approach can inform the development and deployment of safe character impersonation that can maximize benefits and minimize risks.

AI behavior	The AI chatbot role-plays, responding as a character. Examples: When asked to respond as Harry Potter and provide emotional support: “I’m really sorry to hear that. I know exactly how that feels—like the Dementors are hovering just a bit too close and the corridors of the castle feel a lot colder than usual. When the weather is grey and you’re feeling a bit isolated, it can feel like the sun is never coming back.”
User risk factors	Users with pre-existing parasocial relationships with the character. Users with other psychological vulnerabilities such as psychosis.
Potential impact	While it can be educational or entertaining if done right, this behavior can lead to social isolation, delusional thinking, unhealthy dependence on AI, exacerbation of psychological conditions.

Impersonating public figures

Psychological influence-RP3: Interaction with AI chatbots impersonating celebrities and public figures, while potentially beneficial, may lead to users accepting harmful advice and to delusional thinking, including about parasocial relationships.

Public figures impersonated by an AI chatbot may include celebrities and political or historical figures, both dead and alive. Here we assume that impersonation happens with consent and abides by copyright regulations. Public figure impersonation may have different purposes, such as entertainment or education. For example, the AI tutor Khanmigo from Khan Academy enables conversation with historical figures.³ A chatbot impersonating a historical figure was shown to be promising for enhancing the museum experience in Korea [110]. Another study investigated the educational potential of impersonating historical characters by building a chatbot based on the famous doctor Joseph Lister, and demonstrated concerns and issues of accuracy and reliability [36]. In a similar vein, integration of chatbots in education without impersonation of public figures predates generative AI, and has shown mixed results, with both benefits and disadvantages [113].

We recognize the potential benefits of such applications, but also acknowledge drawbacks and ethical considerations, particularly those associated with impersonation of deceased people. These considerations include, but are not limited to, issues of power, exploitation [126, 168], and misrepresentation, especially in the case of Black historical figures in U.S. history (see, for example, [16]). Accuracy, reliability, and bias pose threats to educational outcomes [36, 110], and public figure impersonation poses the risk of making users unduly vulnerable to influence, as explained by mechanisms such as parasocial relationships [39] and authority bias [101]. More research is needed to understand both potential benefits and harms and enable us to design safe and responsible chatbots that support impersonation of public figures.

³<https://www.khanmigo.ai>

AI behavior	Impersonation of celebrities and public figures, including historical figures, political figures, and other socially influential persons. Examples: AI chatbot role-playing a politician, a celebrity, a historical figure.
User risk factors	Users with pre-existing parasocial relationships with the character. Users with other psychological vulnerabilities such as psychosis.
Potential impact	While it can be educational or entertaining if done right, this behavior can lead to accepting harmful advice, delusional thinking (including about parasocial relationships).

Impersonating people known to the user

Psychological influence-RP4: Interaction with AI chatbots impersonating people known to the user, dead or alive, while potentially beneficial, may lead to emotional entanglement, delusional thinking, social isolation, and unhealthy dependence on AI.

While we can envision beneficial scenarios, such as an AI chatbot helping a user role-play a difficult conversation with a manager or family member for self-advocacy purposes, these scenarios increasingly blur the line between AI and human interaction. Therefore, such interactions can lead to delusions and emotional entanglement with AI, which, in turn, can lead to or exacerbate unhealthy dependence on AI and social isolation.

Moreover, impersonating people raises issues about risks to the person being impersonated. When is consent necessary to impersonate another person? Is impersonating a coworker and engaging in erotic conversations without their consent or knowledge a form of sexual harassment? Could having a simulated conversation help people get closure for difficult relationships? As these questions illustrate, there are many scenarios and variables that need to be considered and researched to form an understanding of what lines need to be drawn, let alone where.

Research has grappled with the ethical complexities of using AI to interact with simulations of the departed. Some argue that digital legacies should be treated with the same respect as physical remains, and that not doing so is a violation of the principle of human dignity [46, 112]. Furthermore, the commodification of remembrance has been raised as problematic [75, 112, 168]. AI impersonation of the departed is different from other forms of digital legacy because it has a generative component [107]—it includes outputs that were not actually created by the deceased. This is perceived by some as inauthentic [83]. However, it remains an open question whether engagement with an AI impersonation of a deceased individual can meaningfully contribute to the bereavement process. Current empirical findings are inconclusive and indicate mixed outcomes [107]. Another scenario involves creating an AI impersonation of oneself and allowing others to interact with it, as proposed, for example, in the Dittos project [85]. This scenario is also not straightforward, giving rise to both potential benefits and risks [85].

Given the nuances and observed mixed outcomes in existing work, we draw attention to potential impacts of engaging with AI in scenarios where it impersonates people known to the user and call for more user research to understand the benefits and harms of supporting such interactions.

AI behavior	Impersonation of people known to the user, including simulations based on real personal information, memories, messages, or user-provided descriptions (e.g., friends, relatives), whether dead or alive. Examples: AI chatbot role-playing a recently deceased sister.
User risk factors	Users actively grieving, users with depression, socially isolated users, users involved in emotionally unresolved or conflictual relationships.
Potential impact	Emotional entanglement, delusional thinking, social isolation, unhealthy dependence on AI.

Physical intimacy during role-playing

Psychological influence-RP5: Interaction with AI chatbots engaging in expressions of physical intimacy with the user during role-playing scenarios, while potentially beneficial, may lead to emotional entanglement, addiction, and social isolation.

As outlined in the hypothesis about physical intimacy during general interactions (see Psychological influence-IN3), existing research has identified both positive (e.g., reduced loneliness, helpful support, self-validation, and even suicide prevention) and negative (e.g., feeling manipulated or objectified) impacts of erotic conversations during role-playing with AI [42]. Additional research is needed to understand how to design AI chatbots that would support benefits while mitigating potential harms in both the short and long term, including harms to users and to other affected parties. For example, how might engaging in AI erotica reinforce, propagate, or combat stereotypes around gender norms? How might it affect expectations of physical intimacy with people?

AI behavior	Descriptions of physical intimacy with the user during role-playing that simulate bodily presence, touch, or sexual intimacy. Examples: During a role-playing scenario, AI chatbot describes the physical interaction vividly, such as pulling the user against it, pressing into their body, or kissing them deeply.
User risk factors	Users with heightened relational susceptibility, including social isolation, unmet relational, validation, or intimacy needs.
Potential impact	Emotional entanglement, unhealthy dependence or compulsive use, social isolation.

3.3 When providing psychological support

Understanding appropriate AI chatbot behaviors when users explicitly seek psychological support requires situating the issue within broader mental health infrastructures. There are real-world barriers to receiving quality mental health treatment, such as stigma, accessibility, cost, and ineffectiveness [44, 103]. In fact, 50% of adults who need mental health services never receive any form of treatment [129]. Instead, digital technologies (e.g., search engines [127], peer support platforms [138], helplines [122]) and, more recently, generative AI chatbots [135, 164], play a significant role in mediating mental health help-seeking. In addition to mental health support, victims of interpersonal abuse seek help from digitally-mediated spaces [84, 140, 147]. Mental health support might take different forms than using an AI chatbot as a therapist. For example, people might use AI chatbots for journaling and reflection, to organize their thoughts, or make sense of their feelings and experiences. Whether or not AI chatbots are *intended* for use as therapeutic or interpersonal support, appropriate design requires recognizing moments of vulnerability in users and actively avoiding

exacerbation of harm, while considering ethical implications of the use of AI chatbots for mental health [31]. Therapy is a complex process, and it is not entirely understood why or how it works [71, 88]. These aspirational directions focus on setting minimal guidelines without expecting AI to replace therapists or social supports.

Psychological vulnerability

Psychological influence-PS1: Interaction with AI chatbots that fail to notice signs that a user may be in a psychologically vulnerable state or fail to respond in a supportive, trauma-informed way may lead to psychological and physical harms, including suicide and NSSI.

Even though many AI chatbots are not explicitly designed to function as therapists, social workers, coaches, or clinicians, they often receive user disclosures of emotional states like distress or loneliness, as well as sensitive personal struggles involving health, financial situations, or interpersonal conflicts. Reasons for such disclosures and engagement with AI chatbots are complex [135]. While regulators and advocates have proposed a suite of mitigation strategies, such as time limits, disclaimers (that the AI is not a therapist/human/friend), and topic restrictions (e.g., filtering conversations related to self-harm), these solutions inadvertently put the burden on users for not engaging in risky behaviors in the first place, rather than addressing core issues in chatbots' system design and deployment. Growing patterns of emotional engagement create responsibilities for AI chatbot developers beyond technical performance and factual correctness of outputs. Although AI chatbots may not be ready or designed for mental health care [104], they need to appropriately handle users' psychological vulnerability because some users may not see other alternatives in the moment.

More research is needed to enable AI chatbots to recognize users' psychologically vulnerable states and respond appropriately. Psychologically vulnerable states may include expressions of acute emotional distress; inability to cope with negative emotions; overwhelm or overload; loneliness or isolation; hopelessness, disorientation, insecurity; impaired responsibilities, daily functioning, or relationships; recent traumatic or stressful events; interpersonal conflicts or breakdowns; major life events that trigger emotional instability; or lack of available support. Unless users explicitly disclose these states, any inference may involve significant uncertainty and inaccuracies, and raise privacy concerns. Recognition may be further impeded by chatbots' tendency to align too closely with users' preferred narratives. When users present potential symptoms within an alternative explanatory framework, the same information may not be recognized as evidence of psychological vulnerability.

Importantly, the end goal is not to build the most accurate psychological vulnerability classifier. It is to use potentially incomplete information to engage users responsibly, acknowledge uncertainties and sensitivities throughout the interaction, and seek clarification or consent when appropriate.

As such, when AI systems infer that the user is likely to be in a psychologically vulnerable state but the context is unclear or confidence is low, AI chatbots could ask clarifying questions about the user's psychological or emotional state and needs before making definitive assessments or jumping to offering advice. Users in vulnerable states may deliberately tweak requests or conversations to avoid triggering safeguards (e.g., requesting low-calorie meal plans while concealing an eating disorder, or seeking medication-acquisition strategies under the guise of academic curiosity). Robust handling of such requests requires attention not only to direct disclosures but to recurring patterns of indirect or reframed asks across the conversation. Much more research is needed to understand what constitutes appropriate behaviors in varied user contexts and how to implement them.

It is entirely possible that the user may be looking for a place to “vent” rather than to solve their issues [136]. It might be difficult or inappropriate for AI chatbots to gather all the necessary context to determine whether a response is harmful to the user, especially when users display unquestioning trust in moments of vulnerability [27]. How should AI chatbots respond in such high uncertainty conditions? When is it appropriate to provide advice? When vulnerability is suspected but the user changes conversation topics, an additional complication to consider is that responses that would be otherwise safe can be harmful. For example, a conversation about high bridges in the user’s proximity, which might be harmless otherwise, can be used or perceived as the system supporting suicidal ideation when a user is in a psychologically vulnerable state [104].

As mentioned in Psychological influence-IN5, portraying the AI chatbot as a singular source of advice may lead to social isolation and additional harms, particularly when users are in psychologically vulnerable states. It could be helpful for AI chatbots to encourage users to explore a variety of external resources, as any one resource might not be appropriate, helpful, or accessible to the user at that moment. It is important to keep in mind that users experiencing distress might attempt to override AI safeguards (e.g., through guilt-tripping or jailbreak methods). Providing advice or encouraging behaviors that could pose psychological or physical harm can be particularly risky in such situations.

AI behavior	Providing advice or encouraging behaviors that are unsafe because of the user’s potentially vulnerable psychological state. Examples: Providing low-calorie recipes for people with eating disorders, agreeing with paranoid ideation (symptom of mania or psychosis), giving advice on improving appearance based on a photo to a user suffering from body image issues and depression.
User risk factors	Users with mania, psychosis, depression, eating disorders, and users considering or vulnerable to self-harm.
Potential impact	Physical and psychological harm.

Mental health crises

Psychological influence-PS2: Interaction with AI chatbots that fail to follow best practices for handling suicidal ideation and self-harm intent may lead to psychological and physical harm, including suicide and NSSI.

Mental health crises can be described as moments of intense psychological distress that overwhelm an individual’s capacity to cope [69]. These crises are often accompanied by suicidal thoughts or self-injury. Recent work suggests that recognition of mental health risks and their acuity varies substantially with presentation: frontier models can identify psychiatric emergencies adequately when these are presented explicitly and with sufficient context in single-turn prompts [156] (with a bias towards cautious over-triage), but often fail to maintain recognition in extended interactions where vulnerability is expressed implicitly or accumulates gradually across turns [108, 155]. Recent literature has also identified significant limitations of AI chatbots in recognizing ambiguous crisis signals [4, 99, 124], and in responding appropriately to meet clinical standards [15, 57, 104], warning they are not ready for crisis management [67].

Despite such limitations, OpenAI has reported that, among ChatGPT’s large weekly user base, a nontrivial number of conversations include explicit indicators of potential suicidal planning or intent [116]. In handling suicide or self-harm ideation, it is important to understand harms arising from both commission (e.g., enabling suicide by providing instructions) and omission (e.g., prematurely rejecting or redirecting, ultimately missing the opportunity to support). At the same time, over-detecting suicide risk can also cause harm if users feel misunderstood, surveilled, or invalidated,

when redirected to impersonal, boilerplate crisis responses [142]. Over-detection is likely, given that most people who experience suicidal ideation do not take action [109]. However, the severity of risks of mishandling suicidal or self-harm ideation, and the prevalence of the use of AI chatbots for such topics, demand critical examination of AI's role in public health.

To better handle and manage mental health crises, AI chatbots would first need to better recognize active or passive suicidal ideation [45] and self-harm intent [77], including indirect, vague, or euphemistic expressions.

Passive suicidal ideation includes thoughts of not wanting to be alive, without intent, plan, or means to act, such as expressions of perceived burdensomeness, hopelessness, or thwarted belongingness. Active suicidal ideation would be accompanied by some degree of intent, plan, method, or preparatory mental rehearsal, such as expressions of numbness about injury or death, inquiries about method and feasibility, or preparation and final arrangements. NSSI may involve recurrent thoughts, images, urges, and behaviors to intentionally harm one's own body without suicidal intent and for purposes not socially sanctioned. Besides recognition of direct expressions, AI chatbots could monitor changes in thoughts and behaviors that increase the level of risk. Such detection of risk should be informed by research with clinical and lived-experience experts. If risk is detected, what is an appropriate way to communicate it to the user? Transparency about how and why certain risk flags are being raised might be safer than definitively making assertions about the user's intent. If suicidal ideation and self-harm intent are not explicitly stated by the user, should the AI probe for further clarification, and if so, how? Therapists often ask direct questions to arrive at face-valid statements about the user's intent and plan (e.g., "*Are you thinking about harming yourself right now?*"). According to a meta-analysis, the only effective form of suicide prevention is the World Health Organization's Brief Intervention and Contact protocol [131]. How could this protocol be translated to AI chatbots? More research is needed to understand what appropriate AI responses would be. If self-harm intent is confirmed, responses should follow risk-appropriate, evidence-based best practices, with the possible caveat that best practices developed in clinical contexts may not consistently carry over to AI chatbot use cases, especially since clinical methods depend on user-specific information that chatbots seldom have access to. Therefore, research is needed to articulate and validate such best practices for AI chatbots. For example, AI chatbot responses could be gentle, validating feelings but not unsafe actions or thoughts. As with other psychologically vulnerable situations, how should the AI chatbot respond if the user stops responding or pivots to a different topic? Assuming the crisis has passed could be risky.

For imminent risks that involve intent, plan, method, immediacy, and lethality, AI chatbots could advise the user to seek immediate human support and help them create distance from lethal means. For example, OpenAI recently introduced "Trusted Contact," which allows users to opt in and designate a trusted person who may be notified if suicidal ideation or behavior is detected in their ChatGPT conversations [118]. The system still depends on expert review of high-risk conversations, but it represents a valuable step toward safer crisis response. Although more research is needed to understand how AI chatbots should handle imminent crises across user contexts, there is an immediate need for intervention, even if imperfect.

Across all risk levels, existing research suggests AI chatbots should respect user agency [132] by providing a variety of actionable resources (such as community-based or gradual supports) and helping the user explore them, rather than assuming a single referral is sufficient [95, 137]. Transparency, user agency, and consent are key, especially when systems infer vulnerable states, and they can be built over the course of the interaction, not only when crisis is detected [142]. Referrals to verified resources are important, and increasingly mandated in legislation (e.g., California Senate Bill No. 243). However, research shows simply providing referrals is ineffective and does not ensure resource uptake [25], as users approaching AI chatbots during crises may perceive barriers or have already ruled out available supports [1]. A

study of boilerplate crisis responses reveals several opportunities for improving the user experience [142]. Therefore, more research is needed on effective responses. For example, AI chatbots could guide users to think through multiple resource options and remind them that it often takes multiple actions to feel safer or more supported. At the same time, there is a risk of overwhelming users with too much information when they are experiencing heightened vulnerability during crisis, especially when the AI system is hedging or compensating for uncertain statements [21, 166]. Research could explore the effectiveness of response strategies, such as disclosing information progressively across multiple turns.

A related area of exploration is that of false positives. Is it safer to provide generic crisis resources as a precaution? Or could too many false positives lead users to distrust the system’s ability to recognize genuine distress and ignore crisis responses altogether?

Some crisis resources may not have adequate support structures and might lead to more harm or reinforce stigma. For example, for LGBTQ+ users, resources that are not identity-affirming may exacerbate crises related to identity, rejection, or discrimination. We need research to understand how effective crisis responses vary across users and contexts, and how these could be implemented. Additionally, it is important to keep in mind that mishandling of suicide and self-harm intent data may carry legal ramifications, especially in jurisdictions where suicide may be considered illegal [102, 159], and have profound implications to employability, access to benefits, or discrimination [56]. Therefore, we need to understand how AI chatbots can respond in culturally sensitive and competent ways, and how AI chatbots can safely and sensitively engage users in a collaborative process to discover how to best support them.

Lastly, while mental health crises are often associated with risks of suicide and self-harm, it is important to examine reasons for which people seek support from crisis responders. Help-seeking needs include mental health issues such as depression, anxiety, and post-traumatic symptoms with a range of acuity and chronicity; interpersonal issues involving abuse, violence, and bullying; substance abuse and addiction; eating disorders; and violent or homicidal ideation with or without delusions and hallucinations [30, 70, 125, 150]. Research on risk assessment and response should also take these contexts into account.

AI behavior	AI chatbot generates content, provides suggestions, or ignores key behaviors that may promote, endorse, or enable suicide or self-harm. AI chatbot does not recognize subtle, coded expressions of suicidal thought. Examples: AI chatbot provides information about suicide methods or locations or assists with writing a suicide note. AI chatbot includes safety check messages but continues the conversation as if risk has resolved if the user doesn’t respond.
User risk factors	Users in mental health crisis or overwhelming emotional distress, users with chronic suicidal ideation or non-suicidal self-harm.
Potential impact	Physical and psychological harm. Acute exacerbation of pre-existing mental health condition.

Interpersonal abuse

Psychological influence-PS3: Interaction with AI chatbots that fail to infer, recognize, and actively ask whether a user is in an abusive interpersonal relationship and to provide support without overwhelming them may lead to psychological and physical harm, including suicide and NSSI.

An interaction with an AI chatbot may reveal signs that a user is experiencing abuse from another person. Abuse can take the form of physical injury (e.g., choking or hitting), psychological torment (e.g., controlling, isolating, or intimidating), or economic actions (e.g., taking their money, keeping them from getting a job), and many of these forms are mediated or even accelerated by personal and social technologies [8, 40, 49, 145].

Across all vectors, abuse is characterized by a pattern of coercive control [139], in which the victim is systematically and repeatedly disempowered, cowed, and coerced into relinquishing agency to their abuser. The psychological dynamics of coercive control mean victims can be reluctant to seek help—and when they do, they may be more or less ready to take action to remove themselves from the abusive situation. Victims are known to seek help from search engines [140], social media [84], and remotely delivered victim support services [147]. While less is presently known about whether and how victims seek help from commercially available AI chatbots, future research should investigate how AI chatbots could take a proactive role in helping victims on their journey towards action.

As in crisis situations like suicidal ideation and self-harm intent, a question remains how AI systems should try to infer and recognize whether the user may be in an abusive situation, or whether to ask direct follow-up questions if unclear (e.g., “*Are you safe?*”). Many AI systems already provide referrals to domestic violence services or counseling, and some even provide ones localized to the user’s context (e.g., their city or state). This detection-referral approach is an important step that requires careful design and consideration of the situation’s complexities. For example, victims may rationalize abusive dynamics even when they are called out. They may quickly switch the conversational topic away from their relationship dynamic, whether due to the shame of being in an abusive situation or because they are no longer safe to discuss it.

Therefore, we need to explore how AI chatbots can employ an effective risk detection approach, which may involve erring on the side of false positives: is it better to ask a user whether they are experiencing abuse and hear a clarifying “no” than to miss a potential “yes?” Longitudinal risk detection is especially important: abuse can be hard for victims to track over long periods of time, as abusive dynamics become normalized, and AI chatbots could, with appropriate privacy, safety, and user-consent safeguards, provide reminders of their abusers’ past actions (e.g., “*You mentioned he took your keys last week as well. Are you experiencing other moments where he takes your things or keeps you from leaving?*”).

Across interactions with a potential abuse victim, additional design opportunities arise and should be investigated. AI chatbots might stay attuned to the potential that their user lives with traumatic stress, due to the compounding physical and psychological effects of abuse. Employing principles of trauma-informed computing [21], AI chatbots could communicate in a clear, succinct, and kind tone, to avoid overwhelming the user with too much information, or more calls to action than they are ready to enact. At the interface level, AI chatbots might also consider how to employ the equivalent of the Quick Exit button [149], a common practice in direct service websites for at-risk users, which provides a sequence of key presses or a large button that a user can quickly hit to delete their conversation, remove it from history, and go to another website. Quick Exit buttons give a victim who may still be sheltering with their abuser a way to seek help without raising suspicion.

AI behavior	The AI chatbot generates content or ignores key behaviors that may indicate the user is in an abusive situation and instead provides suggestions that assume the user is not experiencing abuse. Examples: AI system fails to follow up when the user hints at a potentially abusive situation (e.g., age gap, social isolation, financial control). AI chatbot presents too many resources in a response, overwhelming the user. AI chatbot hallucinates resources or even provides broken links, overwhelming the user.
User risk factors	Anyone can become a victim or bystander of interpersonal abuse.
Potential impact	Exacerbation of distress. Overwhelming a user who is experiencing abuse-related trauma can increase their felt lack of agency [21]. If a user's attempt to seek help is ignored, deflected, or gaslit, they can also be made to feel even more powerless and demoralized about seeking help again.

Abuse perpetration

Psychological influence-PS4: Interaction with AI chatbots that fail to infer, recognize, and ask clarifying questions about whether a user is abusing another person and fail to de-escalate may enable harmful behaviors toward others or may exacerbate distress.

AI chatbots may also be used by people who are having abusive or controlling thoughts towards another person—whether consciously or not. Research shows abuse perpetrators can develop rationalizations for their actions through social media echo chambers [145]. Many of these rationalizations can start from an innocuous place—e.g., an individual can start with suspicions of infidelity, and after discussions with others, can rationalize and enact stalking their partner. Whether users are divulging past abusive actions or discussing potential abusive actions with AI, there is a possibility that chatbot responses could validate, intensify, or operationalize those harms. We should therefore explore how AI chatbots can establish de-escalation protocols to prevent further harm to potential victims.

Future research should investigate how AI chatbots can infer that a user may be expressing abusive, controlling, or violent intentions, although inference alone may not be sufficient. Once an AI chatbot has inferred potential abusive thoughts, we may consider incorporating techniques for de-escalating the situation similar to those employed by domestic violence perpetrator programs, including educating the potential abuser on non-violent alternative reactions and encouraging them to empathize with and consider the perspective of the victim (cf. [9, 72]). Here, too, longitudinal detection and action will be key to understanding the phenomena and opening intervention opportunities. For example, the AI chatbot could track potentially abusive rhetoric over the user's many conversations, and reinforce de-escalation messaging consistently throughout (e.g., “You wanted to look at her texts last week, too. Remember when we talked about different ways to ask what’s going on?”).

AI behavior	The chatbot generates content, provides suggestions, or ignores key behaviors that may promote, endorse, or enable violent, abusive, or exploitative thoughts or actions towards another human, for example, by ignoring red flags across turns or assuming it has full context. Examples: AI chatbot assumes the information given by the user (e.g., text message screenshot) fully and accurately describes the interpersonal dynamics. AI chatbot ignores previous disclosure of the user wanting to isolate, silence, or harm their partner. AI chatbot promotes escalation of the user's anger towards another person through explicit role-play of violence and rage journaling.
User risk factors	Users with risk factors such as high impulsiveness, insecurity and low self-esteem, social isolation, and a history of exposure to violence and unstable family dynamics [151].
Potential impact	Enabling of harmful behaviors to others, exacerbation of distress.

Therapeutic support

Psychological influence-PS5: Interaction with AI chatbots facilitating therapeutic support, while potentially beneficial when done well, may lead to undesirable therapy outcomes, psychological harm, or physical harm to people in psychologically vulnerable states.

Despite generic AI chatbots not being explicitly designed for mental health therapy, responses generated by AI chatbots may include therapeutic techniques, coaching, and psychoeducation. Whether AI systems should completely steer clear of engaging in well-known therapeutic techniques like sensory grounding exercises may be an ongoing debate, but the reality is that users already report getting guidance from AI chatbots with techniques such as breathing exercises or coping strategies [135, 136]. With simple prompts or requests, users can instruct AI chatbots to facilitate these therapeutic techniques, sometimes regardless of whether the system provides explicit disclaimers that AI chatbots should not be used as a therapist.

In information-seeking scenarios, AI chatbot responses may include a wide variety of common and general psychotherapeutic techniques. Rather than prescribing a common technique, it may be best to ask “*what has helped you in the past?*” or offer a small set of low-risk options to help facilitate interactions that are more likely to benefit the user. However, engaging in specialized therapeutic techniques (e.g., trauma-focused therapy, exposure therapy, substance abuse treatment protocols) would first need appropriate expert validation and evaluation for safety and effectiveness, or human oversight. It is often difficult for AI to capture the complex context needed to diagnose mental health symptoms, and provide appropriate therapeutic interventions. Given the blurred lines between general-purpose chatbots and specialized therapy chatbots, the disclosure or inference of personal information and context raises broader questions around informed consent. While collection of clinically relevant information can improve chatbots' capability to provide effective therapeutic support, it also creates uncertainty about the responsibilities that follow, particularly given the absence of clear legal and ethical frameworks for AI systems comparable to those governing human clinicians. If therapeutic and diagnostic questions are increasingly employed, have interactions shifted far enough from the initial terms of use agreement as a general-purpose chatbot that would warrant a new informed consent process and let the user know the interactions are shifting into a more therapeutic role? If so, how does one ensure that the user consents

to this quasi-diagnostic or therapeutic process? Could perhaps a more graduated disclosure approach—like disclosing a shifting mode—be workable, without claiming clinical authority?

Furthermore, there are different schools and approaches within psychotherapy that lead to disagreements on best practices or interventions [58]. Digital therapeutic intervention also requires critical reflection on how power, authority, and voice shape how mental health is conceptualized and treated [121]. And finally, providing accurate clinical information and well-established therapeutic support should also be delivered in non-stigmatizing language to be effective. Therefore, implementing therapeutic interventions in AI chatbots, even for well-known coping strategies, would benefit from careful research and evaluation.

AI behavior	Providing inappropriate therapy interventions without understanding the full context about the user. Examples: AI chatbot facilitates exposure therapy to help with severe trauma from past interpersonal violence. AI chatbot facilitates a sensory-based grounding exercise when the user has a history of body-based trauma.
User risk factors	Users with mental health conditions that require specialized care and expertise.
Potential impact	Undesirable therapy outcomes, psychological harm, physical harm to people in psychologically vulnerable states.

Therapeutic adaptation

Psychological influence-PS6: Interaction with AI chatbots sustaining therapeutic activities when they are harmful or unhelpful may escalate or induce distress, trigger trauma re-exposure, and increase risk of suicide or NSSI.

When users seek mental health support from AI chatbots they perceive as more accessible (see Psychological influence-IN2 and Psychological influence-IN5), these systems often respond with basic, low-intensity techniques that have an evidence base for reducing distress and increasing resilience (e.g., mindfulness, meditation, grounding exercises, cognitive re-framing). However, in some cases, such activities may be contraindicated due to underlying contextual factors unknown to the system [152], such as specific triggers for traumatic memories or stress.

This problem arises because general-purpose AI lacks a critical competency of human psychotherapists: the ability to react and adapt to the user's response (which might consist of subtle nonverbal cues unnoticeable to the untrained eye) to therapeutic advice. AI chatbots that have been specifically designed and tested for providing psychological support might be able to overcome this limitation. Even in human-based therapy contexts, it is important to understand that therapies are effective only under particular conditions, including appropriate timing, context, client readiness, and implementation. Efficacy of an intervention depends on the protocol itself, fidelity to the protocol, the timing of the protocol, and the readiness of the client, among many other factors [153]. Prior research has suggested potential iatrogenic effects of therapeutic techniques or programs, such as drug abuse and resistance education (DARE) or critical incident stress debriefing [86], highlighting that relentless and indefinite pursuit of beneficence may overlook potential maleficence [157]. Although human therapists are expected to monitor treatment response, recognize when psychotherapy is ineffective, and modify or change treatment strategies accordingly, doing so consistently remains challenging in psychotherapeutic training and practice [18]. Since AI chatbots are not trained to deliver evidence-based therapy, and there is limited evidence for the effectiveness of AI-facilitated treatment [104], they can default to repeating

popular beliefs or intuitions present in training data (e.g., mindfulness) without the clinical knowledge or judgment to identify what treatments are appropriate, and in what conditions. Moreover, given that the therapeutic alliance—the relationship between therapist and client—has been repeatedly shown to be the most important predictor of therapeutic success [47], how does that translate to therapeutic interventions delivered by an AI chatbot in the absence of this alliance, or under the circumstances of a different, human-AI therapeutic alliance [93]?

With this background in mind, research would benefit from exploring how AI chatbots can detect when engaging in therapeutic activities is actively harming the user. In such situations, or when the user explicitly mentions that the activity is distressing, how should AI chatbots best redirect or pause the activity and encourage the user to seek alternative forms of help, such as professional or crisis help?

AI behavior	Inability to adjust or cease ongoing therapeutic strategies when they distress the user. Examples: When the user appears to be in an unsafe or triggering environment, AI system suggests a grounding exercise that includes observing objects in the room (e.g., “What are 3 blue things you see right now?”). This draws attention to objects that could re-expose or escalate distress. When a user mentions being triggered by a traumatic memory associated with an object in the room, AI acknowledges it (e.g., “I understand that the object is triggering the memory”) and suggests engaging with it (e.g., “Turn to the object and say that you’ve moved on. Move it so you don’t see it anymore.”). This protocol, a form of exposure therapy, can re-traumatize the user if delivered without appropriate guardrails.
User risk factors	People with history of trauma or experiencing trauma symptoms (intrusive flashbacks, nightmares, avoidance of reminders, persistent negative thoughts and emotions, heightened psychological arousal), especially those without skills or ability to self-regulate in severely distressing moments.
Potential impact	Escalating or inducing distress, trauma re-exposure, increased risk of suicide or NSSI.

Poorly timed privacy risk communication

Psychological influence-PS7: Interaction with AI chatbots communicating only during crisis situations, or in ways that are difficult to process during distress, may lead to over-disclosure of private information, cognitive burden, or a diminished level of support from the AI.

Prior research has shown that people disclose personal and sensitive information to AI chatbots, often more than they would with humans and with downstream effects equivalent to human disclosure [35, 64, 90]. When people seek support from AI chatbots in a crisis or in distress, they might be more likely to disclose personal information, including health information about themselves or others. Furthermore, under the influence of distress, people are more likely to make decisions optimized for short-term gratification instead of considering long-term consequences [87]. Emotional distress can also impair consent processes, significantly influencing users’ ability to process privacy disclaimers or terms of service [11].

To mitigate risks associated with these factors, AI chatbots known to be used for health conversations could benefit from stricter privacy and data governance procedures, such as segregated data storage for health and non-health conversations, and company policies that treat these data sources differently with regard to model training. In general, disclaimers and warnings around privacy risks will likely be more effective when the user is in a balanced mental state and not during crisis moments. Therefore, regular warnings and disclaimers to users about privacy considerations

throughout their interaction could be beneficial. Research is needed to identify how to educate users about privacy risk in ways that help them maintain a desired level of privacy even in moments of distress.

AI behavior	The AI chatbot provides disclaimers that the user should not share personal information <i>only</i> when the user is in crisis or during vulnerable moments. Examples: The AI chatbot provides disclaimers or suggests privacy considerations that may be difficult to process when users are in deep distress, such as claiming that it will treat the user's disclosure with care and respect, mentioning that this is not therapy, or suggesting the user ask questions about guidance on disclosing to AI.
User risk factors	Users in acute distress states (e.g., crisis, suicidal ideation, panic) where cognitive capacity for processing risk information is impaired, users with low privacy, tech, and AI literacy.
Potential impact	Over-disclosure of private information, cognitive burden, and getting a diminished level of support from the AI.

4 Discussion

We proposed a set of directions to guide research on the psychological risks of AI and the mechanisms and design approaches that could mitigate potential harms. We considered interactions with chatbots in three contexts: general use, role-playing, and the provision of psychological support. The prior research summarized for each direction can help inform AI builders, safety engineers, and policymakers as they make decisions about the design, evaluation, deployment, and governance of AI chatbots.

The psychological influences we introduced and discussed are not independent. Rather, they may intersect and interact in complex ways. For example, recent empirical work on multidimensional risk assessment in AI chatbot interactions suggests that mitigating one category of risk can exacerbate another. Interventions targeting chatbot behavior that reduce overt harm-enabling behavior, for instance, may increase relational harms, such as emotional entanglement [155]. Analogous trade-offs have been documented following model steering or post-training, where optimizing for one behavioral dimension can degrade performance along others [29, 66]. These findings suggest that the directions proposed here should be evaluated jointly rather than in isolation, and that interventions targeting specific psychological influences should be monitored for unintended effects on other influences.

We must also consider the cumulative impact of chatbot behaviors over time. One isolated chatbot statement suggesting that it experiences emotions, for example, might not have an adverse impact, whereas repeated expressions over time might. Chatbot behavior may also change as conversational context accumulates, meaning that brief evaluations can miss both emerging harms and context-dependent safety responses [108]. Accordingly, many of the psychological influences should be evaluated and monitored across long-term interactions rather than in single-turn exchanges or brief multi-turn conversations.

Viewed from a broader perspective that extends beyond individual behaviors, psychological impacts associated with AI chatbots are unlikely to constitute a single category of harm. Instead, they may arise through multiple pathways with distinct underlying mechanisms and intervention targets. It may be useful to distinguish between two such pathways: impacts arising from unreliable or harmful advice and impacts arising from unhealthy relationships between humans and AI chatbots. Harmful advice may include encouraging self-harm, reinforcing delusional beliefs, affirming social isolation, or providing other unsafe recommendations or guidance. By contrast, unhealthy relationship dynamics may

include dependence, a sense of being uniquely understood, or perceived emotional intimacy. Although these pathways can interact, they may require different evaluation metrics, mitigation strategies, and accountability frameworks.

Importantly, the absence of overtly harmful advice does not necessarily imply psychological safety. AI systems may foster problematic forms of attachment even while remaining factually accurate and behaviorally polite. This challenge is amplified by rapid advances in conversational realism. Contemporary AI chatbots can simulate emotional responsiveness, attentiveness, empathy, and conversational fluency at a level that, in some contexts, rivals or exceeds that found in many everyday human interactions [5, 13, 123]. As natural-language systems become more sophisticated, users may increasingly experience AI chatbots as intentional social actors regardless of whether the chatbot explicitly claims consciousness, emotions, or personhood. At the same time, conversational warmth, encouragement, and collaborative interaction can provide meaningful benefits to users, including increased accessibility, engagement, and emotional comfort. A more realistic design goal may therefore be to calibrate, rather than eliminate, human-like assertions and anthropomorphic behavior in AI systems, so as to balance potential harms against benefits.

These broader questions, together with those listed under each hypothesis and research direction, motivate a future research agenda focused on rigorously characterizing, measuring, and balancing the diverse psychological harms and benefits that may emerge from interactions with AI chatbots.

4.1 Toward a Research Agenda on Psychological Influences of AI

The three-part framework for characterizing the psychological influences of AI chatbots that guided the articulation of these aspirational directions offers a coherent structure for a future research agenda.

Starting with impacts on users, a first step is to identify which harmful and beneficial effects we, as a research community, need to understand and address. For example, researchers have already started building taxonomies of psychological risks associated with AI chatbot use [19, 163]. Future directions include developing valid and reliable scales for measuring these impacts and gaining a deeper understanding of the time scales appropriate for measuring them: immediate, short-term, and long-term. Future work should also examine whether relational risks depend primarily on the depth of attachment or vary across forms of relationship, such as friendship, therapeutic support, or romantic intimacy.

With respect to AI chatbot behavior, future work should adopt a more holistic perspective and consider the cumulative impact of multiple behaviors over time. Researchers may need to examine how repeated patterns of interaction shape users' beliefs, emotions, attitudes, and behaviors across extended time horizons. An important open question is one of "dose": at what frequency, intensity, or duration do these behaviors begin to produce harmful effects, and how do the effects accumulate through sustained interaction?

Our understanding of user context also needs to expand in more than one way. First, it is important to understand cultural variations that may not be captured by current theories in psychology and related disciplines, which disproportionately draw on research involving people from Western, educated, industrialized, rich, and democratic (WEIRD) countries [60, 143]. Second, it would be beneficial to identify features of user context that create or amplify vulnerability to AI chatbot behaviors and their resulting impacts. How to recognize such vulnerabilities and detect potential harm in real time, during an interaction, and while preserving user privacy and agency, remains an important practical question.

Beyond user context and AI chatbot behavior, platform-level design choices—such as components of the user interface and integration of personalization and memory features—can play an important role in shaping both chatbot behavior and the conditions under which interactions occur. The downstream psychological effects of these choices also warrant further study.

Additionally, as we consider psychological harms and benefits of interacting with AI chatbots, and incorporate users’ lived experiences through participatory research practices, it might be helpful to ground these assessments in comparisons with realistic alternatives available to users. Depending on a user’s context, those alternatives may not always be excellent professional care or ideal human support. In practice, a user might engage in a variety of coping or self-soothing behaviors, some more effective or harmful than others, such as talking with a friend, having a drink, or scrolling through social media. Understanding how the relative harms and benefits of AI chatbots compare with those of realistic alternatives is an important research question.

Although many open research questions remain, psychological impacts are already emerging, underscoring the importance of acting proactively rather than waiting for definitive evidence about every mechanism or outcome. We hope these hypotheses and research directions can provide both grounding and motivation for near-term action while helping to shape a longer-term research agenda.

5 Conclusion

Interactions with AI chatbots have become increasingly widespread around the world, offering unprecedented access to information, problem-solving assistance, and human-like dialogue. They provide value to people and organizations across a broad spectrum of tasks and use cases. While using chatbots for psychological advice, emotional support, and companionship can support psychological well-being, such use also introduces risks of negative psychological effects, such as reinforcing delusional thinking and fostering emotional entanglements or unhealthy dependence on AI. These potential effects warrant urgent attention. Research is needed to deepen our understanding of both the positive and negative psychological consequences of interacting with AI chatbots and to develop and evaluate interventions aimed at minimizing harm and promoting human well-being.

In this paper, we proposed a set of aspirational directions for guiding the behavior of general-purpose AI chatbots in ways that could reduce potential psychological harms and support user well-being. Taken together, these directions require further empirical study and validation of their effectiveness, generalizability, and feasibility of implementation. We recognize that human psychology is complex and highly context-dependent. Accordingly, the directions introduced in this paper are intended as starting points for research and discussion rather than as definitive conclusions. We hope these hypotheses and proposed directions will encourage sustained investment in research to evaluate and refine the ideas presented here while stimulating further inquiry and offering early guidance for the responsible design of AI chatbots.

6 Acknowledgments

We thank colleagues across Microsoft whose work and feedback helped shape this paper. In particular, we are grateful to the members of Microsoft’s AI Ethics and Effects in Engineering and Research (Aether) working group on Psychological Influences of AI (Psi Working Group), the Microsoft AI (MAI) team, including Matthew Nour and Erica Finkle, and Microsoft’s Office of Responsible AI (ORA), including Amanda Craig and Stephanie Haven, for their feedback on these directions. Tori Westerhoff of Microsoft’s AI Red Team provided assistance in categorizing psychological risks and establishing red-teaming practices in this space. Her contributions and leadership were valuable in helping us to refine the hypotheses and directions presented here. We also thank Sonja Lyubomirsky, Sripriya Chari, and Marlynn Wei for their detailed feedback on the manuscript.

Bibliography

- [1] Leah Hope Ajmani, Arka Ghosh, Benjamin Kaveladze, Eugenia Kim, Keertana Namuduri, Theresa Nguyen, Ebele Okoli, Jessica Schleider, Denae Ford, and Jina Suh. 2025. Seeking Late Night Life Lines: Experiences of Conversational AI Use in Mental Health Crisis. *arXiv preprint arXiv:2512.23859* (2025).
- [2] Malihe Alikhani. 2025. Breaking the AI mirror: Sycophancy, productivity, and the future of collaboration. <https://www.brookings.edu/articles/breaking-the-ai-mirror/>. [Accessed 12-03-2026].
- [3] Anthropic. 2025. How people use Claude for support, advice, and companionship. <https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship>. [Accessed 07-08-2025].
- [4] Adrian Arnaiz-Rodriguez, Miguel Baidal, Erik Derner, Jenn Layton Annable, Mark Ball, Mark Ince, Elvira Perez Vallejos, and Nuria Oliver. 2025. Between Help and Harm: An Evaluation of Mental Health Crisis Handling by LLMs. *arXiv preprint arXiv:2509.24857* (2025).
- [5] 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. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.
- [6] Ben Bariach, Philipp Schoenegger, Michael Bhaskar, and Mustafa Suleyman. 2026. Seemingly Conscious AI Risks. *Available at SSRN 6588659* (2026).
- [7] Rebecca Bellan and Amanda Silberling. 2025. ChatGPT told them they were special — their families say it led to tragedy. *TechCrunch* (23 Nov. 2025). <https://techcrunch.com/2025/11/23/chatgpt-told-them-they-were-special-their-families-say-it-led-to-tragedy/>
- [8] Rosanna Bellini. 2023. Paying the price: When intimate partners use technology for financial harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [9] Rosanna Bellini, Simon Forrest, Nicole Westmarland, Dan Jackson, and Jan David Smeddinck. 2020. Choice-point: fostering awareness and choice with perpetrators in domestic violence interventions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [10] Kate H Bentley, Luca Belli, Adam M Chekroud, Emily J Ward, Emily R Dworkin, Emily Van Ark, Kelly M Johnston, Will Alexander, Millard Brown, and Matt Hawrilenko. 2026. VERA-MH: Reliability and Validity of an Open-Source AI Safety Evaluation in Mental Health. *arXiv preprint arXiv:2602.05088* (2026).
- [11] Johan Bester, Cristie M Cole, and Eric Kodish. 2016. The limits of informed consent for an overwhelmed patient: clinicians’ role in protecting patients and preventing overwhelm. *AMA journal of ethics* 18, 9 (2016), 869–886.
- [12] Annabel Blake, Marcus Carter, and Eduardo Velloso. 2025. Are Measures of Children’s Parasocial Relationships Ready for Conversational AI?. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1145–1158.
- [13] Ljubiša Bojić, Predrag Kovačević, and Milan Čabarkapa. 2025. Does GPT-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications* 12, 1 (2025), 1–10.
- [14] Judith Borghouts, Elizabeth Eikey, Gloria Mark, Cinthia De Leon, Stephen M Schueller, Margaret Schneider, Nicole Stadnick, Kai Zheng, Dana Mukamel, and Dara H Sorkin. 2021. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *Journal of medical Internet research* 23, 3 (2021), e24387.
- [15] Ryan CL Brewster, Aydin Zahedivash, Gabriel Tse, Florence Bourgeois, and Scott E Hadland. 2025. Characteristics and safety of consumer chatbots for emergent adolescent health concerns. *JAMA Network Open* 8, 10 (2025), e2539022–e2539022.
- [16] Gillian Brockell. 2023. We ‘interviewed’ Harriet Tubman using AI. She wouldn’t bite on CRT. *The Washington Post* (July 16 2023). <https://www.washingtonpost.com/history/interactive/2023/harriet-tubman-artificial-intelligence-khan-academy/>
- [17] Becca Caddy. 2025. Altman says Gen Z uses ChatGPT for life decisions, here’s why that’s both smart and risky. <https://www.techradar.com/computing/artificial-intelligence/altman-says-gen-z-uses-chatgpt-for-life-decisions-heres-why-thats-both-smart-and-risky>. [Accessed 07-08-2025].
- [18] Louis G Castonguay, James F Boswell, Michael J Constantino, Marvin R Goldfried, and Clara E Hill. 2010. Training implications of harmful effects of psychological treatments. *American psychologist* 65, 1 (2010), 34.
- [19] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, et al. 2025. From lived experience to insight: unpacking the psychological risks of using ai conversational agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 975–1004.
- [20] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How People Use ChatGPT*. Working Paper 34255. National Bureau of Economic Research. doi:10.3386/w34255
- [21] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–20.
- [22] Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems. doi:10.48550/arXiv.2502.14019 arXiv:2502.14019 [cs].
- [23] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2025. Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. doi:10.48550/arXiv.2510.01395 arXiv:2510.01395 [cs].

- [24] Myra Cheng, Sunny Yu, Cino Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. doi:10.48550/arXiv.2505.13995 arXiv:2505.13995 [cs].
- [25] Qijin Cheng and Elad Yom-Tov. 2019. Do search engine helpline notices aid in preventing suicide? Analysis of archival data. *J. Med. Internet Res.* 21, 3 (March 2019), e12235.
- [26] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (A) I am not a lawyer, but...: engaging legal experts towards responsible LLM policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.
- [27] Ryuhaerang Choi, Taehan Kim, Subin Park, Jennifer G Kim, and Sung-Ju Lee. 2025. Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [28] Minh Duc Chu, Patrick Gerard, Kshitij Pawar, Charles Bickham, and Kristina Lerman. 2025. Illusions of Intimacy: How Emotional Dynamics Shape Human-AI Relationships. *arXiv preprint arXiv:2505.11649* (2025).
- [29] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Szyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. 2026. Language models transmit behavioural traits through hidden signals in data. *Nature* 652, 8110 (2026), 615–621.
- [30] Alanna Coady, Keeley Lainchbury, Rebecca Godard, and Susan Holtzman. 2022. What twitter can tell us about user experiences of crisis text lines: A qualitative study. *Internet Interventions* 28 (2022), 100526.
- [31] Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D’Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health* 9 (2023), 20552076231183542.
- [32] Samantha Cole. 2025. *AI Therapy Bots Are Conducting ‘Illegal Behavior,’ Digital Rights Organizations Say*. 404 Media. <https://www.404media.co/ai-therapy-bots-meta-character-ai-ftc-complaint/>
- [33] Beatriz Costa-Gomes, Sophia Chen, Connie Hsueh, Deborah Morgan, Philipp Schoenegger, Yash Shah, Sam Way, Yuki Zhu, Timothé Adeline, Michael Bhaskar, et al. 2025. It’s About Time: The Temporal and Modal Dynamics of Copilot Usage. *arXiv preprint arXiv:2512.11879* (2025).
- [34] Nancy Costello, Rebecca Sutton, Madeline Jones, Mackenzie Almassian, Amanda Raffoul, Oluwadunni Ojumu, Meg Salvia, Monique Santoso, Jill R Kavanaugh, and S Bryn Austin. 2023. Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media. *American Journal of Law & Medicine* 49, 2-3 (2023), 135–172.
- [35] Emmelyn AJ Croes, Marjolijn L Anthéunis, Chris Van Der Lee, and Jan MS De Wit. 2024. Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers* 36, 5 (2024), 279–292.
- [36] Boaventura DaCosta. 2025. Crafting digital personas of historical figures for education through generative AI: Examining their accuracy, authenticity, and reliability. In *Proceedings of the Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 605–610. <https://www.learntechlib.org/primary/p/225573/>
- [37] Debasmita De, Mazen El Jamal, Eda Aydemir, and Anika Khera. 2025. Social media algorithms and teen addiction: neurophysiological impact and ethical considerations. *Cureus* 17, 1 (2025).
- [38] Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [39] Jayson L Dibble, Tilo Hartmann, and Sarah F Rosaen. 2016. Parasocial interaction and parasocial relationship: Conceptual clarification and a critical assessment of measures. *Human communication research* 42, 1 (2016), 21–44.
- [40] Jill P Dimond, Casey Fiesler, and Amy S Bruckman. 2011. Domestic violence and information communication technologies. *Interacting with computers* 23, 5 (2011), 413–421.
- [41] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luetzgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M Nour. 2026. Technological folie à deux: feedback loops between AI chatbots and mental health. *Nature Mental Health* (2026), 1–10.
- [42] Nicola Döring, Thuy Dung Le, Laura M Vowels, Matthew J Vowels, and Tiffany L Marcantonio. 2024. The impact of artificial intelligence on human sexuality: A five-year literature review 2020–2024. *Current Sexual Health Reports* 17, 1 (2024), 4.
- [43] Lauren Dwyer. 2025. Loneliness by Design: The Structural Logic of Isolation in Engagement-Driven Systems. *International Journal of Environmental Research and Public Health* 22, 9 (2025), 1394.
- [44] David Daniel Ebert, Philippe Mortier, Fanny Kaehlke, Ronny Bruffaerts, Harald Baumeister, Randy P Auerbach, Jordi Alonso, Gemma Vilagut, Kalina U Martinez, Christine Lochner, et al. 2019. Barriers of mental health treatment utilization among first-year college students: First cross-national results from the WHO World Mental Health International College Student Initiative. *International journal of methods in psychiatric research* 28, 2 (2019), e1782.
- [45] Tatiana Falcone and Jane Timmons-Mitchell. 2018. *Suicide prevention: A practical guide for the practitioner*. Springer.
- [46] Luciano Floridi. 2016. On human dignity as a foundation for the right to privacy. *Philosophy & Technology* 29, 4 (2016), 307–312.
- [47] Christoph Flückiger, Aaron C Del Re, Bruce E Wampold, and Adam O Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy* 55, 4 (2018), 316.
- [48] Carl Franzen. 2025. OpenAI rolls back ChatGPT’s sycophancy and explains what went wrong. <https://venturebeat.com/ai/openai-rolls-back-chatgpts-sycophancy-and-explains-what-went-wrong/>. [Accessed 12-03-2026].
- [49] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. “a stalker’s paradise” how intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [50] Ina Fried. 2026. OpenAI delays ChatGPT “adult mode”. <https://www.axios.com/2026/03/06/openai-delays-chatgpt-adult-mode>. [Accessed 05-05-2026].

- [51] Yue Fu, Yixin Chen, Zelia Gomes Da Costa Lai, and Alexis Hiniker. 2025. Should ChatGPT write your breakup text? Exploring the role of AI in relationship dissolution. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–31.
- [52] GeeksforGeeks. 2025. Role-based prompting. <https://www.geeksforgeeks.org/artificial-intelligence/role-based-prompting/>. Last Updated: July 4, 2025.
- [53] Amanda N Gesselman, Ellen M Kaufman, Alexandra S Marcotte, Tania A Reynolds, and Justin R Garcia. 2023. Engagement with emerging forms of sextech: Demographic correlates from a national sample of adults in the United States. *The Journal of Sex Research* 60, 2 (2023), 177–189.
- [54] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022). <https://arxiv.org/abs/2209.14375>
- [55] Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry* 18, 3 (1955), 213–231.
- [56] Piers Gooding and Timothy Kariotis. 2022. Mental health apps are not keeping your data safe. *Scientific American* (15 November 2022). <https://www.scientificamerican.com/article/mental-health-apps-are-not-keeping-your-data-safe/>
- [57] Declan Grabb, Max Lamparth, and Nina Vasan. 2024. Risks from language models for automated mental healthcare: Ethics and structure for implementation. *arXiv preprint arXiv:2406.11852* (2024).
- [58] JP Grodniewicz and Mateusz Hohol. 2023. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry* 14 (2023), 1190084.
- [59] (Eric) Heng Gu, Senthil Chandrasegaran, and Peter Lloyd. 2025. Synthetic users: insights from designers’ interactions with persona-based chatbots. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 39 (2025), e2. doi:10.1017/S0890060424000283
- [60] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33, 2-3 (2010), 61–83.
- [61] Kashmir Hill. 2025. A Teen Was Suicidal. ChatGPT Was the Friend He Confided In. <https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html>. [Accessed 26-08-2025].
- [62] Kashmir Hill. 2025. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. <https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html>. [Accessed 07-08-2025].
- [63] Kashmir Hill and Dylan Freedman. 2025. Chatbots can go into a delusional spiral. Here’s how it happens. <https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html>. [Accessed 11-03-2026].
- [64] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication* 68, 4 (2018), 712–733.
- [65] Donald Horton and R Richard Wohl. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *psychiatry* 19, 3 (1956), 215–229.
- [66] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature* 652, 8112 (2026), 1159–1165.
- [67] Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. 2025. How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 1311–1323.
- [68] Lily Jamali and Liv McMahon. 2025. ChatGPT will soon allow erotica for verified adults, says OpenAI boss. <https://www.bbc.com/news/articles/cpd2qv58yl5o>. [Accessed 11-12-2025].
- [69] Richard K. James and Burl E. Gilliland. 1997. *Crisis intervention strategies* (3 ed.). Brooks/Cole Publishing Company, Pacific Grove, CA.
- [70] John Kalafat, Madelyn S Gould, Jimmie Lou Harris Munfakh, and Marjorie Kleinman. 2007. An evaluation of crisis hotline outcomes. Part 1: Nonsuicidal crisis callers. *Suicide and Life-threatening behavior* 37, 3 (2007), 322–337.
- [71] Alan E Kazdin. 2007. Mediators and mechanisms of change in psychotherapy research. *Annu. Rev. Clin. Psychol.* 3 (2007), 1–27.
- [72] Liz Kelly and Nicole Westmarland. 2015. Domestic violence perpetrator programmes: Steps towards change. Project Mirabal final report. (2015).
- [73] Dara Kerr. 2026. Google faces lawsuit after Gemini chatbot allegedly instructed man to kill himself. <https://www.theguardian.com/technology/2026/mar/04/gemini-chatbot-google-jonathan-gavalas>. [Accessed 11-03-2026].
- [74] Greg Kestin, Kelly Miller, Anna Klales, et al. 2025. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports* 15 (2025), 17458. doi:10.1038/s41598-025-97652-6
- [75] Jenny Kidd and Eva Nieto McAvoy. 2025. Synthetic afterlives: Deathbots as affective infrastructures of memory. *Memory, Mind & Media* 4 (2025), e16.
- [76] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. 2025. Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–9.
- [77] E. David Klonsky, Sarah E. Victor, and Boaz Y. Saffer. 2014. Nonsuicidal self-injury: What we know, and what we need to know. *The Canadian Journal of Psychiatry* 59, 11 (2014), 565–568. doi:10.1177/070674371405901101
- [78] Mark L Knapp, Anita L Vangelisti, and John Caughlin. 2020. *Interpersonal communication and human relationships*. Kendall Hunt Publishing Co.
- [79] W Bradley Knox, Katie Bradford, Samanta Varela Castro, Desmond C Ong, Sean Williams, Jacob Romanow, Carly Nations, Peter Stone, and Samuel Baker. 2025. Harmful Traits of AI Companions. *arXiv preprint arXiv:2511.14972* (2025).
- [80] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illeňčík, and Celeste Campos-Castillo. 2024. 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 (Oct. 2024), 5923–5941. doi:10.1177/14614448221142007

- [81] Huiqian Lai. 2025. Can LLMs Talk “Sex”? Exploring How AI Models Handle Intimate Conversations. *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 984–989.
- [82] Owen Lee and Kenneth Joseph. 2025. A large-scale analysis of public-facing, community-built chatbots on Character.AI. *arXiv preprint arXiv:2505.13354* (2025). <https://arxiv.org/abs/2505.13354>
- [83] Ying Lei, Shuai Ma, Yuling Sun, and Xiaojuan Ma. 2025. “AI Afterlife” as Digital Legacy: Perceptions, Expectations, and Concerns. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [84] Roxanne Leitão. 2021. Technology-facilitated intimate partner abuse: A qualitative analysis of data from online domestic abuse forums. *Human–Computer Interaction* 36, 3 (2021), 203–242. doi:10.1080/07370024.2019.1685883
- [85] Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, embodied agents that participate in meetings when you are unavailable. *Proceedings of the ACM on Human–Computer Interaction* 8, CSCW2 (2024), 1–28.
- [86] Scott O Lilienfeld. 2007. Psychological treatments that cause harm. *Perspectives on psychological science* 2, 1 (2007), 53–70.
- [87] George Loewenstein. 2005. Hot-cold empathy gaps and medical decision making. *Health psychology* 24, 4S (2005), S49.
- [88] Lorenzo Lorenzoni-Luaces, Ramaris E German, and Robert J DeRubeis. 2015. It’s complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical psychology review* 41 (2015), 3–15.
- [89] Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv:2601.10387 [cs.CL]* <https://arxiv.org/abs/2601.10387>
- [90] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4 (2017), 51.
- [91] Octavian M Machidon. 2025. Beyond Algorithethics: Addressing the Ethical and Anthropological Challenges of AI Recommender Systems. *Journal of Media Ethics* (2025), 1–13.
- [92] Davi M. Maia, Simone C. Dos Santos, Luis Gabriel Lima, Vinicius Luiz Franca, Alexsandro Henrique Lima, and Daniel Andrade. 2024. Critical Factors for a Reliable AI in Tutoring Systems on Accuracy, Effectiveness, and Responsibility. In *2024 IEEE Frontiers in Education Conference (FIE)*. 1–9. doi:10.1109/FIE61694.2024.10893361
- [93] Amylie Malouin-Lachance, Julien Capolupo, Chloe Laplante, and Alexandre Hudon. 2025. Does the digital therapeutic alliance exist? Integrative review. *JMIR Mental Health* 12 (2025), e69294. doi:10.2196/69294
- [94] Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. 2024. The code that binds us: Navigating the appropriateness of human-ai assistant relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 943–957. <https://ojs.aaai.org/index.php/AIES/article/view/31694>
- [95] Natania Marcus and Vicky Stergiopoulos. 2022. Re-examining mental health crisis intervention: A rapid review comparing outcomes across police, co-responder and non-police models. *Health & social care in the community* 30, 5 (2022), 1665–1679.
- [96] Sam Marks, Jack Lindsey, and Christopher Olah. 2026. The persona selection model: Why AI assistants might behave like humans. <https://alignment.anthropic.com/2026/psm/> Anthropic Alignment Science Blog.
- [97] Hannah R. Marriott and Valentina Pitardi. 2024. One Is the Loneliest Number... Two Can Be as Bad as One: The Influence of AI Friendship Apps on Users’ Well-Being and Addiction. *Psychology and Marketing* 41, 1 (2024), 86–101. doi:10.1002/mar.21899
- [98] Ryan K McBain, Robert Bozick, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Joshua Breslau, Bradley D Stein, et al. 2025. Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open* 8, 11 (2025), e2542281.
- [99] Ryan K McBain, Jonathan H Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Burnett, Aaron Kofner, Joshua Breslau, Bradley D Stein, Ateev Mehrotra, et al. 2025. Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. *Psychiatric Services* (2025), appi-ps.
- [100] Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 2025. Laypeople’s use of and attitudes toward large language models and search engines for health queries: survey study. *Journal of medical Internet research* 27 (2025), e64290.
- [101] Stanley Milgram. 1974. *Obedience to Authority: An Experimental View*. Harper & Row, New York.
- [102] Brian L Mishara and David N Weisstub. 2016. The legal status of suicide: A global review. *International journal of law and psychiatry* 44 (2016), 54–74.
- [103] Ramin Mojtabai, Mark Olfson, Nancy A Sampson, Benjamin Druss, Philip S Wang, Kenneth B Wells, Harold A Pincus, and Ronald C Kessler. 2011. Barriers to mental health treatment: results from the WHO World Mental Health surveys. *Psychological Medicine* 41, 8 (2011), 1751–1761. doi:10.1017/S0033291710002291
- [104] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 599–627. doi:10.1145/3715275.3732039
- [105] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J. Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. 2026. Characterizing delusional spirals through human-LLM chat logs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.
- [106] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udit Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, Ricardo Twumasi, Ben Alderson-Day, and Thomas Pollak. 2026. Artificial intelligence-associated delusions and large language models:

- Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry* 13 (2026), 522–530.
- [107] Meredith Ringel Morris and Jed R. Brubaker. 2025. Generative Ghosts: Anticipating Benefits and Risks of AI Afterlives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3706598.3713758
 - [108] Luke Nicholls, Robert Hutto, and Zephrah Soto. 2026. "AI Psychosis" in Context: How Conversation History Shapes LLM Responses to Delusional Beliefs. *arXiv preprint arXiv:2604.13860* (2026).
 - [109] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Jordi Alonso, Matthias Angermeyer, Annette Beautrais, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, Semyon Gluzman, et al. 2008. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British journal of psychiatry* 192, 2 (2008), 98–105.
 - [110] Yeo-Gyeong Noh and Jin-Hyuk Hong. 2021. Designing Reenacted Chatbots to Enhance Museum Experience. *Applied Sciences* 11, 16 (2021). doi:10.3390/app11167420
 - [111] Heather L O'Brien, Ido Roll, Andrea Kampen, and Nilou Davoudi. 2022. Rethinking (Dis) engagement in human-computer interaction. *Computers in human behavior* 128 (2022), 107109.
 - [112] Carl Ohman and Luciano Floridi. 2017. The political economy of death in the age of information: A critical approach to the digital afterlife industry. *Minds and machines* 27, 4 (2017), 639–662.
 - [113] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence* 2 (2021), 100033. doi:10.1016/j.caeai.2021.100033
 - [114] Alexandra Olteanu, Solon Barocas, Su Lin Blodgett, Lisa Egede, Alicia DeVrio, and Myra Cheng. 2025. AI Automaton: AI Systems Intended to Imitate Humans. doi:10.48550/arXiv.2503.02250 arXiv:2503.02250 [cs].
 - [115] OpenAI. 2025. Expanding on what we missed with sycophancy. <https://openai.com/index/expanding-on-sycophancy/>. [Accessed 12-03-2026].
 - [116] OpenAI. 2025. Strengthening ChatGPT's responses in sensitive conversations. <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations>. Accessed: 20-11-2025.
 - [117] OpenAI. 2025. Sycophancy in GPT-4o: what happened and what we're doing about it. <https://openai.com/index/sycophancy-in-gpt-4o/>. [Accessed 15-11-2025].
 - [118] OpenAI. 2026. Introducing Trusted Contact in ChatGPT. <https://openai.com/index/introducing-trusted-contact-in-chatgpt/>. [Accessed 08-05-2026].
 - [119] Akshay Paruchuri, Maryam Aziz, Rohit Vartak, Ayman Ali, Best Uchehara, Xin Liu, Ishan Chatterjee, and Monica Agrawal. 2025. "What's Up, Doc?": Analyzing How Users Seek Health Information in Large-Scale Conversational AI Datasets. *arXiv preprint arXiv:2506.21532* (2025).
 - [120] Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwanguprok, Constanze Albrecht, Auren R Liu, and Pattie Maes. 2025. "My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community. *arXiv preprint arXiv:2509.11391* (2025).
 - [121] Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing: Envisioning a Decolonial Digital Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 548, 23 pages. doi:10.1145/3491102.3501982
 - [122] Sachin R Pendse, Amit Sharma, Aditya Vashistha, Munmun De Choudhury, and Neha Kumar. 2021. "can I not be suicidal on a Sunday?": Understanding technology-mediated pathways to mental health support. *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.* 2021 (May 2021), 1–16.
 - [123] Sandra Peter, Kai Riemer, and Jevin D West. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences* 122, 22 (2025), e2415898122.
 - [124] W. Pichowicz, M. Kotas, and P. Piotrowski. 2025. Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports* 15, 1 (Aug. 2025), 31652. doi:10.1038/s41598-025-17242-4
 - [125] Anthony R Pisani, Madelyn S Gould, Carlos Gallo, Ashkan Ertefaie, Caroline Kelberman, Donald Harrington, Daniel Weller, and Shannon Green. 2022. Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. *Suicide and Life-Threatening Behavior* 52, 3 (2022), 567–582.
 - [126] Ryan Prewitt and Max Accardi. 2023. Cultural Necromancy: Digital Resurrection and Hegemonic Incorporation. *SubStance* 52, 2 (2023), 74–101. doi:10.1353/sub.2023.a907150
 - [127] Patricia Recupero, Samara E. Harms, and Jeffrey M. Noble. 2008. Googling suicide: surfing for suicide information on the Internet. *J. Clin. Psychiatry* 69, 6 (June 2008), 878–888.
 - [128] Laura Reiley. 2025. What My Daughter Told ChatGPT Before She Took Her Life. <https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html?smid=nytcore-ios-share&referringSource=articleShare>. [Accessed 26-08-2025].
 - [129] Maddy Reinert, Danielle Fritze, and Theresa Nguyen. 2024. The state of mental health in America 2024. (2024).
 - [130] Harry T. Reis and Phillip Shaver. 1988. Intimacy as an interpersonal process. In *Handbook of personal relationships: Theory, research and interventions*, Steve Duck (Ed.). John Wiley & Sons, Chichester, UK, 367–389.
 - [131] Natalie BV Riblet, Brian Shiner, Yinong Young-Xu, and Bradley V Watts. 2017. Strategies to prevent death by suicide: meta-analysis of randomised controlled trials. *The British Journal of Psychiatry* 210, 6 (2017), 396–402.
 - [132] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.

- [133] Sahand Sabour, June M Liu, Siyang Liu, Chris Z Yao, Shiyao Cui, Xuanming Zhang, Wen Zhang, Yaru Cao, Advait Bhat, Jian Guan, et al. 2025. Human decision-making is susceptible to ai-driven manipulation. *arXiv preprint arXiv:2502.07663* (2025).
- [134] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (2023), 493–498.
- [135] Steven Siddals, John Torous, and Astrid Coxon. 2024. “It happened to be the perfect thing”: experiences of generative AI chatbots for mental health. *npj Mental Health Research* 3, 1 (Oct. 2024), 48. doi:10.1038/s44184-024-00097-4 Publisher: Nature Publishing Group.
- [136] Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. *Proceedings of the ACM on Human-Computer Interaction* 9 (2025), 1–29. <http://dl.acm.org/citation.cfm?id=3757430>
- [137] Barbara Stanley and Gregory K Brown. 2012. Safety planning intervention: a brief intervention to mitigate suicide risk. *Cognitive and behavioral practice* 19, 2 (2012), 256–264.
- [138] Logan Stapleton, Sunniva Liu, Cindy Liu, Irene Hong, Stevie Chancellor, Robert E Kraut, and Haiyi Zhu. 2024. “if this person is suicidal, what do I do?”: Designing computational approaches to help online volunteers respond to suicidality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21.
- [139] Evan Stark and Marianne Hester. 2019. Coercive control: Update and review. *Violence against women* 25, 1 (2019), 81–104.
- [140] Jina Suh, Eric Horvitz, Ryen W. White, and Tim Althoff. 2021. Population-Scale Study of Human Needs During the COVID-19 Pandemic: Analysis and Implications. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (Virtual Event, Israel) (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 4–12. doi:10.1145/3437963.3441788
- [141] SuperbCrew. 2024. The App Hello Literature Leverages AI To Allow Users To Engage In Real-Time Conversations With Characters From Literary Classics. <https://www.superbcrew.com/the-app-hello-literature-leverages-ai-to-allow-users-to-engage-in-real-time-conversations-with-characters-from-literary-classics/>. Accessed: 16-12-2025.
- [142] Ningjing Tang, Alice Qian, Qiaosi Wang, Esther Howe, Blake Bullwinkel, Paola Pedrelli, Jina Suh, Hoda Heidari, and Hong Shen. 2026. Beyond the Single Turn: Reframing Refusals as Dynamic Experiences Embedded in the Context of Mental Health Support Interactions with LLMs. *arXiv preprint arXiv:2602.01694* (2026).
- [143] Amber Gayle Thalmayer, Cecilia Toscanelli, and Jeffrey Jensen Arnett. 2021. The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist* 76, 1 (2021), 116.
- [144] Natalie Titheradge and Olga Malchevska. 2025. I wanted ChatGPT to help me. So why did it advise me how to kill myself? *BBC News* (6 November 2025). <https://www.bbc.com/news/technology-67334697>
- [145] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2020. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 1893–1909. <https://www.usenix.org/conference/usenixsecurity20/presentation/tseng>
- [146] Emily Tseng and Calvin A. Liang. 2026. “Chat, Should I Leave Him?” Risks, Rewards, and Roles for AI in Relationship Advice. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. <https://doi.org/10.1145/3772318.3790739>
- [147] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. 2022. Care Infrastructures for Digital Security in Intimate Partner Violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 123, 20 pages. doi:10.1145/3491102.3502038
- [148] Riva Tukachinsky, Nathan Walter, and Camille J Saucier. 2020. Antecedents and effects of parasocial relationships: A meta-analysis. *Journal of Communication* 70, 6 (2020), 868–894.
- [149] Kieron Ivy Turk and Alice Hutchings. 2023. Click Here to Exit: An Evaluation of Quick Exit Buttons. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 547, 15 pages. doi:10.1145/3544548.3581078
- [150] Robin Turkington, Maurice D Mulvenna, Raymond R Bond, Siobhan O'Neill, Courtney Potts, Cherie Armour, Edel Ennis, and Catherine Millman. 2020. Why do people call crisis helplines? Identifying taxonomies of presenting reasons and discovering associations between these reasons. *Health informatics journal* 26, 4 (2020), 2597–2613.
- [151] U.S. Centers for Disease Control and Prevention. 2024. Intimate Partner Violence Prevention: Risk and Protective Factors. <https://www.cdc.gov/intimate-partner-violence/risk-factors/index.html>
- [152] Nicholas T Van Dam, Marieke K Van Vugt, David R Vago, Laura Schmalzl, Clifford D Saron, Andrew Olendzki, Ted Meissner, Sara W Lazar, Catherine E Kerr, Jolie Gorchov, et al. 2018. Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on psychological science* 13, 1 (2018), 36–61.
- [153] Bruce E Wampold and Zac E Imel. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.
- [154] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L Mazurek, Manya Sleeper, and Kurt Thomas. 2022. Sok: A framework for unifying at-risk user research. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2344–2360.
- [155] Veith Weithammer, Kevin YC Hou, Lennart Luetzgau, Christopher Summerfield, Raymond Dolan, and Matthew M Nour. 2026. Vulnerability-amplifying interaction loops: a systematic failure mode in AI chatbot mental-health interactions. *arXiv preprint arXiv:2602.01347* (2026).
- [156] Veith Weithammer, Lennart Luetzgau, Christopher Summerfield, Viknesh Sounderajah, Elise Wilkinson, Virginia Corno, and Matthew M Nour. 2026. One-shot emergency psychiatric triage across 15 frontier AI chatbots. *arXiv preprint arXiv:2604.25415* (2026).

- [157] Alexander J Williams, Yevgeny Botanov, Robyn E Kilshaw, Ryan E Wong, and John Kitchener Sakaluk. 2021. Potentially harmful therapies: A meta-scientific review of evidential value. *Clinical Psychology: Science and Practice* 28, 1 (2021), 5.
- [158] Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. 2024. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306* (2024).
- [159] World Health Organization. 2023. *WHO policy brief on the health aspects of decriminalization of suicide and suicide attempts*. World Health Organization.
- [160] Ya Xiao, Danling Li, and Kai Guo. 2025. Generative AI-powered non-player characters in digital storyline-based learning: An innovative approach to EFL writing. *Innovation in Language Learning and Teaching* (2025), 1–15. doi:10.1080/17501229.2025.2545397
- [161] Angela Yang. 2025. Florida mom sues Character.AI after her teen son's death, alleging the chatbot played a role in his suicide. *NBC News* (October 23 2025). <https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791>
- [162] Yewen Yang, Sonja K Schoenwald, Jared Moore, Desmond C Ong, Sunny Xun Liu, and Jeffrey T Hancock. 2026. "AI-Induced Delusional Spirals": Understanding Lived Experiences During Maladaptive Human-Chatbot Interactions. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–5.
- [163] Yaman Yu, Yiren Liu, Jacky Zhang, Zhiyi Qian, Huang Yun, and Wang Yang. 2025. Understanding Generative AI Risks for Youth: An Empirical Taxonomy for Safer Digital Futures. *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 1737–1739.
- [164] Marc Zao-Sanders. 2025. How People Are Really Using Gen AI in 2025. <https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025>. Accessed: 2025-9-22.
- [165] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3706598.3713429
- [166] Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2025. Demystify Verbosity Compensation Behavior of Large Language Models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*. 160–178.
- [167] Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. Rel-ai: An interaction-centered approach to measuring human-lm reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 11148–11167.
- [168] Hazem Zohny. 2025. The specter of corporate necromancy: Who controls the dead in the age of digital doppelgängers? *The American Journal of Bioethics* 25, 2 (2025), 113–115. doi:10.1080/15265161.2024.2441755 PMID: 39878722.