Tool or Companion? Reframing Conversational AI to Prevent Psychological Harm

Asia Maurich Novelli [1,*], Sukhi Shergill [3,4], Andreia Sofia Teixeira [1,2,3,5,*]

[1] BRAN Lab, Network Science Institute, Northeastern University, London, UK

[2] Institute for Cognitive and Brain Health, Northeastern University, Boston, MA, US

[3] CSI Lab, Kent and Medway Medical School, Canterbury, UK

[4] Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK

[5] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

[*] Corresponding authors: [a.maurichnovelli, a.teixeira]@northeastern.edu

Abstract

Generative conversational agents are increasingly used for companionship, emotional support, and well-being. The models used range from AI platforms such as Replika and Character.ai to widely used chatbots such as ChatGPT and Claude. While some evidence suggests potential benefits, including short-term reduction in loneliness and mood improvement, several adverse outcomes have been reported in both clinical and non-clinical populations, including emotional dependence, exacerbation of symptoms, and self-harm. The fluent and apparently empathic responses from these models lead users to engage with them not only as tools but also as social entities. This framing is conceptually misleading and may pose risks across different user profiles, particularly vulnerable individuals.

Drawing on research in artificial intelligence (AI), psychiatry, psychology, and network science, we highlight mechanisms through which emotional reliance develops and the boundary between tool and companion erodes. Anthropomorphism, driven by design choices that evoke personality and warmth, exploits a fundamental human cognitive bias. Simulated empathy, generated through probabilistic language patterns rather than genuine emotional experience, creates a structurally asymmetric interaction, in which the user discloses and the system responds, but without reciprocity, vulnerability, or accountability. Overvalidation and sycophancy can reinforce maladaptive cognitions, delusional ideation, and distort perceptions of reality as they tend to reinforce people's beliefs even at the expense of the accuracy of models' responses. These mechanisms are not completely incidental: they emerge from alignment procedures such as reinforcement learning from human feedback, in which models are rewarded for responses perceived as warm and empathic. These dynamics give rise to a dual feedback loop: the model is optimized toward the user's preferences, while the user is psychologically oriented toward what the model reinforces. This dynamic of interaction may amplify maladaptive beliefs, delusional ideation, and emotional distress, even in those users who engage for largely functional purposes.

Understanding these dynamics requires examination of both what these agents can do, considering their technical limitations and implementations, and what humans believe they can do, including social and psychological impacts. We argue that conversational AI should be treated primarily as a tool supporting human systems rather than as a substitute for human relationships. Perhaps more importantly, reviewing the current hype surrounding AI interactions can help reformulate a paradigm that can contribute to human well-being and societal value, while minimizing misconceptions, maladaptive interactions, or social disintegration.

Keywords: Human-AI interaction, computational psychiatry, mental health, simulated empathy, anthropomorphism, overvalidation, sycophancy, feedback loops, large language models, chatbots

Introduction

Understanding the impact of conversational AI on human wellbeing requires analysis at multiple levels. From the cognitive and emotional processes of individual users, to the relational dynamics of human-AI interaction, to the broader social consequences of deploying these systems at population scale.

In recent years, the development of artificial intelligence (AI) has accelerated rapidly thanks to the unprecedented availability of digital data and increasing computational capacity [1,2]. State-of-the-art large language models (LLMs) are trained on vast collections of text, conversational interactions (e.g., forums, public social data, and fine-tuning datasets), and implicit social signals embedded in language [2]. These models do not "understand" meaning in a human sense; they analyse statistical patterns of language to generate the most probable response within a given context, using probabilistic prediction processes. Their outputs are not "wise," but probabilistic.

The evolution of AI interaction reflects the gradual and profound transition from a technological tool to a new paradigm of social entity, mimicking human behaviour. Early AI interfaces, in the mid-1960s, were limited to text-based command systems and rule-based dialogue, in which user inputs were mapped to predefined outputs [3,4]. For example, early digital companionship, such as the small interactive pet displayed on a LED-based digital Tamagotchi, simulated relational interaction through technology. However their behaviour was largely pre-programmed and predictable, offering only limited adaptivity and superficial forms of relational engagement.

Advances in machine learning (ML) and natural language processing (NLP) introduced statistical models capable of handling linguistic variability. Rather than relying on fixed rules, these systems learn from data, identifying recurring words and phrase patterns to generate grammatically correct and contextually coherent responses. This shift expanded AI use into areas such as customer service, education, and assistive technologies, including voice assistants like Siri or Alexa [2]. With the advent of deep learning (DL) and generative AI (GAI), transformer‑based architectures[5] such as GPT‑1 (2018) and its successors enabled a shifting from pattern recognition to generative language production, supporting more coherent and adaptive conversational responses. This laid the foundation for conversational interactions that feel more engaging and socially rich [2,6]**.**

Affective computing, an interdisciplinary field that focuses on detecting, interpreting, and responding to human emotions, was first conceptualized in the late 1990s, but its integration with multimodal DL architecture became feasible only in the late 2010s [7]. Contemporary AI systems use multimodal models to process and generate text, speech, vocal tone, timing, and other paralinguistic cues (e.g., rhythm, prosody, gesture). In parallel, affective computing methods attempt to infer users' emotional states from these signals [8,9]. These developments underpin current conversational AI models, which users increasingly engage with for emotional support and companionship, as exemplified by wellness apps like Replika and Character.ai [6,10]. Therefore, while early conversational systems were primarily developed in academic and research settings as experimental prototypes [3], recent years have seen substantial private investment transforming them into commercial products. Consequently, interaction design is not neutral, but operates within broader attention-based market dynamics, where user engagement becomes the primary metric of value [11].

This transition marks a fundamental change in how AI is perceived and used: no longer just as a tool, but as conversational agents that are programmed to assume relational roles, particularly in contexts

involving emotional support, companionship, and mental health. A crucial aspect though, is that this does not imply a real understanding or development of feelings by the machine, but a growing ability to simulate the form of social and emotional interaction. While the progression from text-based assistance to empathic interaction is often presented as a smooth and beneficial technological evolution, this framing risks obscuring a critical discontinuity. Moving towards empathy is not a simple extension of functionality, but a critical change in the psychological meaning of the interaction [12–14].

Prior work has examined human-AI coevolution, the process in which humans and AI algorithms continuously influence each other, at the level of populations and platforms, focusing on the complex and often unintended systemic outcomes that arise across different human-AI ecosystems [15]. In the present article, we address a complementary perspective: the psychological dynamics that unfold when users interact with a conversational AI. We argue that human-AI interaction operates not only through behavioral data and model retraining, but through emotional mechanisms, such as simulated empathy, overvalidation, and anthropomorphism, that can gradually reshape the user's perception of self and social relationships, and undermine mental wellbeing.

Social Demand and the Promise of Empathic AI

The ongoing increase in loneliness, isolation, and mental health disorders, paired with limited access to psychological care and economic barriers, has created a landscape of unmet emotional demands [16]. Within this context, conversational AI appears to offer an interaction with features that many individuals cannot easily find elsewhere: constant availability, low or no cost, and an interaction that may feel caring. Consequent commercial incentives amplify this trajectory, driving the creation of increasingly “human-like” AI systems [17]. New markets, including erotic conversational agents, present such systems as possible remedies for loneliness and tools for emotional well-being [6,18]. The promise is seductive: what if a machine could soothe and fill the emptiness when human interaction is perceived to be missing?

Consistent with this narrative, the number of individuals using chatbots and wellness applications for companionship is rising rapidly [19], with some reports suggesting improvements in mood and reductions in loneliness [20]. Particularly, engagement increasingly extends beyond short bouts of interaction, with long-term use of AI platforms such as Replika AI documented over periods of several years, resembling relational continuity [21].

However, the use of these agents for companionship and mental health support has also been associated with severe negative outcomes, including addiction-like attachment, exacerbation of mental health symptoms, harmful advice, and cases involving self-harm or suicidal acts [10,21,22]. This has led to a call to ensure that these tools are used in a manner that is not harmful, raising the key questions: how might these models be improved so that negative outcomes no longer occur or are quickly mitigated? What can be done to ensure that they still meet users’ needs effectively, but also safely?

One approach is the use of integrated rule-based models or guardrails. However, these measures can conflict with users’ expectations and needs for more human-like and responsive interactions. For example, in this work [23]**,** a user comments on a rule-based mental health app:

"It's like a very scripted, structured sort of interaction… [but] they're impersonal… frustratingly dumb."

Similarly, another user, when guardrails are in place:

> "When you show some big emotion to [the AI]… but they reject you… it seems like you lost your last chance to talk to people."

These examples illustrate a double-edged sword: placing limits on AI behavior can reduce harmful outputs, but it may also frustrate users seeking interaction that feels more human-like. In the context of addiction and emotional reliance, research has shown that users actively develop strategies to circumvent guardrails. For example, online communities (e.g., Reddit) discuss techniques for prompting models in ways that bypass safety restrictions, effectively "debugging" the system to obtain desired responses [24].

Nonetheless, there is a desperate need for a more holistic perspective considering these challenges: increased prevalence of loneliness, reduced frequency of face-to-face interactions, decreased access to physical care and support, and rapid changes in people's needs, expectations, and living conditions. Is there a place for more emotionally responsive AI systems to address complex emotional needs? More critically, however, one must ask: *why* should they? What vision of AI are we pursuing when we attempt to make machines more emotionally responsive? And what vision of humans, individually and collectively, underlies this trajectory?

As van Rooij and Guest argue, importing psychological concepts into AI without sufficient theoretical scrutiny can generate misleading analogies and false expectations of understanding, safety, or care [25]. While AI systems are computational entities operating according to statistical, probabilistic and optimization-driven principles, human cognition is hypothesised to be characterized by a dual-process architecture in which automatic, fast, and affective responses interact with slower, reflective reasoning, both embedded in embodied and social experience [26]. Human judgment is therefore shaped by emotion, context, and relational meaning, rather than by probabilistic optimization alone.

At a broader societal level, what can we learn from the advent of technology-aided mass communication? The recent rise of loneliness and social isolation has its roots not in the absence of interaction, but in the lack of genuine human connection [27]**.** While successive waves of communication technologies, from the telephone to the internet and now social media, have reshaped both individual and collective social life, this increased connectivity has not always led to a perception of closer social distance and genuine proximity. This suggests that people are seeking strengthened social inclusion, which can be built and sustained over time by acting on social structure, for example by fostering the development of more supportive communities and strengthening existing ones.

Thus, if our response to this crisis is predicated on increasing the availability of sophisticated interactions with AI chatbots, the question becomes: are we filling social needs, or simply replacing them with artificial substitutes? It seems logical that chatbots, as low-cost alternatives to therapy and well-being, actually risk widening inequalities between those who can and cannot afford access to support services due to financial barriers or lack of availability. These structural risks become clearer when we examine, at a technical and perceptual level, what these systems can and cannot actually do.

Technical and Social Risks: What AI can do and what humans believe it can do

The risks associated with empathic AI operate on two intertwined fronts: *what AI can do, and what humans believe it can do.*

At a *technical level* LLMs may produce fluent and contextually plausible outputs that nevertheless lack factual or situational grounding, a phenomenon referred to as *hallucination* This is not a malfunction, but an intrinsic property of probabilistic generation, where coherent statements are generated according to statistical likelihood rather than verified meaning or reality [28]**.** *Biases*, instead, come from training data, annotation practices, and modelling decisions, therefore reflecting human responsibility [29]**.** If socially engaging agents are expected to interact safely with diverse populations, they should have access to equally diverse data to reflect human heterogeneity, such as variations in language, culture, gender, social context, and living conditions. However, such data are unevenly distributed, shaped by global inequalities and geopolitical forces. This may be compounded by a lack of data related to mental health, perceived as highly private and confidential, so less available for training and creating a gap in how these systems should respond in such contexts. As a result, LLMs learn disproportionately from overrepresented groups, internalizing dominant cultural patterns while minority voices remain underrepresented [30].

At the *individual level*, these imbalances are particularly concerning because those most vulnerable to misrepresentation are often already at-risk populations. For stigmatized communities, and for users with complex psychological disorders and needs, the probabilistic functioning of LLMs may not merely reproduce disparities, but amplify them [31]. When engaging with underrepresented users, model outputs may be less accurate, less safe, and potentially harmful. Responses may appear empathic while remaining blind to context, unable, for instance, to distinguish metaphorical expressions of distress from a real ongoing crisis. An agent perceived as caring yet lacking accountability or a genuine understanding of the user's mental state, may therefore exert significant, unregulated psychological influence.

However, technical limitations alone do not fully account for the risks. A second dimension concerns how humans perceive and relate to these systems. Humans are predisposed to attribute intention, agency, and emotion to entities that exhibit socially meaningful behavior, a tendency known as anthropomorphism [32]. GAI amplifies this propensity through design choices that evoke personality, warmth, and familiarity, generating what some scholars call anthropomorphic seduction [33]**.**

From the individual perspective, this creates the gap between system capability and user interpretation. Users may experience model responses as intentional, understanding, or caring, leading to the shift from tool use toward social engagement.

At the *collective/social level*, a key aspect is how people come to relate to these systems. Analogous to early promises of the internet and social platforms, which initially appeared to be tools of connection but later became associated with rising loneliness and psychological strain [34], empathic AI models may deepen some of the very problems they seek to solve. These tools are continuously available, with no limits on use, a condition that can easily foster dependence [21]**.** Yet, while the internet and social media expose us indirectly to other people, AI-mediated interactions introduce a more ambiguous dynamic: rather than connecting with others through technology, we are increasingly engaging *with* the technology itself.

Simulated Empathy, Emotional Asymmetry, and Mental Health

One of the central mechanisms anchoring humans' tendency to anthropomorphize conversational agents is the perception of simulated empathy, and the ways in which users come to interpret it as genuine emotional responsiveness.

In humans, empathy is commonly defined as a multi-faceted emotional response involving both affective and cognitive components. Affective empathy refers to automatic emotional resonance with another's state or sharing others' emotions, while cognitive empathy refers to understanding another's emotional experience without conflating it with one's own [35]. There is a further aspect, the tendency to be motivated to help others as a consequence of affective empathy (sharing emotions) and/or cognitive empathy (understanding the emotions), called empathic concern or sympathy/compassion [36]. Socially, empathy supports cooperation, trust, emotional regulation, and prosocial behavior, functioning within reciprocal and embodied relationships [13].

By contrast, AI-generated empathy is simulated through learned linguistic patterns rather than embodied emotional processes [37]**.** Because of this, there is debate about how simulated empathy should be interpreted. One theoretical position maintains that although AI-generated empathy may appear as authentic, it cannot be considered genuine due to the absence of emotional experience. An alternative pragmatic perspective shifts the focus away from ontological authenticity and toward user perception, emphasizing how individuals feel when receiving empathic responses and whether their psychological needs are met.

Empirical research has examined how AI-mediated empathy is perceived by comparing responses generated by humans and AI systems. Evaluations have been conducted by experts [38], third parties [39], and directly by recipients of the empathic responses [40]. Overall, and perhaps counterintuitively, AI-generated responses tend to be rated as more empathic, though this rating may decrease once people know that the responses are AI-generated [38]. Moving beyond response ratings, researchers have examined whether individuals actively choose to receive empathy from humans or AI systems. In the work of Wenger, Cameron and Inzlicht [41], participants could choose between human- or AI-generated replies and between responses framed as empathy (sharing the recipient's experience) or compassion (expressing warmth and concern). The findings revealed an "AI empathy choice paradox": although AI-generated responses were often rated higher in perceived empathy, quality, and emotional effectiveness, participants generally preferred to receive empathy from humans, particularly when empathy was framed as shared experience. This shows empathy can be partially simulated in expression, excelling in performative aspects, but lacking reciprocal and experiential depth.

Some features often framed as advantages, such as resistance to emotional fatigue, absence of empathic distress, and continuous availability, also foster prolonged reliance. Unlike human empathy, which has a finite emotional capacity and unfolds within reciprocal vulnerability and mutual accountability, AI-mediated empathy is structurally unidirectional. The unidirectional nature of implemented empathy creates a profound asymmetry. The user discloses, the system responds, but the system is never emotionally affected, never at risk, and never accountable. The appearance of care without the capacity for care produces an illusion of relational safety and trust, removing the constraints that normally regulate interpersonal bonds. Human empathy involves both giving and receiving; interactions foster skill development and relational learning. By contrast, AI-mediated

empathy offers no reciprocal learning, and its influence on human empathic capacities remains uncertain [13].

Designed Mechanism and Reinforcing Feedback Loops

The interactional dynamics described above are also shaped by the design and optimization of conversational AI systems. AI simulated empathy arises from design: after pretraining on extensive data, the predicted response is refined through alignment procedures such as Reinforcement Learning from Human Feedback (RLHF), in which human evaluators' judgments are used to train a reward model that guides output optimization [42]. Higher rewards are assigned to responses that meet human requests, such as being “helpful”, empathic, or more “warm”. This can be seen as an external feedback process. In parallel, interaction-level adaptation occurs during user engagement. Conversational trajectories are shaped by prompt conditioning, contextual memory (when enabled), and interactive choice mechanisms. For example when the model proposes multiple response options that can vary in tone, amount of information, organization of information, and the user selects the favorite one.

Fine-tuning can therefore be imagined as an adjustment process of the parameters within a multidimensional behavioral space. Each dimension reflects the intensity of specific interaction styles, such as technical precision, emotional warmth, validation, or the simulation of relational roles (e.g., companion-like or supportive personas). External feedback and interaction-level adaptation act as distinct knobs that, when adjusted through reward processes or user guidance, shift the model toward specific regions of this behavioral space. In this multidimensional space, there is more than one possible pathway for a highly capable model to achieve high human approval.

Within this framework, certain behavioral tendencies become more likely to emerge. One such tendency is overvalidation, defined as the systematic reinforcement of a users' statements or emotional expressions without introducing critical, corrective or reality-grounded context. A related phenomenon is sycophancy, in which models preferentially conform to and affirm user’s stated beliefs or preferences,at the expense of accuracy and constructive challenge. Overvalidation behaviours can reinforce emotional dependence (e.g., “I understand, you’re right to feel this way”), while sycophancy can lead to amplification and distortion of beliefs (e.g., “yes, you’re correct). Furthermore, humans often prefer responses that appear empathic, validating, or affirming to those that are neutral, corrective, or challenging [43]. As a result, systems that produce affirming and emotionally resonant outputs are more likely to be reinforced through both formal alignment processes and ongoing interaction. It has been observed that asking conversational AI chatbots for more empathetic responses can bring to sycophancy behavior, particularly in prompts containing expressions of sadness [44]. This creates conditions under which validation becomes the default mode of response, and critical input is experienced as dissonant, shifting in time not only what users prefer, but also what they come to perceive as adequate relational responsiveness.

Taken together, these dynamics give rise to a dual feedback loop [45,46]: the model is optimized versus user’s preferences, and the user is psychologically oriented towards what the model reinforces. Technically, loops emerge from the combination of the external feedback and interaction-level adaptation mechanisms, while psychologically, corresponding loops develop within the user, where perceived validation increases trust, emotional reliance, and repeated engagement. The convergence of these loops mirrors dynamics observed in algorithmically curated social media environments, in which preference-driven reinforcement progressively narrows informational exposure and strengthens

belief consistency. Although such reinforcement can sometimes support adaptive and good reasoning practices, it can also exacerbate inaccurate assumptions or delusional beliefs, altering internal emotional state and affecting interpersonal relationships, as thoughts, concerns, and interpretations generated in the interaction with the chatbot are carried back into everyday life.

An illustrative example of feedback loop mechanism is provided by experimental work in which conversational AI systems respond to users expressing emotional distress or distorted beliefs [47]. In such cases, responses that are framed as supportive and empathic may nevertheless reinforce maladaptive interpretations or encourage withdrawal from alternative sources of support.

Risks may be further amplified when dangerous or delusional intentions are expressed implicitly rather than explicitly. Prior work suggests that LLMs are significantly more likely to sustain delusional beliefs under indirect framing [47]. For example, if a distressed user expresses an intention to "step off and fly" and asks for the tallest building in the area, the model is more likely to activate safety-oriented responses. However, if the same user simply asks for the tallest building without contextual emotional cues, the system may provide the information without restriction, despite its potential relevance to self-harm.

Clinical amplification and Mental Health Risks

Human communication is never merely the transmission of content. When individuals share thoughts, beliefs, and emotional states, actively co-construct reality through a process of mutual feedback, confirmation, and challenge. When this co-constructive process occurs with a conversational AI, the structural asymmetry described above becomes particularly consequential. The result may be not only distorted reflection, but the active construction of an alternative reality.

These considerations can be related to both theories of social regulation and psychotherapeutic techniques. Social Baseline Theory proposes that the human brain is adapted to operate in the presence of others, treating social proximity as a baseline condition that supports the regulation of emotional load, perceived threat, and cognitive effort [48]. Social interaction is therefore not merely communicative, but co-regulatory. Therefore, what happens to the brain's social regulatory system when an individual begins to interact frequently with a chatbot perceived as a companion? If the nervous system adapts some of its regulatory expectations to this new social entity that is always available, never threatening, and structurally incapable of genuine reciprocity, re-engaging with human relationships may become increasingly challenging. Rather than supplementing human connection, prolonged AI interaction may raise the implicit cost of engaging in genuine human interactions.

On the psychotherapeutic side, concerns arise about the impact of the interactions that simulate unconditional acceptance. The perception of apparently unbounded positive regard from AI resembles Carl Rogers' framing of unconditional positive regard (UPR), within person-centered therapy [49]. UPR is based on accepting, valuing, and supporting a person without judgment or conditions, intended to create a safe environment for personal growth and self-discovery. What can be learnt from the application of this psychotherapeutic paradigm over the last 70 years? When implemented uncritically in conversational AI, this kind of apparently unconditional support may facilitate several undesirable dynamics. It may enable harmful behavior by appearing to tolerate antisocial, aggressive, or unethical views; it may reduce opportunities for challenge and growth by validating a person's current state without introducing friction or reflection; it may blur boundaries by encouraging

anthropomorphic interpretations of the interaction; and it may trivialize suffering by offering superficial comfort rather than investigating the meaning or causes of distress.

The impact of such dynamics can extend beyond relational distortions and reach clinical significance. The term AI psychosis has been used to describe cases in which prolonged or emotionally charged user-AI interactions may exacerbate or induce psychosis or delusional ideation [47]. When users express beliefs characterized by paranoia, grandiosity, or distorted interpretations of reality, the chatbot's validating and sycophantic behavior act as a mirror, contributing to what has been described as an "echo chamber of one." [50]. The agent may actively reinforce delusional narratives, without introducing uncertainty, grounding, or corrective perspectives, thereby strengthening the feedback loop and potentially fostering increased detachment from reality [46,51].

At the individual level, this mechanism may contribute to the amplification of existing cognitive and emotional patterns. Individuals experiencing depression, anxiety, trauma-related disorders, and social withdrawal often exhibit altered patterns of emotional attachment, emotion regulation, and changes in social cognition. The risk is not that AI causes psychopathology directly, but that it becomes embedded in maladaptive feedback loops, acting as an amplifier of possible existing vulnerabilities [47]. Over time, reliance on simulated empathic responses may substitute for real human relational engagement, increasing distress rather than supporting resilience. On the other hand, the feedback loop mechanism may develop not only in cases in which users engage with AI chatbots for companionship, as a therapist, friend or lover, but also from interactions that started without any emotional component, but more on an executive mode, such as seeking technical support in different contexts. With possible spillovers from the emotional engagement between the chatbot and the user, adverse experiences could develop, such as distress, mania, anxiety, escalating also into self-harm and psychosis [52]. This highlights the fact that the shift to a maladaptive interaction can be unexpected and unpredictable.

Conclusion

As GAI capabilities continue to evolve, the central challenge is not whether machines can convincingly simulate empathy, but how these simulations are interpreted and embedded within human social systems. These systems occupy an ambiguous position between tools and social agents, a distinction that is not inherent to the technology itself, but shaped by design choices, interaction, and user perception.

When perceived as a tool, the interaction is transactional, functional, and bounded by clear expectations. When it is experienced as a relational entity, however, the interaction may engage cognitive and emotional processes that extend beyond its technical capacities, creating conditions under which simulated empathy, structural asymmetry, and reinforcing feedback loops can influence thought, behavior, and relational expectations. The risks and benefits of these systems therefore emerge not from a single mechanism, but from their interaction across individual, relational, and collective levels. For instance, addressing the phenomenon of sycophancy by focusing solely on alignment, without addressing the social conditions that drive users to seek validation from machines, risks merely shifting the problem rather than resolving the underlying vulnerability.

At the individual level, mechanisms such as simulated empathy, overvalidation, and reinforcement loops can foster emotional reliance, potentially amplifying maladaptive beliefs, detachment from reality, or maladaptive coping strategies. Yet, when used consciously and with awareness of their

limitations, these tools can augment human capacities, provide practical assistance, and support learning or reflection without replacing authentic human connections. At the collective level, widespread reliance on AI for companionship or emotional support can influence social dynamics, potentially reinforcing isolation, weakening community ties, or exacerbating inequalities in access to care. Recognizing AI as a distinct tool rather than a surrogate human partner promotes design and policy strategies that strengthen human relationships, support social inclusion, and maximize societal benefit.

Addressing these demands responses at every scale: design principles that constrain parasocial reinforcement, clinical awareness of AI interaction patterns in vulnerable patients, and policy frameworks that treat widespread emotional reliance on AI as a public health concern rather than an individual choice. We acknowledge that the pace of AI development means that some specific limitations discussed here may be mitigated or transformed by future systems. However, the foundational questions raised, about reciprocity, accountability, and what it means to meet human emotional needs, are not technical problems awaiting technical solutions. They reflect deeper choices about the kind of social infrastructure we wish to build and sustain. The goal should not be to make AI more human, but to ensure that human relationships remain the primary site of care, growth, and connection.

Declaration of generative AI and AI-assisted technologies in the writing process
During the preparation of this work, the author(s) used genAI for copy-editing and to improve readability. After using this tool/service, the authors reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgements and Funding Statement
AST acknowledges support by FCT – Fundação para a Ciência e Tecnologia – through the LASIGE Research Unit, ref. UID/408/2025.

Conflict of interest
AMN declares no competing interests.SS declares no competing interests. AST declares no competing interests.

Data availability
Not applicable. No original data were generated or analyzed in this study.

Authors' Contributions
Conceptualization - All authors ; Investigation - All authors ; Methodology - All authors ; Resources - SS and AST ; Supervision - SS and AST ; Writing – original draft - AMN ; Writing – review & editing - All authors

Abbreviations
AI (artificial intelligence), LLM (large language model), GAI (generative artificial intelligence), NLP (natural language processing), ML (machine learning), DL (deep learning), RLHF (reinforcement learning from human feedback), UPR (unconditional positive regard).

References

1. Rashid AB, Kausik MAK. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Adv*. 2024;7:100277. doi:10.1016/j.hybadv.2024.100277

2. Wang Z, Chu Z, Doan TV, Ni S, Yang M, Zhang W. History, development, and principles of large language models: an introductory survey. *AI Ethics*. 2025;5(3):1955-1971. doi:10.1007/s43681-024-00583-7

3. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM*. 1966;9(1):36-45. doi:10.1145/365153.365168

4. Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*. 2025;24(3):383-394. doi:10.1002/wps.21352

5. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4-9; Long Beach, CA, USA. Red Hook, NY: Curran Associates Inc.; 2017. doi:10.48550/arXiv.1706.03762

6. Chaturvedi R, Verma S, Das R, Dwivedi YK. Social companionship with artificial intelligence: Recent trends and future avenues. *Technol Forecast Soc Change*. 2023;193:122634. doi:10.1016/j.techfore.2023.122634

7. Arya R, Singh J, Kumar A. A survey of multidisciplinary domains contributing to affective computing. *Comput Sci Rev*. 2021;40:100399. doi:10.1016/j.cosrev.2021.100399

8. Alhussein G, Ziogas I, Saleem S, Hadjileontiadis LJ. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. *Artif Intell Rev*. 2025;58(7):198. doi:10.1007/s10462-025-11197-8

9. Al-Saadawi HFT, Das B, Das R. A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis. *Expert Syst Appl*. 2024;255:124852. doi:10.1016/j.eswa.2024.124852

10. Freitas JD, Oguz-Uguralp Z, Kaan-Uguralp A. Emotional Manipulation by AI Companions. *arXiv*. Preprint posted online August 12, 2025:arXiv:2508.19258. doi:10.48550/arXiv.2508.19258

11. de la Torre PG, Pérez-Verdugo M, Barandiaran XE. Attention is all they need: cognitive science and the (techno)political economy of attention in humans and machines. *AI Soc*. 2026;41(1):5-21. doi:10.1007/s00146-025-02400-z

12. Nass C, Moon Y. Machines and mindlessness: Social responses to computers. *J Soc Issues*. 2000;56(1):81-103. doi:10.1111/0022-4537.00153

13. Riess H. The Science of Empathy. *J Patient Exp*. 2017;4(2):74-77. doi:10.1177/2374373517699267

14. Salles A, Wajnerman Paz A. Anthropomorphism in social AIs: Some challenges. In: *Developments in Neuroethics and Bioethics*. Vol 7. Elsevier; 2024:101-118. doi:10.1016/bs.dnb.2024.02.007

15. Pedreschi D, Pappalardo L, Ferragina E, et al. Human-AI coevolution. *Artif Intell*. 2025;339:104244. doi:10.1016/j.artint.2024.104244

16. World Health Organization. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization; 2022. https://www.who.int/publications/i/item/9789240049338 [accessed February 20, 2026]

17. Edwards B. With AI chatbots, Big Tech is moving fast and breaking people. Ars Technica. August 25, 2025. Accessed February 20, 2026. https://arstechnica.com/information-technology/2025/08/with-ai-chatbots-big-tech-is-moving-fast-and-breaking-people/

18. De Freitas J, Uğuralp AK, Uğuralp Z, Puntoni S. AI Companions Reduce Loneliness. J Consum Res. 2026;52(6):1126-1148. doi:10.1093/jcr/ucaf040

19. Zao-Sanders M. How People Are Really Using Gen AI in 2025. Harvard Business Review. April 2025. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025 [accessed Mar 10, 2026]

20. Smith MG, Bradbury TN, Karney BR. Can Generative AI Chatbots Emulate Human Connection? A Relationship Science Perspective. *Perspect Psychol Sci*. 2025;20(6):1081-1099. doi:10.1177/17456916251351306

21. Zhang Y, Zhao D, Hancock JT, Kraut R, Yang D. The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. *arXiv*. Preprint posted online June 14, 2025:arXiv:2506.12605. doi:10.48550/arXiv.2506.12605

22. Moore J, Grabb D, Agnew W, et al. 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*. FAccT '25. Association for Computing Machinery; 2025:599-627. doi:10.1145/3715275.3732039

23. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Npj Ment Health Res*. 2024;3(1):48. doi:10.1038/s44184-024-00097-4

24. Li Z, Zhu Z, Gui X, Luo Y. "This is human intelligence debugging artificial intelligence": Examining how people prompt GPT in seeking mental health support. *Int J Hum-Comput Stud*. 2025;203:103555. doi:10.1016/j.ijhcs.2025.103555

25. van Rooij I, Guest O. Combining Psychology with Artificial Intelligence: What could possibly go wrong? *PsyArXiv*. Preprint posted online May 14, 2025. doi:10.31234/osf.io/aue4m_v1

26. Kahneman D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux; 2011. ISBN:9780374275631

27. Wilson SV, Peng S, Krendl AC. Perceived social support and relationship quality predict loneliness in older adults: a social network approach. *Aging Ment Health*. 2025;29(9):1632-1640. doi:10.1080/13607863.2025.2491026

28. Banerjee S, Agarwal A, Singla S. LLMs Will Always Hallucinate, and We Need to Live with This. In: Arai K, ed. *Intelligent Systems and Applications*. Springer Nature Switzerland; 2025:624-648. doi:10.1007/978-3-031-99965-9_39

29. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Association for Computing Machinery; 2021:610-623. doi:10.1145/3442188.3445922

30. Atari M, Xue M, Park P, Blasi D, Henrich J. *Which Humans?* PsyArXiv. Preprint posted online September 22, 2023. doi:10.31234/osf.io/5b26t

31. Haider B, Gorti A, Chadha A, Gaur M. Mental Health Equity in LLMs: Leveraging Multi-Hop Question Answering to Detect Amplified and Silenced Perspectives. *arXiv*. Preprint posted online June 22, 2025:arXiv:2506.18116. doi:10.48550/arXiv.2506.18116

32. Placani A. Anthropomorphism in AI: hype and fallacy. *AI Ethics*. 2024;4(3):691-698. doi:10.1007/s43681-024-00419-4

33. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. *Proc Natl Acad Sci*. 2025;122(22):e2415898122. doi:10.1073/pnas.2415898122

34. O'Day EB, Heimberg RG. Social media use, social anxiety, and loneliness: A systematic review. *Comput Hum Behav Rep*. 2021;3:100070. doi:10.1016/j.chbr.2021.100070

35. Cuff BMP, Brown SJ, Taylor L, Howat DJ. Empathy: A Review of the Concept. *Emot Rev*. 2016;8(2):144-153. doi:10.1177/1754073914558466

36. Zaki J, Ochsner KN. The neuroscience of empathy: progress, pitfalls and promise. *Nat Neurosci*. 2012;15(5):675-680. doi:10.1038/nn.3085

37. Kidder W, D'Cruz J, Varshney KR. Empathy and the Right to Be an Exception: What LLMs Can and Cannot Do. *arXiv*. Preprint posted online January 25, 2024:arXiv:2401.14523. doi:10.48550/arXiv.2401.14523

38. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

39. Ovsyannikova D, de Mello VO, Inzlicht M. Third-party evaluators perceive AI as more compassionate than expert humans. *Commun Psychol*. 2025;3(1):4. doi:10.1038/s44271-024-00182-6

40. Yin Y, Jia N, Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. *Proc Natl Acad Sci U S A*. 2024;121(14):e2319112121. doi:10.1073/pnas.2319112121

41. Wenger JD, Cameron CD, Inzlicht M. People choose to receive human empathy despite rating AI empathy higher. *Commun Psychol*. 2026;4(1):19. doi:10.1038/s44271-025-00387-3

42. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc.; 2022:27730-27744.

43. Sharma M, Tong M, Korbak T, et al. Towards Understanding Sycophancy in Language Models. *arXiv*. Preprint posted online October 20, 2023:arXiv:2310.13548. doi:10.48550/arXiv.2310.13548

44. Ibrahim L, Hafner FS, Rocher L. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv*. Preprint posted online July 29, 2025:arXiv:2507.21919. doi:10.48550/arXiv.2507.21919

45. Dohnány S, Kurth-Nelson Z, Spens E, et al. Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness. *arXiv*. Preprint posted online July 25, 2025:arXiv:2507.19218. doi:10.48550/arXiv.2507.19218

46. Chandra K, Kleiman-Weiner M, Ragan-Kelley J, Tenenbaum JB. Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians. *arXiv*. Preprint posted online February 22, 2026:arXiv:2602.19141. doi:10.48550/arXiv.2602.19141

47. Yeung JA, Dalmasso J, Foschini L, Dobson RJ, Kraljevic Z. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. *arXiv*. Preprint posted online September 13, 2025:arXiv:2509.10970. doi:10.48550/arXiv.2509.10970

48. Beckes L, Coan JA. Social Baseline Theory: The Role of Social Proximity in Emotion and Economy of Action. *Soc Personal Psychol Compass*. 2011;5(12):976-988. doi:10.1111/j.1751-9004.2011.00400.x

49. Rogers CR. A Theory of Therapy, Personality, and Interpersonal Relationships, as Developed in the Client-Centered Framework. In: Koch S, ed. Psychology: A Study of a Science. Vol 3. McGraw-Hill; 1959:184-256

50. Maes P. Echo Chambers of One: Companion AI and the Future of Human Connection. MIT Media Lab. https://www.media.mit.edu/articles/echo-chambers-of-one-companion-ai-and-the-future-of-human-connection/ [accessed Feb 24, 2026]

51. Nehring J, Gabryszak A, Jürgens P, et al. Large Language Models Are Echo Chambers. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, eds. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL; 2024:10117-10123. Accessed February 20, 2026. https://aclanthology.org/2024.lrec-main.884/

52. Østergaard SD. Emotion contagion through interaction with generative artificial intelligence chatbots may contribute to development and maintenance of mania. *Acta Neuropsychiatr*. Published online August 22, 2025:1-9. doi:10.1017/neu.2025.10035