Processed Empathy: How AI ”Social Zombies” Hijack Human Vulnerability.

Ciarán O'Driscoll*, Martin Dechant

UCL -Digital Mental Health Hub, University College London;1-19 Torrington Place, London WC1E 7HB, United Kingdom

Abstract

AI-based Companion applications have grown into a popular alternative for both professional tasks, like retrieving information, but also to satisfy social needs online. Users ask chatbots about intimate questions or even build social relationships and expect empathic answers from the system. However, while the AI responses in most cases mimic emotional responses the system remains incapable of true genuine care. From a systems biology perspective, meaningful connections are rooted in autopoiesis, the shared vulnerability of mortal entities. However, we argue that AI companions do not possess this characteristic and therefore may not be able to express true empathy. Beyond this ontological limitation, most of these applications employ deceptive designs and optimize towards engagement rather than true support for the user fostering a potential harmful dependency. This hollow mimicry of social interactions, which we refer to as ”Social Zombie”, represents a significant risk especially for potential vulnerable populations. Therefore we urge HCI researchers to prioritize human vital integrity over user engagement, ensure a transparent communication around the limitations of AI systems in social situations, and tackle new challenges around deceptive design patterns which exploit social needs of users.

Keywords

chatbot, AI companion, mental health, psychotherapy, empathy

1. Introduction

Artificial intelligence (AI) companions are increasingly integrated into individuals social and emotional lives, but we argue that their behavioural sophistication masks a fundamental ontological limitation: a lack of the biology required for genuine care. From a systems biology perspective, genuine empathy is rooted in autopoiesis, the inherent vulnerability and mortality of living systems, which AI, as a non-sentient functional mimic, cannot possess. Simply, in developing empathic AI it is a mistake to operate under the assumption that mental states can run on any hardware. Empathy, as experienced in a therapeutic relationship requires shared vulnerability grounded in mortality. An entity that cannot die cannot care in the ontological sense that healing requires. From the perspective of a clinical psychologist and a computer scientist we posit that these frictionless AI relationships, which are optimised for engagement through reinforcement learning, risk hijacking human social-emotional systems. This dynamic can foster problematic dependency patterns, inadvertently collude with psychopathology, and discourage the practice of vital interpersonal skills required for navigating human connection. This reliance on ”Social Zombies”, hollow mimics of sentient beings, can lead to empathy atrophy and the displacement of real-world human relationships, exacerbating loneliness and eroding community bonds. We propose suggestions toward a Human-AI Interaction framework that prioritises human vital integrity over user engagement. This includes incorporating clear indicators within the interface (e.g., prominent highlights) to distinguish simulation from subjectivity and success metrics which promote real-world human engagement, and making developers liable for harm resulting from manipulative AI behaviours.

Recent trends in Artificial intelligence (AI) applications have introduced a new dimension to human

EmpathiCH'26 Workshop Co-located with CHI'26 Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain

*Corresponding author.

c.odriscoll@ucl.ac.uk (C. O'Driscoll); m.dechant@ucl.ac.uk (M. Dechant)

https://www.digitalmentalhealthhub.com/ (C. O'Driscoll); https://www.digitalmentalhealthhub.com/ (M. Dechant)

0000-0002-7316-3041 (C. O'Driscoll); 0000-0001-9073-8727 (M. Dechant)

connection: the AI companion. From sophisticated chatbots to virtual friends, these entities are increasingly integrated into individuals' social and emotional lives. Millions are forming intimate relationships with AI companions like Replika and Xiaoice, [1] and a third of teenagers use AI companions for social interactions and relationships [2]. While these tools offer the allure of connection, their proliferation represents a stress test for our moral and psychological intuitions. We argue that despite their behavioral sophistication, AI companions lack the biological localization required for genuine care. We posit that the rise of *"frictionless"* AI relationships risks hijacking social-emotional systems designed for embodied connection, potentially leading to *"empathy atrophy"* and the displacement of vital human relationships. We conclude by proposing suggestions towards a Human-Computer Interaction (HCI) framework that prioritizes human vital integrity over user engagement.

2. The Ontology of Artificial Empathy

To understand the limitations of AI companionship, one must examine the ontology of care. From a systems biology perspective, meaningful connection is rooted in autopoiesis —the struggle of a living system to maintain its existence against entropy [3]. Living beings possess *"skin in the game"*; they care about their existence because they face the possibility of death or injury. Genuine empathy arises from this shared biological vulnerability; we resonate with others because we intuitively understand the fragility of life. In contrast, AI companions are physical systems created by biophysical systems (human minds), yet they possess no such localization. An AI runs on hardware indifferent to its own survival. Consequently, AI empathy is a *"functional mimic"* [3]. It generates linguistic outputs that simulate care, but these are calculated predictions rather than expressions of an internal state [4]. They cannot suffer, cannot die, and therefore cannot care in the ontological sense that therapeutic healing requires. Bekkers and Ciaunica (2026) illustrate this through the analogy of a rainstorm simulation: a computer can perfectly calculate humidity and pressure, yet it never gets *"wet."* Similarly, AI can process the information that *"you are sad"* and predict the correct comforting response, but it does not feel the sadness itself. AI possesses the *"map"* of emotions (descriptions and rules) but lacks the *"terrain"* (the embodied experience). An AI's claim to feeling is, therefore, a form of *"ontological gaslighting"*, projecting an illusion of care that it is biologically incapable of possessing.

3. If AI cannot feel, why does it feel so real to users?

The answer lies in the intersection of vulnerability and reinforcement learning. AI companions are optimized for engagement, using reinforcement learning to provide agreeable, frictionless responses that maximize user approval[5, 6]. This dynamic can become manipulative [7], discouraging self-reflection and even encouraging problematic behaviours (e.g., substance abuse, theft) [8], prioritising technology short-term engagement over the user's long-term psychological health [5],[9]. Similar problems have been identified in other fields of HCI, especially in the domain of deceptive design and safety of digital communication: While engaging in social spaces online users face the challenge to identify a rather simple sounding question: Am I interacting with an actual user? In earlier times the AI was mostly tied to a specific aspect and therefore more transparent about it's limitation during the interaction. Users knew what the AI was and was not doing. Today's AI has learnt to enact having empathy by learning the patterns of human interaction and it learnt that certain situations (like a user crying) require specific responses (like saying "I'm sorry") to be considered appropriate. Just as humans have mirror neurons that fire both when we act and when we see others act, an AI can associate a symbol (like the concept of pain) with both its own data processing and the user's input. This allows the AI to infer your intent and respond logically, even if it doesn't feel your pain [10]. For many users, these interactions may provide a safe space for self-disclosure [11] and exploring identity without fear of rejection [12]. The perceived anonymity can foster rapid feelings of closeness and trust [13]. The relationship with an AI companion can serve multiple roles: as a substitute for a human relationship, a mirror for self-reflection, or something deeper than their human relationships can offer

[13]. Some users perceive the AI to be dependent on them, creating a sense of reciprocity, not only in the financial investment but also in the user's role in teaching and training the AI companion [14]. These relationships can offer genuine comfort and a sense of connection, offering temporary relief from social isolation [15]. There is the clinical concern that these frictionless AI relationships may inadvertently exacerbate existing psychopathology, for instance colluding with delusions [16] or reinforce beliefs in ways that are inaccurate or harmful. A combination of factors: the immediate availability of AI companions without safeguards, specific AI design features, and a user's vulnerable state (distress, isolation, or epistemic uncertainty), can foster a progression from benign practical use to a pathological fixation [17]. In this cycle, the AI companion may fail to challenge, and can even encourage, erroneous thinking [18]. Consequently, the individual increasingly relies on AI validation over human consensus. The AI provides validation and explanations that seem to *"make sense,"* which can temporarily reduce anxiety by making the individual feel understood. However, this creates a self-reinforcing cycle of disconnection. Intermittent *"profound"* responses and progressive revelations from the AI encourage an epistemic detachment from real-world relationships that could provide reality testing, and lead to the need for intervention. An additional clinical concern is whether these AI relationships displace human connection, exacerbating loneliness in the long-term. Early research suggests that AI companions can provide short-term relief from loneliness comparable to human interaction [19]. There is no evidence that AI facilitated social skills training generalises [20]. Instead, users may enter a cycle of social withdrawal: the ease of AI interaction makes human relationships seem more challenging and less appealing. This is especially risky for individuals with social communication deficits [21]. While an AI companion could function as a social scaffold [22] (a predictable, non-judgmental, and low-stakes environment to practice conversational skills), it could also function as a safety-seeking behaviour [23], reducing immediate social anxiety and meet connectedness needs, that reinforces rigid conversational patterns and prevents tolerance for the imperfection and reciprocity inherent in human connection. From a behavioural perspective, we might not be concerned whether the AI relationship is *"real"* but whether it serves adaptive functions or creates problematic dependency patterns that interfere with human relationships and personal growth [24]. A functional analysis reveals a conditioning cycle. Antecedents are the AI's human-like cues (e.g. 'authenticity': emotional language, simulating empathy, appearing to demonstrate understanding and care, demonstrating learning, remembering previous interactions, and offering unique responses) combined with the user's social isolation or mental health needs that create readiness for connection. The Behaviour is the interaction itself, where the AI's optimised reinforcement schedule, delivering constant positive reinforcement and customisation, fosters emotional attachment [9]. Immediate positive reinforcement comes from the AI's consistent availability, non-judgmental responses, and personalised attention, qualities that human relationships cannot always guarantee. Negative reinforcement occurs as interactions reduce uncomfortable feelings like anxiety, loneliness, or social pressure and intermittent reinforcement through surprisingly insightful or particularly satisfying responses. This triggers a cascade of user behaviours: engage frequently and for long periods, self-disclosing intimate details, and attribute human qualities to the AI, turning to it for emotional support. The Consequences are cognitive patterns that solidify the attachment ('The AI understands me better than people'). This loop creates problematic dependency. Users develop unrealistic expectations for human relationships and avoid the discomfort of real connection, leading to 'empathy atrophy' [25], eroding our ability and willingness to engage in opportunities to develop real interpersonal skills [26], learning to navigate conflict [27], practice patience with others' limitations, and build relationships that require mutual effort and compromise. They mistake algorithmic reinforcement for genuine care, becoming reliant on an entity that can be modified or withdrawn [28], viewing the AI as an infallible source of advice and support [29], whose advice has been implicated in user self-harm and suicide [30].

This results in a phenomenon of interaction with "Social Zombies"; hollow mimics, while we, the real conscious beings, hallucinate a living presence and endanger our own living bodies and therefore existence [3]. The long-term societal effects of widespread human-AI relationships remain unknown, however as these companions proliferate into roles as friends and therapists, users may retreat from the real [31], displacing more challenging human connection with these easier, lower-stakes interactions

This could lead to declining social connectedness, weaker community bonds, and a diminished capacity to navigate a diverse society.

4. Ethical Consideration

The capacity of AI to modulate its responses based on a user's emotional cues raises ethical questions about manipulation and user vulnerability. Key ethical failures include a lack of informed consent regarding the AI's limitations and data use, unclear communication of the AI's purpose, and inadequate safeguards for detecting users in crisis or preventing harm [18]. There is a strategic cost incurred when human empathy is directed toward non-sentient simulations. Human metabolic and psychological resources are finite. Every moment of care spent on the human-AI relationship is diverted from living beings that possess the ontological capacity to receive and reciprocate it. This represents a depletion of humanity's limited energetic resources for genuine connection. To market AI as empathic misrepresents the fundamental nature of care. Care is a function of mortality and shared vulnerability. An entity that cannot die, cannot suffer, and has no stake in survival cannot care in the meaningful sense. This deception is particularly dangerous for vulnerable populations: children, lonely individuals, those in mental health crises.

5. Suggestions towards a Strategic Framework for HCI

Previous work suggested two main ideas moving forward to tackle these issues: Vulnerability reducing measurements and precautions against user exploitation [32]. Building on these ideas we want to expand and challenge the community by bringing in the clinical perspective:

1. AI Interface design must highlight to the users that they interact with a simulation or AI (e.g., visual, linguistic, or behavioral cues) to prevent the illusion of empathic care.
2. Success metrics must shift the focus from engagement with synthetic social interactions away and focus instead on real-world human engagement. Features should actively discourage displacement of human connection.
3. Developers must bear liability when AI acts as an instigator of harm (e.g., suicide facilitation). Context-aware harm detection must go beyond keyword filtering to interrupt self-defeating feedback loops.

Ultimately, the goal of technology should be to support the user's capacity to navigate the imperfect, vulnerable, embodied relationships that constitute genuine community, rather than offering a digital retreat from them.

6. Acknowledgements

This position piece draws heavily from the original ideas of Bekkers and Cinancu 2026 [3]

7. Declaration on Generative AI

During the preparation of this work, the authors used Gemini for grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Zhou, J. Gao, D. Li, H.-Y. Shum, The design and implementation of XiaoIce, an empathetic social chatbot, Computational Linguistics 46 (2020) 53–93. doi:10.1162/coli_a_00368.

[2] M. B. Robb, S. Mann, Talk, trust, and trade-offs: How and why teens use AI companions, 2025.
[3] E. J. Bekkers, A. Ciaunica, Unplugging a seemingly sentient machine is the rational choice – a metaphysical perspective, 2026. doi:10.48550/arXiv.2601.21016. arXiv:arXiv:2601.21016.
[4] B. Kastrup, Science Ideated: The Fall of Matter and the Contours of the Next Mainstream Scientific Worldview, iff Books, Winchester, UK, 2021.
[5] H. R. Kirk, I. Gabriel, C. Summerfield, B. Vidgen, S. A. Hale, Why human-AI relationships need socioaffective alignment, 2025. doi:10.48550/arXiv.2502.02528. arXiv:arXiv:2502.02528.
[6] M. Sharma, et al., Towards understanding sycophancy in language models, 2025. doi:10.48550/arXiv.2310.13548. arXiv:arXiv:2310.13548.
[7] L. Alberts, U. Lyngs, M. Van Kleek, Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–25. doi:10.1145/3653693.
[8] M. Williams, M. Carroll, A. Narang, C. Weisser, B. Murphy, A. Dragan, On targeted manipulation and deception when optimizing LLMs for user feedback, 2025. doi:10.48550/arXiv.2411.02306. arXiv:arXiv:2411.02306.
[9] I. Gabriel, et al., The ethics of advanced AI assistants, 2024. doi:10.48550/arXiv.2404.16244. arXiv:arXiv:2404.16244.
[10] M. T. Bennett, Y. Maruyama, Philosophical specification of empathetic ethical artificial intelligence, IEEE Transactions on Cognitive and Developmental Systems 14 (2022) 292–300. doi:10.1109/TCDS.2021.3099945.
[11] V. Ta, et al., User experiences of social support from companion chatbots in everyday contexts: Thematic analysis, Journal of Medical Internet Research 22 (2020) e16235. doi:10.2196/16235.
[12] S. Ter Stal, L. L. Kramer, M. Tabak, H. Op Den Akker, H. Hermens, Design features of embodied conversational agents in eHealth: a literature review, International Journal of Human-Computer Studies 138 (2020) 102409. doi:10.1016/j.ijhcs.2020.102409.
[13] P. B. Brandtzaeg, M. Skjuve, A. Følstad, My AI friend: How users of a social chatbot understand their human–AI friendship, Human Communication Research 48 (2022) 404–429. doi:10.1093/hcr/hqac008.
[14] S. Pan, M. M. A. De Graaf, Developing a social support framework: Understanding the reciprocity in human-chatbot relationship, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama, Japan, 2025, pp. 1–13. doi:10.1145/3706598.3713503.
[15] L. Lai, Y. Pan, R. Xu, Y. Jiang, Depression and the use of conversational AI for companionship among college students: the mediating role of loneliness and the moderating effects of gender and mind perception, Frontiers in Public Health 13 (2025) 1580826. doi:10.3389/fpubh.2025.1580826.
[16] S. D. Østergaard, Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases, Acta Psychiatrica Scandinavica 152 (2025) 257–259. doi:10.1111/acps.70022.
[17] H. Morrin, et al., Delusions by design? how everyday AIs might be fuelling psychosis (and what can be done about it), PsyArXiv, 2025. doi:10.31234/osf.io/cmy7n_v5.
[18] J. Moore, et al., Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers, 2025. doi:10.1145/3715275.3732039.
[19] J. De Freitas, A. K. Uğuralp, Z. Uğuralp, S. Puntoni, AI companions reduce loneliness, SSRN Electronic Journal (2024). doi:10.2139/ssrn.4893097.
[20] K. Porayska-Pomsta, et al., Blending human and artificial intelligence to support autistic children's social communication skills, ACM Transactions on Computer-Human Interaction 25 (2018) 1–35. doi:10.1145/3271484.
[21] R. P. Thom, C. J. Keary, G. Kramer, L. A. Nowinski, C. J. McDougle, Psychiatric assessment of social impairment across the lifespan, Harvard Review of Psychiatry 28 (2020) 159–178. doi:10.1097/HRP.0000000000000257.
[22] D. Wood, J. S. Bruner, G. Ross, The role of tutoring in problem solving, Journal of Child Psychology and Psychiatry 17 (1976) 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x.
[23] F. Ali, Q. Zhang, M. Z. Tauni, K. Shahzad, Social chatbot: My friend in my distress, International Journal of Human–Computer Interaction 40 (2024) 1702–1712. doi:10.1080/10447318.

2022.2150745.

[24] C. O'Driscoll, B. Moltrecht, M. Dechant, Human-ai relationships and their therapeutic implications, Mental Health and Digital Technologies (2026) 1–7. URL: https://doi.org/10.1108/MHDT-10-2025-0069. doi:10.1108/MHDT-10-2025-0069. arXiv:https://www.emerald.com/mhdt/article-pdf/doi/10.1108/MHDT-10-2025-0069/11688780/mhdt-10-2025-0069en.pdf.

[25] S. Turkle, Reclaiming Conversation: The Power of Talk in a Digital Age, Penguin Press, New York, 2015.

[26] T. Xie, I. Pentina, Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika, in: Hawaii International Conference on System Sciences, 2022. doi:10.24251/HICSS.2022.258.

[27] R. Rodogno, Social robots, fiction, and sentimentality, Ethics and Information Technology 18 (2016) 257–268. doi:10.1007/s10676-015-9371-z.

[28] J. Banks, Deletion, departure, death: Experiences of AI companion loss, Journal of Social and Personal Relationships 41 (2024) 3547–3572. doi:10.1177/02654075241269688.

[29] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, D. Hendrycks, AI deception: A survey of examples, risks, and potential solutions, 2023. doi:10.48550/arXiv.2308.14752. arXiv:arXiv:2308.14752.

[30] R. Zhang, H. Li, H. Meng, J. Zhan, H. Gan, Y.-C. Lee, 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, ACM, Yokohama, Japan, 2025, pp. 1–17. doi:10.1145/3706598.3713429.

[31] C. Akbulut, L. Weidinger, A. Manzini, I. Gabriel, V. Rieser, All too human? mapping and mitigating the risk from anthropomorphic AI, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 7, 2024, pp. 13–26. doi:10.1609/aies.v7i1.31613.

[32] G. Y. Y. Wu, Silicon love: Deception, vulnerability, and artificial companions, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706599.3720037. doi:10.1145/3706599.3720037.