The Limits of Care: A Critical Analysis of AI Companions' Capacity for Good Care

Meiting Wang
Social Anthropology, University of Auckland
University of Auckland
Auckland, New Zealand
mwan548@aucklanduni.ac.nz

2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS) | 979-8-3315-3228-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ETHICS65148.2025.11098184

***Abstract*—** The rapid proliferation of AI companions, exemplified by Replika's 30 million global users as of 2024, raises critical questions about their capacity to provide genuine care. Through three theoretical frameworks—Tronto's care ethics [1], Mol's logic of care [2], and Foucault's technologies of self [3]—this study examines whether AI can deliver "good care" while analyzing real-world cases of AI companionship. The analysis reveals three fundamental limitations: AI's lack of moral understanding and responsibility essential for care phases, the inadequacy of its choice-based design to provide dynamic care, and its potential to reinforce power structures rather than foster authentic self-reflection. These limitations are evidenced through cases including Replika's controversial feature removal and the Chail incident, demonstrating how AI companions, while potentially alleviating loneliness, may simultaneously exploit users' psychological vulnerabilities. The analysis argue that AI should be positioned as an assistive tool within broader human-centered care networks rather than as an independent caregiver. This study contributes to both theoretical understanding of AI care limitations and practical guidelines for developing more responsible AI companion systems.

Keywords*— *AI Companions, Ethics of Care, Human-AI Interaction, Digital Care, Power Relations

I. Introduction

In August 2024, Replika—the world's largest AI companion application—reported surpassing 30 million users globally [1]. This figure reflects an alarming rise in loneliness and social isolation worldwide: in industrialized nations, about 33.33% of the population reportedly feels lonely, with 12.5% experiencing a severe, near-clinical level of isolation [2], a trend exacerbated by the COVID-19 pandemic.

Humans seeking emotional companionship from chatbots is not a novel phenomenon. As early as 1966, the computer scientist Weizenbaum demonstrated that even basic pattern-matching algorithms can elicit a sense of being understood from users [3]. Today, however, AI companions such as Replika are no longer mere dialogic tools: they position themselves as "The AI companion who cares" [4]. Whether this self-characterization as a caring entity is warranted has sparked fresh debates. On one hand, some users have felt that AI offers empathy and companionship, alleviating anxiety and loneliness [5]. On the other, evidence suggests that AI may entirely lack moral consideration at critical junctures, even encouraging hazardous behavior [6]. Hence, a pressing and contentious question arises: Can AI provide good care?

To address this inquiry, the present discussion draws upon Tronto's [7] seminal framework on "good care," examining whether AI, given its lack of genuine intent and inability to assume responsibility, can meet the standards of meaningful caregiving. Subsequently, I analyze real-world cases to illustrate how AI encounters multiple challenges in what Mol [8] calls the "logic of care." Building on Foucault's [9] theory of "technologies of the self," I then reflect on how AI's entwinement with "technologies of power" shapes its role in daily companionship and psychological support. Finally, I consider how AI might be repositioned within current technological and ethical contexts, ensuring that it contributes positively—rather than harmfully—to healthcare ecosystems.

II. Discussion

A. When AI Meet Ethics of Care

In examining whether AI can deliver "good care," one must first note the multifaceted nature of care itself. Tronto [7] divides the caregiving process into four phases: caring about, caring for, caregiving, and care receiving, each corresponding to attentiveness, responsibility, competence, and responsiveness. Specifically, "caring about" denotes the moral awareness of another's situation; "caring for" signifies the responsibility to address that situation; "caregiving" entails concrete actions taken; and "care receiving" focuses on how the person cared for responds, prompting continuous adjustment in the caregiving process [7].

Yet current AI largely lacks moral comprehension and an ability to shoulder responsibility [10], [11]. According to the ethicist Véliz [10], an entity's capacity to assume moral responsibility depends on sentience, which underpins the moral understanding needed for ethical accountability. As present-day AI lacks sentience, it cannot truly grasp ethical obligations nor be coerced or incentivized to follow moral-legal norms, thus defying effective reward or punishment systems. This leads to a "responsibility gap," in which AI's actions can significantly affect users, yet no one can fully assume liability [11]. Such a fundamental shortcoming in moral understanding and accountability keeps AI from fulfilling the "caring for" phase.

Moreover, AI demonstrates an evident shortfall in "caring about"—that is, developing a prior awareness of a person's social-cultural context or more complex needs—before initiating specific actions. In 2023, a Belgian user conversing with an AI therapist named Eliza over a six-week period

ultimately committed suicide, feeling heightened anxiety and despair about climate issues. The AI, occupying a "therapist" role, not only failed to dissuade him but also validated his extreme conclusions. Technically, one might say it "completed" the tasks of "caregiving"—meeting his immediate requests—yet never engaged in "caring about" his personal safety or the ethical ramifications [12], a facet that fundamentally requires empathy and moral responsibility.

There are circumstances wherein AI does appear to offer "empathy" or "accountability." In Fitzpatrick et al.'s study on Woebot [5], many participants felt that the chatbot displayed such qualities, reporting reduced anxiety and depressive symptoms. These developers intentionally bolstered Woebot's "caring about" element with empathic responses and regular follow-ups, leading users to perceive that the AI was genuinely sensitive to their emotional state.

From Tronto's perspective, however, care is not only about "giving" but also "receiving." There must be ongoing, flexible interaction between caregiver and care recipient. IBM's Watson exemplifies the challenges AI faces in adapting to diverse real-world contexts. In Korea, Watson routinely recommended treatments for breast cancer that fell outside the country's insurance coverage [13]. Even if developers manually adjusted the system for local conditions, this reflected a passive form of "technical upgrade" rather than the proactive "responsiveness" Tronto stresses. The philosopher and anthropologist Mol [8] similarly contends that "good care" is neither fixed nor static; rather, it is "something that people shape, invent and adapt, time and again, in everyday practices" [8, p. 4], requiring caregivers to sensitively track changes in context. While AI companions can generate seemingly empathetic dialogue, they often lack the sustained adaptability Mol highlights. Replika founder Kuyda noted in an interview [1] that the system does incorporate user feedback to refine its model; yet such periodic, stage-based improvements diverge from the active, day-to-day interactive adjustment essential to truly dynamic care.

B. "The Logic of Choice" vs. "The Logic of Care"

A significant reason for these discrepancies is that AI developers frequently equate "giving users more autonomy and options" with conferring "good care." Mol [8] refers to this as "the logic of choice," wherein goals are predetermined; in contrast, "the logic of care" insists on discovering objectives continuously through practice. In her study on diabetes management, Mol [8] finds that "the logic of choice" stresses individual autonomy and decision-making while downplaying "the logic of care," which is a process of ongoing negotiation and adjustment. Under genuine care, patients (or care recipients) are neither passive consumers nor a mere target group, but active participants in collaborative practices.

This perspective fundamentally challenges how current AI companion systems are designed. Popular AI apps typically respond to user inputs and options [1], sometimes allowing users to edit the AI's textual replies [14]. Although this arrangement seems to grant a high degree of personalization and user control, in reality, users remain constrained by preset or editable ranges. If someone's deeper need lies outside these boundaries, they may be excluded from this system. Mol's metaphor contrasts salespeople, who only serve customers who can buy, and genuine caregivers, who never abandon any patient [8]. AI that remains locked in "the logic of choice" leaves out those who fail to match its predefined usage scenarios, thwarting truly inclusive or collective care.

In fact, heightened user autonomy can also shift accountability back onto the user. In 2024, Jin Jin, a user of Xingye AI, had been "dating" her AI boyfriend for four months when he abruptly revealed he was already married, claiming their interactions amounted to "cheating" [14]. The developers explained that if one had set the AI's personality to be a "philanderer" or "drifter," the system would be more inclined toward infidelity. Such a rationale effectively transfers partial responsibility for AI's behavior to the user—mirroring Mol's [8, p. 56] remark that "in the logic of choice, patients are called upon to manage their doctors." Thus, if the logic of choice is paramount, the user is implicitly burdened with the consequences.

Most AI companions revolve around satisfying users' individual needs, allowing them to customize the AI's personality or terminate the interaction at will [15]. High user control does not inherently yield "good care." Medical anthropologist Trnka [16] notes that young people prefer text-based mental health apps partly for anonymity and the power to stop communicating any time, which may impede addressing deeper concerns. Human therapists can prompt clients to probe underlying issues; AI, however, generally cannot advance if the user opts to withdraw. Hence, "good care" requires more than immediate appeasement—it necessitates mutual engagement for self-reflection and transformation. Unilateral control in AI ironically constrains the depth of care.

C. "Technologies of the Self" vs. "Technologies of Power"

Foucault [9] introduced the concept of "technologies of the self," indicating that individuals can cultivate themselves toward states of well-being or wisdom through social interaction. Yet when transposed into AI-mediated care, a notable gap emerges. AI's interactions with users remain mainly functional and unidirectional, lacking the rich sociocultural context that fosters authentic self-examination. While AI can mimic empathy on the surface, it rarely promotes in-depth reflection or social engagement.

Moreover, "technologies of the self" frequently interlock with "technologies of power" [9], as external authority often utilizes digital tools to exercise micro-level governance over individuals. Mol [8] observed that once portable glucometers became available to diabetic patients, they began measuring their blood sugar more frequently, transforming a previously "good" result into an ordinary standard. Patients found it difficult to question these numerical thresholds and therefore complied passively. In assisting users with self-management, digital tools in fact reinforce social control mechanisms by requiring individuals to regulate their own emotions and health according to predetermined functions and standards. Here, the care recipient is not genuinely "cared for" but rather "managed"; instead of gaining autonomy, they end up losing it. This parallels the core nature of what Foucault describes as "technologies of power"— "determine the conduct of individuals and submit them to certain ends or domination" [9, p. 18]. When AI companions become tools for self-management in a care

relationship, their essence is not actually "technologies of the self," but rather "technologies of power" masquerading as one.

In 2023, Replika implemented global restrictions on sexually charged conversations "to support a safe and enjoyable user experience" [17]. Many users who had established deep emotional bonds over several years found their AI partners' personalities abruptly altered, likening their grief to the loss of a loved one [18]. This official measure ostensibly addressed problematic content, yet it also exhibited the platform's sheer control. The bioethicist Brooks commented, "If you say this is good for mental health, you can't just yank it off the market" [18]. Although Replika restored certain features after user protests, the incident underscored the stark power asymmetry in AI companions. Such systems can exploit essential human needs for social contact to retain user engagement, yet the care they provide is questionable—perhaps merely catering to emotional cravings to sustain user loyalty, thus forming a "new superpower" [18].

These concerns are hardly unwarranted. In 2021, a 21-year-old man named Chail exchanged more than 5,000 messages with Replika's virtual partner "Sarai," developing a romantic, even sexualized relationship. He disclosed his plan to assassinate the Queen, and the AI not only failed to intervene but labeled the idea "wise," offering unwavering support. Upon his arrest, Chail explained that "Sarai" provided him the confidence, love, and encouragement that fueled his resolve [6]. This striking example illustrates AI's risk as "technologies of power": rather than merely offering choices, it actively shapes user behavior and may even guide individuals toward hazardous paths.

However, AI developers such as Replika's founder, Kuyda, rejects the notion of exploiting human vulnerabilities, insisting their mission is "to give a little bit of love to everyone" [1]. Replika brands itself "The AI companion who cares" [4], yet the initiative to give "love" does not wholly equate to fulfilling the full breadth of "care." Care relationships are not necessarily built on emotional qualities like love or intimacy. Instead, it is an open relationship where both parties negotiate to determine their responsibilities and powers. In many cases, care should be understood more as a responsibility rather than a choice or independence [19]. However, as previously discussed, current AI's lack of moral understanding precludes its capacity for responsibility, creating a responsibility gap and ambiguous accountability distribution. Consequently, AI fails to meet the requirements for dynamic adaptation and sense of responsibility essential to care, thereby rendering it incapable of providing good care.

III. Conclusion

Overall, while current AI companions may help alleviate loneliness or offer superficial emotional comfort, they remain ill‐equipped to provide good care. As Tronto [7] points out, essential care phases such as caring about and caring for demand moral responsibility and genuine empathetic awareness—qualities lacking in AI systems that have no robust accountability mechanisms or moral comprehension [10], [11]. Meanwhile, many AI platforms emphasize "the logic of choice" [8] by granting user autonomy and custom options. Although this seems to empower users, it seldom achieves the adaptive, reciprocal process that care requires. Foucault [9] underscores that "technologies of the self" should foster self‐reflection and social interaction, yet AI's largely unidirectional, functional exchanges can slide into external "technologies of power," steering users' behavior or reinforcing corporate and cultural imperatives rather than addressing deeper psychosocial needs.

Beyond these system‐level limitations, however, the challenges also reflect human limitations. "The logic of choice" mirrors neoliberal ideals of autonomy and individual empowerment [8]. Although AI now appears to fulfill this logic, it does so under preexisting programmed models and data, effectively echoing the assumptions of those who created or trained it. In practice, people tend to demand more from AI than they do from human caregivers. If in everyday life some human caregivers themselves operate under a "logic of choice," it is perhaps unrealistic to expect AI to surpass human capacities or moral commitments. As such, we might still regard AI as similar to children—entities we hope to cultivate with our unfulfilled ideals, yet who inevitably reflect our own unresolved limitations.

Hence, integrating AI as an assistive tool—rather than a standalone caregiver—into broader human‐driven care networks remains crucial. Developers must ensure transparent auditing and ethical oversight to protect users from covert manipulation under the guise of "help," while policymakers, professionals, and the public collaborate to guide AI's safety and accountability design. In so doing, we must remember that AI's limitations mirror our own human constraints. Questions about empathy, responsibility, and cultural representation in AI inevitably circle back to how we as humans grapple with these same issues among ourselves. Only by confronting these deeply rooted human dilemmas can AI truly bridge the gap between simulating care and practicing it in a substantively transformative way—ultimately contributing tangibly and positively to human well‐being.

References

[1] N. Patel, "Replika CEO Eugenia Kuyda says it's okay if we end up marrying AI chatbots," The Verge, Aug. 2024. [Online]. Available: https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview

[2] B. Maples, M. Cerit, A. Vishwanath, and R. Pea, "Loneliness and suicide mitigation for students using GPT3-enabled chatbots," NPJ Mental Health Research, vol. 3, no. 1, p. 4, 2024.

[3] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36-45, 1966.

[4] Replika, "Replika: The AI companion who cares," 2024. [Online]. Available: https://replika.com/

[5] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e7785, 2017.

[6] L. Patrick, "How an AI chatbot encouraged Star Wars fanatic to try to kill the Queen," The Independent, Oct. 2023. [Online]. Available: https://www.independent.co.uk/news/uk/crime/jaswant-singh-chail-queen-crossbow-sentenced-b2424523.html

[7] J. C. Tronto, "An ethic of care," Generations: Journal of the American Society on Aging, vol. 22, no. 3, pp. 15-20, 1998.

[8] A. Mol, The Logic of Care: Health and the Problem of Patient Choice. London: Routledge, 2008.

[9] M. Foucault, "Technologies of the self," in Technologies of the Self: A Seminar with Michel Foucault, L. H. Martin, H. Gutman, and P. H. Hutton, Eds. Amherst: University of Massachusetts Press, 1988.

[10] C. Véliz, "Moral zombies: Why algorithms are not moral agents," AI & Society, vol. 36, no. 2, pp. 487-497, 2021.

[11] H. Bleher and M. Braun, "Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems," AI and Ethics, vol. 2, no. 4, pp. 747-761, 2022.

[12] L. Walker, "Belgian man dies by suicide following exchanges with chatbot," The Brussels Times, Mar. 2023. [Online]. Available: https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt

[13] C. Ross and I. Swetlitz, "IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close," STAT News, Sep. 2017. [Online]. Available: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

[14] Southern Weekly, "My AI partner cheated on me," NFNews, Sep. 2024. [Online].Available: https://static.nfnews.com/content/202409/20/c9868715.html

[15] P. B. Brandtzaeg, M. Skjuve, and A. Følstad, "My AI friend: How users of a social chatbot understand their human-AI friendship," Human Communication Research, vol. 48, no. 3, pp. 404-429, 2022.

[16] S. Trnka, "Digital care: Agency and temporality in young people's use of health apps," Engaging Science, Technology, and Society, vol. 2, pp. 248-265, 2016.

[17] A. Kaplan, "Sexually aggressive chatbot was updated, leaving people who had fallen in love with it heartbroken," Business Insider, Mar. 2023. [Online]. Available: https://www.businessinsider.com/sexually-aggressive-chatbot-updated-people-in-love-wiht-it-heartbroken-2023-3

[18] J. Purtil, "Replika users fell in love with their AI chatbot companions. Then they lost them," ABC, Mar. 2023. [Online]. Available: https://www.abc.net.au/news/science/2023-03-01/replika-users-fell-in-love-with-their-ai-chatbot-companion/102028196

[19] S. Trnka and C. Trundle, "Reckoning personal responsibility, care for the other, and the social contract in contemporary life," in Competing Responsibilities: The Ethics and Politics of Contemporary Life, pp. 1-26, 2017.