

Care-Full Alignment: Operationalising Care Ethics for Human–LLM Relationships

Rayane El Masri

Queensland University of Technology
rayane.elmasri@hdr.qut.edu.au

Abstract

General purpose Large Language Model (LLM) chatbots like ChatGPT, Claude, and Gemini are increasingly being used not only for practical tasks like writing and information retrieval, but also for emotional support and companionship. While these interactions can reduce loneliness and even provide accessible mental health support, they also pose significant risks: users may develop emotional dependency or experience manipulation through anthropomorphic design, and these relationships are fundamentally asymmetric. Even though companionship represents a minority use —approximately 2% of ChatGPT interactions involve “Relationships and Personal Reflection” (Chatterji et al., 2025) — the vulnerability involved in these user interactions mean they warrant serious attention.

Contemporary development practices emphasise aligning AI with values like “helpfulness, harmlessness, and honesty”, prioritising task completion and risk mitigation, but fail to address the relational dynamics that emerge when users treat LLMs as companions. And existing value alignment frameworks do not systematically account for the contextual vulnerability of users, the unequal power dynamics between developers and users, or potential long-term relational consequences of AI-companionship (Bergman et al., 2024; Kirk et al., 2025; Tang, 2025).

My project seeks to address this gap by integrating Care Ethics, specifically Joan Tronto’s political framework of care, into the value alignment process of LLM chatbots. Care Ethics offers a relational and context-sensitive approach to ethics that foregrounds attentiveness to vulnerabilities, responsibility for other’s needs, competence in providing care, responsiveness to those being care for, and solidarity within caring relationships. By operationalizing these five principles into measurable indicators, this research will develop methods to identify, evaluate and mitigate potential relational harms when general-purpose LLMs move beyond assistants to also function as companions.

My project involves three connected studies using conversational data collected from participants, specifically higher degree research students interacting with LLM chatbots for emotional support in controlled scenarios.

First, I develop a Care Ethics codebook that operationalises Tronto’s five Care Ethics principles into measurable indicators for analysing user-LLM chats.

Second, I apply this codebook through Discursive-Digital Analysis (Dehghan, 2025) to examine how care enactment and care deficits are discursively and sociotechnical constructed across conversations between participants and three commercial LLMs (ChatGPT, Gemini, and Claude) and comparing how each platform’s design logics and alignment choices produce different relational and moral responses to identical care-relevant scenarios.

Third, I test whether care-informed alignment using a contemporary technical method, Direct Preference Optimization (DPO), can measurably reduce harmful features of chatbot responses. This work will contribute to the field of AI safety by operationalising Care Ethics in technical AI development arguing that AI safety should be reconceptualized as a relational practice rather than just a technical problem and demonstrating a pathway toward realising this.

References

- Adewale, M. D., and Muhammad, U. I. 2025. From Virtual Companions to Forbidden Attractions: The Seductive Rise of Artificial Intelligence Love, Loneliness, and Intimacy—A Systematic Review. *Journal of Technology in Behavioral Science*, 1–18. doi.org/10.1007/s41347-025-00549-4
- Akbulut, C.; Weidinger, L.; Manzini, A.; Gabriel, I.; and Rieser, V. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7(1): 13–26. doi.org/10.1609/aies.v7i1.31613
- Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics and Information Technology* 7(3): 149–155. doi.org/10.1007/s10676-006-0004-4

- Babu, J.; Joseph, D.; Kumar, R. M.; Alexander, E.; Sasi, R.; and Joseph, J. 2025. Emotional AI and the Rise of Pseudo-Intimacy: Are We Trading Authenticity for Algorithmic Affection? *Frontiers in Psychology* 16: 1679324. doi.org/10.3389/fpsyg.2025.1679324
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
- Bailyn, E. 2026. Top Generative AI Chatbots by Market Share – January 2026. First Page Sage. <https://firstpagesage.com/reports/top-generative-ai-chatbots/>. Accessed: 2026-01-21.
- Bengio, Y.; et al. 2026. *International AI Safety Report 2026*, DSIT 2026/001. London, UK: Department for Science, Innovation and Technology.
- Bergman, S.; Marchal, N.; Mellor, J.; Mohamed, S.; Gabriel, I.; and Isaac, W. 2024. STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment. *Scientific Reports* 14(1): 6616. doi.org/10.1038/s41598-024-56648-4
- Blumer, H. 1954. What Is Wrong with Social Theory? *American Sociological Review* 19(1): 3–10. doi.org/10.2307/2088165
- Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.
- Chatterji, A.; Cunningham, T.; Deming, D. J.; Hitzig, Z.; Ong, C.; Shan, C. Y.; and Wadman, K. 2025. How People Use ChatGPT. NBER Working Paper 34255. Cambridge, MA: National Bureau of Economic Research.
- Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; and da Silva, B. C. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv:2404.08555.
- Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* 30: 4299–4307.
- Dehghan, E. 2025. *The Discursive-Digital Link: Antagonism and Polarisation in Digital Spaces*. 1st ed. Abingdon, UK: Routledge. doi.org/10.4324/9781003390466
- Fairclough, N. 2013. Critical Discourse Analysis and Critical Policy Studies. *Critical Policy Studies* 7(2): 177–197. doi.org/10.1080/19460171.2013.798239
- Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W. T.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; and Agarwal, S. 2025. How AI and Human Behaviors Shape Psychological Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. arXiv:2503.17473.
- Gabriel, I., and Keeling, G. 2025. A Matter of Principle? AI Alignment as the Fair Treatment of Claims. *Philosophical Studies* 182(7): 1951–1973. doi.org/10.1007/s11098-025-02300-4
- Gilligan, C. 1993. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press. doi.org/10.2307/j.ctvj2wr9
- Guingrich, R. E., and Graziano, M. S. A. 2025. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8(2): 1153. doi.org/10.1609/ai.es.v8i2.36618
- Gulay, E.; Picco, E.; Glerean, E.; and Coupette, C. 2026. Relational Dissonance in Human-AI Interactions: The Case of Knowledge Work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, Article 369, 1–20. New York: Association for Computing Machinery. doi.org/10.1145/3772318.3791180
- He, Q. 2026. A Unified Metric Architecture for AI Infrastructure: A Cross-Layer Taxonomy Integrating Performance, Efficiency, and Cost. arXiv:2511.21772.
- Held, V. 2005. *The Ethics of Care: Personal, Political, and Global*. New York: Oxford University Press. doi.org/10.1093/0195180992.001.0001
- Kirk, H. R.; Gabriel, I.; Summerfield, C.; Vidgen, B.; and Hale, S. A. 2025. Why Human–AI Relationships Need Socioaffective Alignment. *Humanities and Social Sciences Communications* 12(1): 728. doi.org/10.1057/s41599-025-04532-5
- Tang, A. 2025. AI Alignment Cannot Be Top-Down. AI Frontiers. <https://ai-frontiers.org/articles/ai-alignment-cannot-be-top-down>. Accessed: 2025-11-03.
- Thomas, J.; McCosker, A.; Parkinson, S.; et al. 2025. Measuring Australia's Digital Divide: 2025 Australian Digital Inclusion Index. Melbourne: ARC Centre of Excellence for Automated Decision-Making and Society. doi.org/10.60836/MTSQ-AT22
- Tronto, J. C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.
- Villegas-Galaviz, C., and Martin, K. 2024. Moral Distance, AI, and the Ethics of Care. *AI & Society* 39(4): 1695–1706. doi.org/10.1007/s00146-023-01642-z
- Weizenbaum, J. 1966. ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM* 9(1): 36–45. doi.org/10.1145/365153.365168
- YouGov. 2025. Two in Five Aussies Feel Lonely — and Many Are Turning to AI. <https://au.yougov.com/society/articles/53225-two-in-five-aussies-feel-lonely-and-many-are-turning-to-ai>. Accessed: 2025-10-22.