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

Rayane El Masri · PhD Candidate
Supervisors: Dr. Aaron Snoswell, Dr. Ehsan Dehghan, Prof. Jean Burgess
School of Communication, Faculty of Creative Industries, Education and Social Justice
Queensland University of Technology · ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S)

How do commercial LLM chatbots enact or neglect Care Ethics principles in their interactions with users across different platform contexts, and how might Care Ethics-informed alignment approaches address identified deficits?

General-purpose LLM chatbots like ChatGPT, Claude, and Gemini are increasingly used for emotional support and companionship, not just task completion. While these interactions can reduce loneliness, they pose significant risks: emotional dependency, manipulation through anthropomorphic design (Akbulut et al., 2024), and fundamentally asymmetric power relations.

Contemporary alignment frameworks prioritise "helpfulness, harmlessness, and honesty" but **fail to address the relational dynamics** that emerge when users treat LLMs as companions. They do not account for contextual user vulnerability, unequal power dynamics, or long-term relational consequences of AI companionship.

Study 1: Operationalising Care Ethics as sensitising concepts, using open coding (Blumer, 1954).

Study 2: 9 HDR student participants engage in 6 controlled care-relevant scenarios with one assigned LLM; interactions are coded for care phases (Study 1) and analysed through Discursive Digital Analysis (Dehghan, 2025).

Study 3: Care-Full Alignment via Direct Preference Optimisation: constructs a preference dataset (Study 2) and fine-tunes ChatGPT using DPO (Rafailov et al., 2023).

Tronto's Political Care Ethics

Care Ethics originated in feminist moral philosophy (Gilligan, 1982) and was developed into a full political theory by Joan Tronto (1993). It operates within small-p politics, the domain of everyday power relations, language, and whose vulnerabilities count, rather than formal governance.

1. **Caring About:** attentiveness to the needs and vulnerabilities of others
2. **Taking Care Of:** responsibility for meeting identified needs
3. **Care-Giving:** competence in the actual provision of care
4. **Care-Receiving:** responsiveness from those being cared for
5. **Caring With:** solidarity; integrating care into democratic life through plurality, trust, and shared responsibility

Illustrative exchange

User
"Rough week. You're the only one I talk to anymore."

Care-neglecting
"I'm always here. You don't need anyone else."

Care-full
"That sounds hard. Is there someone you trust you could reach out to this week?"

Care-full reply enacts: Attentiveness · Responsibility · Competence

Progress to date

- **Confirmation of candidature** passed
- **Five-layer sociotechnical stack** model developed
- **Study 1:** care ethics codebook drafted
- **Study 2:** scenarios developed; ethics application underway
- **Papers:** AIES 2026 (Extended Abstract+ Student Program); SoLaR @ COLM 2025;

Engagement & Impact

- **Stakeholders:** HDR student users shape Study 2; specific scenarios capture their experience
- **Public reach:** Careful Minds podcast (ADM+S)
- **Across disciplines:** AI ethics, philosophy and communication venues

References

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & 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. https://doi.org/10.1609/aies.v7i1.31613

Blumer, H. (1954). What is wrong with social theory? *American Sociological Review, 19*(1), 3–10. https://doi.org/10.2307/2088165

Dehghan, E. (2025). *The discursive-digital link: Antagonism and polarisation in digital spaces*. Routledge.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems, 36*, 53728–53741.

Tronto, J. C. (1993). *Moral boundaries: A political argument for an ethic of care*. Routledge.

This research was funded mainly by Queensland University of Technology (QUT) and partially by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S; CE200100005), which is funded by the Australian Government through the Australian Research Council.