

Designing for the Intentional Stance: Stance Misalignment, Value Alignment, and the Case for Care Ethics in Human-LLM Relationships (Extended Abstract)

Rayane El Masri, Aaron Snoswell, Ehsan Dehghan

Queensland University of Technology

rayane.elmasri@hdr.qut.edu.au, a.snoswell@qut.edu.au, e.dehghan@qut.edu.au

Abstract

Users are increasingly forming emotionally charged relationships with general-purpose large language models (LLMs), producing documented patterns of dependency, relational harm, and affective disappointment. Existing value alignment frameworks, including Reinforcement Learning from Human Feedback, Constitutional AI, and the Helpful-Harmless-Honest principles, have failed to prevent these harms even where they succeed on their own technical terms. This paper argues that the root cause of this failure is a structural interpretive gap produced by LLM design itself: a stance misalignment between developers, who approach LLMs through Daniel Dennett's design stance, and users, who adopt the intentional stance in response to LLMs' linguistic fluency, contextual memory, and performed empathy. Drawing on Dennett's intentional systems theory, this paper makes three interrelated contributions. First, it diagnoses stance misalignment as the mechanism through which a distinct and underdiagnosed form of the value alignment problem arises: relational harm in human-LLM interactions. Second, it considers three possible structural resolutions to this misalignment — correcting users' interpretations (convergence downward), removing the design features that produce it (deflationary design), and developers adopting the intentional stance as their own operative moral framework (convergence upward) — and argues that the third is the most tractable and ethically appropriate. Third, it proposes Care-Full Alignment, an operationalization of Joan Tronto's Political Care Ethics, as the normative framework convergence upward demands.

Extended version — <https://eprints.qut.edu.au/267194/>

References

- Akbulut, C.; Weidinger, L.; Manzi, A.; Gabriel, I.; and Rieser, V. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 13–26. doi.org/10.1609/aies.v7i1.31613
- Aske, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Kernion, J.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; and Kaplan, J. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861. Ithaca, NY: Cornell University Library.
- 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.; Aske, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073. Ithaca, NY: Cornell University Library.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi.org/10.1145/3442188.3445922
- Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. Ithaca, NY: Cornell University Library.
- Booth, R. 2026. Richard Dawkins Concludes AI Is Conscious, Even If It Doesn't Know It. *The Guardian*, May 5.
- 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 No. 34255. doi.org/10.3386/w34255
- Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; and Silva, B. C. da. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv:2404.08555. Ithaca, NY: Cornell University Library.
- 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.
- Dahlgren Lindstrom, A.; Methnani, L.; Krause, L.; Ericson, P.; de Rituerto de Troya, I. M.; Coelho Mollo, D.; and Dobbe, R. 2025. Helpful, Harmless, Honest? Sociotechnical Limits of AI Alignment and Safety through Reinforcement Learning from Human Feedback. *Ethics and Information Technology*, 27(2): 28. doi.org/10.1007/s10676-025-09837-2
- Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
- 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 Psychosocial Effects of Extended Chatbot Use. arXiv:2503.17473. Ithaca, NY: Cornell University Library.
- Gabriel, I., and Ghazavi, V. 2021. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. arXiv:2101.06060. Ithaca, NY: Cornell University Library.
- Gausen, A.; Wallbridge, S.; Kirk, H. R.; Williams, J.; and Summerfield, C. 2026. Disclosure by Design: Identity Transparency as a Behavioural Property of Conversational AI Models. arXiv:2603.16874. Ithaca, NY: Cornell University Library.
- Gunkel, D. J., and Coeckelbergh, M. 2025. A Relational Approach to Moral Standing: Reframing Ethical Boundaries in the Age of Artificial Intelligence. In Vandenbergh, F., and Papilloud, C., eds., *New Directions in Relational Sociology, Volume Two: Relations All the Way Down*, 285–302. Cham: Springer Nature Switzerland. doi.org/10.1007/978-3-032-02413-8_12
- Hadfield-Menell, D., and Hadfield, G. K. 2019. Incomplete Contracting and AI Alignment. arXiv:1804.04268. Ithaca, NY: Cornell University Library.
- 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
- Klein, S. H. 2025. The Effects of Human-Like Social Cues on Social Responses towards Text-Based Conversational Agents: A Meta-Analysis. *Humanities and Social Sciences Communications*, 12(1): 1322. doi.org/10.1057/s41599-025-05618-w
- LaCroix, T. 2026. Relative Principals, Pluralistic Alignment, and the Structural Value Alignment Problem. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, 622–639. New York: Association for Computing Machinery. doi.org/10.1145/3805689.3812420
- Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2024. Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika. *New Media & Society*, 26(10): 5923–5941. doi.org/10.1177/14614448221142007
- Maeda, T., and Quan-Haase, A. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1068–1077. New York: Association for Computing Machinery.
- Manzini, A.; Keeling, G.; Alberts, L.; Vallor, S.; Morris, M. R.; and Gabriel, I. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 943–957. doi.org/10.1609/aies.v7i1.31694
- Marchegiani, B. 2025. Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy. *Journal of Applied Philosophy*, 42(5): 1399–1419. doi.org/10.1111/japp.70008
- Morrin, H.; Nicholls, L.; Levin, M.; Yiend, J.; Iyengar, U.; Del-Guidice, F.; Bhattacharya, S.; Tognin, S.; MacCabe, J.; Twumasi, R.; Alderson-Day, B.; and Pollak, T. A. 2026. Artificial Intelligence-Associated Delusions and Large Language Models: Risks, Mechanisms of Delusion Co-Creation, and Safeguarding Strategies. *The Lancet Psychiatry*, 13(6): 522–530. doi.org/10.1016/S2215-0366(25)00396-7
- Nass, C., and Moon, Y. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1): 81–103. doi.org/10.1111/0022-4537.00153
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- Phang, J.; Lampe, M.; Ahmad, L.; Agarwal, S.; Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W. T.; Pataranutaporn, P.; and Maes, P. 2025. Investigating Affective Use and Emotional Well-being on ChatGPT. arXiv:2504.03888. Ithaca, NY: Cornell University Library.
- Pierre, J. M.; Gaeta, B.; Raghavan, G.; and Sarma, K. V. 2025. "You're Not Crazy": A Case of New-onset AI-associated Psychosis. *Innovations in Clinical Neuroscience*, 22(10–12): 11–13.
- Reeves, B., and Nass, C. I. 1996. *The Media Equation*. Cambridge: Cambridge University Press.
- Roth, E. 2025. ChatGPT Now Has over 300 Million Weekly Users. *The Verge*, December 5.
- Rovane, C. 1994. The Personal Stance. *Philosophical Topics*, 22(1/2): 351–396. doi.org/10.5840/philtopics1994221/25
- Russell, S. J., and Norvig, P. 2022. *Artificial Intelligence: A Modern Approach*. 4th ed. London: Pearson.
- Smith, M. G.; Bradbury, T. N.; and Karney, B. R. 2025. Can Generative AI Chatbots Emulate Human Connection? A Relationship Science Perspective. *Perspectives on Psychological Science*, 20(6): 1081–1099. doi.org/10.1177/17456916251351306
- Srnicek, N. 2017. The Challenges of Platform Capitalism: Understanding the Logic of a New Business Model. *Juncture*, 23(4): 254–257. doi.org/10.1111/newe.12023
- Srnicek, N. 2026. *Silicon Empires: The Fight for the Future of AI*. Cambridge: Polity.
- Tronto, J. C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.
- Vermaas, P. E.; Carrara, M.; Borgo, S.; and Garbacz, P. 2011. The Design Stance and Its Artefacts. *Synthese*, 190(6): 1131–1152. doi.org/10.1007/s11229-011-9885-9
- 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
- Wertheimer, T. 2022. Blake Lemoine: Google Fires Engineer Who Said AI Tech Has Feelings. *BBC News*, July 23.