AI Should Develop Human Empathy, Not Replace It

Ethan Landes and Jim A.C. Everett
University of Kent

Correspondence Email: E.Landes@Kent.ac.uk

Forthcoming in A. Perry & C. D. Cameron (Eds.), *Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Considerations.* Cambridge University Press.

Empathizing with another is perhaps the most human emotion of all and a long shot from the cold calculations of vectorized tokens that drive the inner workings of contemporary LLMs. This chapter looks beyond the questions of what to make of AI-generated "empathy" and instead asks whether AI can be used to develop our capacities for genuine human empathy. Empathy is not a static trait, instead capable of growth and development, and this paper explores whether AI can and, more importantly, should be used to increase one's empathy. Nothing in principle stands in the way of AI improving our empathy, but the possibility raises unanswered questions about whether such an approach would be effective or would backfire in unexpected ways, such as encouraging the commodification of empathy as a technological tool that can make companies money, rather than a fundamental part of the human experience.

Keywords: Artificial Moral Enhancement, Empathy Training, Virtue Ethics, Artificial Empathy

Empathy - the ability to understand and share the feelings of another person - is in many ways at the heart of our social lives. Our feelings of empathy are thought to help us connect with others, build trust, and form meaningful relationships (Decety & Ickes, 2009; de Waal, 2008). It is empathy that allows us to respond with care and understanding, and it is empathy that can often reduce conflict and foster cooperation (Vanman, 2016). Empathy is clearly fundamental. But in an increasingly AI-mediated world where people not only interact with each other through technologies like smartphones and computers but also interact more and more with non-human "agents" powered by LLMs (large language models), how should we think about empathy? Here, in contrast to discussions that have tended to dominate the literature, we distinguish *three* sorts of empathy-involving relationships between humans and AI.

First, we might *empathize with AI*, where AI is the object of human empathy. Insofar as this is a genuine mental state of a human involving the (perceived or imagined) mental states of something else, empathy directed towards AI is *genuine* empathy. One might worry empathy is misplaced when directed towards AI and that our limited empathetic resources should instead be directed towards humans. However, as anyone who has cried at the end of the movies Iron Giant, Big Hero Six, or Terminator 2 knows, not only can our feelings towards robots be very real, but we regularly empathize with things like fictional entities that lack any genuine mental life to empathize with. One might also worry that the amount of anthropomorphization required to empathize with AI is fundamentally damaging to human users, elevating the risk of alienation, false beliefs, gaslighting, or even psychosis (Kleinman, 2025; Østergaard, 2023), but the mental state of empathy is nonetheless occurring. Therefore, even if, normatively, one believes AI is not the sort of thing that we should empathize with, descriptively, AI can certainly be the target of empathy.

Second, AI may *"empathize" with us,* where a human is the object of the empathy (Inzlicht et al., 2024). This is *fake* empathy, as contemporary forms of AI neither understand nor share our feelings (Perry, 2023; Shteynberg et al., 2024). When an LLM chatbot tells us they are sorry for our situation, their output is merely the result of their underlying transformers recursively predicting the next word in a sequence based on statistical representations of their training data. Even if these outputs are interpreted as warm and moving (Ayers et al., 2023; Wenger et al., 2025), they are unreal pastiches of the genuine empathy felt by humans (Vallor, 2024).

While much discussion in the recent literature has focused on the distinction between real and fake empathy and the potential for AI as a provider of empathy, there is a further third possible relationship between humans, empathy, and AI: AI may help users *develop their own capacity for empathy*. Empathy can be developed, enhanced, and trained (Lam et al., 2011; Teding van Berkhout & Malouff, 2016; Wu et al., 2024), and AI might be able to aid in this process. Rather than a specific moment, thought, or emotion directed at a specific person or thing, here, the involvement of AI serves to develop the moral motivations, skills, or sensitivities of the human user. While the AI's outputs are phony, the resulting empathy is not. If used correctly, AI could, at least in theory, inspire the human user to improve their moral motivations and develop moral insight by inspiring moral growth in users (Landes et al., 2025). AI could serve as a sort of moral scaffolding, inspiring empathy-developing reflection and thought on behalf of the users, resulting in genuine empathy directed at things other than AI.

Could AI be effective at developing our empathy?

Even if empathy can be trained, undirected LLM use is extraordinarily unlikely to improve empathy in users. Moral growth from LLM use likely requires directed and effortful engagement (Landes et al., 2025), and LLMs appear to decrease the amount of effort users spend on tasks (Lee et al., 2025).

Rather than giving people access to AI and hoping it improves their empathy, a more effective approach is likely to be one in which AI is used as a tool embedded in procedures known to improve empathy. For example, human-based simulations have been found to successfully improve empathy, where trainees either take the role of the person being empathized with (Larti et al., 2018) or practice interacting with the target of empathy as themselves (Gholamzadeh et al., 2018). LLMs may be able to augment in-person training by playing either the subject or object of empathy in text-based role-playing exercises with the trainee, with hidden prompts to hit certain beats or emphasize certain topics. Similarly, other empathy training protocols augment lectures and forms of online training with discussion boards or writing tasks (Mueller et al., 2018; Shapiro et al., 2006). Here, LLMs may be incorporated to encourage further reflection on discussion board posts, generating follow-up questions that tie together aspects of participants'

posts and key themes preselected by the course designer. This could potentially fruitfully be combined with other approaches, such as simulation approaches, that employ reflective cycles.

Deploying LLMs in these, or even more sophisticated ways, is well within the scope of contemporary technology. The growing sophistication of LLMs – particularly increases in context windows (the LLM counterpart to human working memory) – mean that LLMs can be prompted to follow sophisticated training regimens. These include personalizing responses to users' demographics, carrying forward information from earlier interactions with users', and following procedures with specific checkpoints or triggers.

Would AI be effective at developing our empathy?

There are reasons to be optimistic that AI could be used as a tool to enhance empathy. Many empathy-increasing practices include steps that work with the primarily text-based medium of contemporary AIs. Moreover, the current body of research has found that AI, if anything, is more capable than humans at producing moving and apparently empathetic text-based responses (Ayers et al., 2023; Wenger et al., 2025). To the extent that empathy training works to improve someone's ability to communicate compassion or perspective-taking via language, we could be optimistic about the role of LLMs as trainers.

There are also, however, reasons to be pessimistic about AI's ability to be a tool for improving human empathy. First, people trust AI less than human decision-makers in a phenomenon known as "algorithm aversion" (Dietvorst et al., 2015). Even if AI can generate more apparently moving empathetic statements when people do not know the source, when people know they are generated by AI (which, ethically, they must), they show resistance (Wenger et al., 2025). Therefore, humans using AI-driven empathy training tools may, regardless of the quality of AI outputs, fail to engage with the training to the extent they would with other humans. That said, in some limited cases, people can trust AI more than humans (Logg et al., 2019), and in studies of algorithmic aversion in the moral domain trust in AI is not zero, just comparatively less than humans (e.g., Myers & Everett, 2025). Therefore, even if algorithmic aversion prevents engagement with AI-driven training, it will likely not completely eliminate it.

LLMs may additionally struggle to be an effective tool of developing human empathy because of the primarily text-based medium of modern LLMs. Text-based exchanges with LLMs lack depth and miss important elements like facial expressions and vocal tone that are central to human interactions (Crockett, 2025; Perry, 2023), and developing empathy toward others through AI conversations seems unlikely to foster every real-world embodied skill that empathy depends on. It misses key elements like reading facial expressions, sensing hesitation, and offering gestures like a smile or a gentle touch - crucial parts of genuine human connection. This does not mean AI-based training is likely to be completely ineffective, as text can still be an effective medium for expressing and exchanging empathy, and AI models are likely to become increasingly multimodal as technology advances.

Finally, AI use, regardless of the format or medium, may *degrade* human empathy. Building on ideas from virtue ethics, we might think that empathy is a skill that must be practiced, a form of connective or emotional labour that ties us together. If AI steps in too often as a support system, even to encourage reflection, it is plausible to imagine that we become less proficient at spontaneously practicing empathy ourselves in the same way that other kinds of deskilling occur when relying too much on AI (Sambasivan & Veeraraghavan, 2022). Similarly, if people grow accustomed to using an LLM that always listens and responds without judgment, they might lose patience for the messiness and complexity of real human emotions, which require effort, vulnerability, and compromise. These are real possibilities, and we simply do not know yet what the long-term effects of AI-mediated empathy training would be (for more discussion, see Shastry et al., this volume; Lee et al., this volume).

Should AI be used to develop empathy?

Overall, there are reasons to be optimistic that modern LLMs could successfully be employed as part of larger procedures to improve the empathy of human trainees. It is a separate question of whether they *should* be used to improve the empathy of users. We take the answer not only to be a qualified yes but that it is a better lens through which to focus on AI-related empathy than is typically considered.

As discussed in Section 1, discussion of the role of empathy in human-AI relationships has typically focused on whether AI can or should replace human empathizers. Endorsing the use of AI-produced empathy requires thinking that while it may be phony, AI "empathy" is better than nothing - it is an ersatz substitute but an improvement, nonetheless. For example, because human empathizers tire out and are not good at empathizing with marginalized groups, for use cases where empathy fatigue or marginalization are a risk, AI empathizing could be an improvement on the state of affairs we have now (Inzlicht et al., 2024). Taking this strategy, the possibilities for AI are nearly endless. AI empathizers could free up the emotional workload of doctors, augment HR departments, or provide companionship to lonely people (Pugh, 2024; see also Lee et al., this volume).

The ethical problem with focusing on AI "empathy" is in how it cheapens empathy. As something produced by AI, "empathy" is no longer a fundamentally human virtue and way of connecting with others. Empathy is reduced to nothing more than a means to an end – a design problem to solve. Because many people are lonely, AI can keep them company. Because doctors are not always the best at bedside manner, AI can replace them to better comfort humans. Because training HR caseworkers is expensive and time-consuming, AI can take over some of the workload. Through this lens, empathy is a way of mollifying customers, creating the perception of friendship, or removing the need for genuine human-to-human connections in trying times. Empathy is a commodity to be purchased from tech companies, one API credit at a time.

The pressures towards cheapening empathy will doubtless only grow stronger as AI becomes more sophisticated and AI companies offer new "solutions" related to empathy. This can be fought by shifting focus away from replacing human empathy with AI "empathy" towards instead using AI to *develop* human empathy, thereby retaining empathy's value and its fundamental humanity. If AI can be used to successfully increase a person's empathy, the resulting growth in empathy would be genuine, even if the means to arrive at that empathy were fake. The origin of one's empathy-related skills and motivations do not change whether or not we genuinely empathize in the same way the $5 in our pocket is still $5, regardless of whether it ended up there as part of our monthly salary, we found it on the street, or we stole it in a violent robbery (Landes et al., 2025). Used as a tool for our moral development, AI doesn't limit us or

cut us off from others and limit our moral growth (pace Vallor, 2024; for a related point, see Berio, this volume). Instead, AI helps us develop our own virtues and flourish as a community of moral and social agents. We thereby preserve our autonomy in the face of AI (Floridi & Cowls, 2019) because we are not yielding ground to AI but using it as scaffolding.

Even if using AI as a tool seems less fraught than using AI to replace humans, it is not without potential downsides. First, replacing humans with AI in empathy training risks alienating humans in the training because humans will know (and ethically should know) that they are interacting with an AI that is not displaying genuine empathy and is not genuinely deserving of the human's empathy, potentially causing existential concerns. Relatedly, AI being used as a tool – even a tool for growth – could encourage a superficial or performative focus on merely saying the right things rather than forging genuine emotional connections with others. Contingent features of the AI models employed may also cause problems. Moreover, given the cultural bias in LLMs' training data (Tao et al., 2024), teaching empathy through AI might subtly impose biased or narrow versions of what "good" empathy looks like, directing more empathy to those who already receive it, while further marginalizing other cultural and moral perspectives. The corporate actors responsible for developing and maintaining the deployed AI models may simultaneously be incentivized to manipulate AI models to develop empathy in ways that further their interests, by, for example, encouraging empathy towards corporations or their AI products.

Perhaps most critically, though, we must consider the question of how AI could improve empathy against the broader backdrop: one in which AI companies use their money and influence to change politics to align with their own economic interests, where the pursuit of "efficiency" leads to job displacement and economic precarity, and where the labour of training AI models is often outsourced to underpaid and exploited workers in the global South. Perhaps we are willing to ignore this broader context and accept these consequences as a reasonable cost for the greater good if AI could meaningfully and strongly enhance our empathy and make a real difference in the world. There nonetheless remains a deeper risk for how widespread use of this technology could change the way we understand ourselves and others. To even entertain the idea that AI could be used to train or enhance empathy is to risk opening a discursive space in which, by talking about whether it is possible, we implicitly assume that it is desirable in the first place. By moving to the idea that we should rely on AI for either providing or enhancing empathy, we

risk empathy becoming another product or service: something that could be delivered via subscription or branded as a self-help solution rather than an integral, shared part of the human experience.

Funding Declaration

This work was generously supported by a Philip Leverhulme Prize (PLP-2021-095), and the Horizon Europe UK Guarantee via the UKRI (EP/Y00440X/1) awarded to JACE.

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Berio, L. (in press). Stories of empathy: Empathy, interactions with AI, and identity building. In A. Perry & C. D. Cameron (Eds.), Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Considerations. Cambridge University Press.

Crockett, M. (2025, February 27). AI is 'beating' humans at empathy and creativity. But these games are rigged. The Guardian. https://www.theguardian.com/commentisfree/2025/feb/28/ai-empathy-humans

de Waal, F. B. M. (2008). Putting the Altruism Back into Altruism: The Evolution of Empathy. Annual Review of Psychology, 59(Volume 59, 2008), 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

Decety, J., & Ickes, W. J. (Eds.). (2009). The Social Neuroscience of Empathy. MIT Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology. General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Gholamzadeh, S., Khastavaneh, M., Khademian, Z., & Ghadakpour, S. (2018). The effects of empathy skills training on nursing students' empathy and attitudes toward elderly people. BMC Medical Education, 18(1), 198. https://doi.org/10.1186/s12909-018-1297-9

Inzlicht, M., Cameron, C. D., D'Cruz, J., & Bloom, P. (2024). In praise of empathic AI. Trends in Cognitive Sciences, 28(2), 89–91. https://doi.org/10.1016/j.tics.2023.12.003

Kleinman, Z. (2025, August 20). Microsoft boss troubled by rise in reports of 'AI psychosis'. BBC News. https://www.bbc.com/news/articles/c24zdel5j18o

Lam, T. C. M., Kolomitro, K., & Alamparambil, F. C. (2011). Empathy training: Methods, evaluation practices, and validity. Journal of Multidisciplinary Evaluation, 7(16), 162–200.

Landes, E., Voinea, C., & Uszkai, R. (2025). Rage against the authority machines: How to design artificial moral advisors for moral enhancement. AI & SOCIETY, 40(4), 2237–2248. https://doi.org/10.1007/s00146-024-02135-3

Larti, N., Ashouri, E., & Aarabi, A. (2018). The effects of an empathy role-playing program for operating room nursing students in Iran. Journal of Educational Evaluation for Health Professions, 15. https://doi.org/10.3352/jeehp.2018.15.29

Lee, H.-P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. CHI Conference on Human Factors in Computing Systems. CHI 2025.

Lee, E. H., Yin, Y., & Wakslak, C. J. (in press). Machines that care: On receiving and providing AI-driven empathy. In A. Perry & C. D. Cameron (Eds.), Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Considerations. Cambridge University Press.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mueller, K., Prins, R., & de Heer, H. (2018). An Online Intervention Increases Empathy, Resilience, and Work Engagement Among Physical Therapy Students. Journal of Allied Health, 47(3), 196–203. https://www.jstor.org/stable/48722085

Myers, S., & Everett, J. A. C. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. Cognition, 256, 106028. https://doi.org/10.1016/j.cognition.2024.106028

Østergaard, S. D. (2023). Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis? Schizophrenia Bulletin, 49(6), 1418–1419. https://doi.org/10.1093/schbul/sbad128

Perry, A. (2023). AI will never convey the essence of human empathy. Nature Human Behaviour, 7(11), 1808–1809. https://doi.org/10.1038/s41562-023-01675-w

Pugh, A. J. (2024). The Last Human Job: The Work of Connecting in a Disconnected World. Princeton University Press.

Sambasivan, N., & Veeraraghavan, R. (2022). The deskilling of domain expertise in AI development. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–14.

Shapiro, J., Rucker, L., Boker, J., & Lie, D. (2006). Point-of-view writing: A method for increasing medical students' empathy, identification and expression of emotion, and insight. Education for Health: Change in Learning & Practice, 19(1), 96–105. https://doi.org/10.1080/13576280500534776

Shastry, M., Fernandes, S., & Gray, K. (in press). Empathic AI will undermine human kindness. In A. Perry & C. D. Cameron (Eds.), Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Considerations. Cambridge University Press.

Shteynberg, G., Halpern, J., Sadovnik, A., Garthoff, J., Perry, A., Hay, J., Montemayor, C., Olson, M. A., Hulsey, T. L., & Fairweather, A. (2024). Does it matter if empathic AI has no empathy? Nature Machine Intelligence, 6(5), 496–497. https://doi.org/10.1038/s42256-024-00841-7

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Teding van Berkhout, E., & Malouff, J. M. (2016). The efficacy of empathy training: A meta-analysis of randomized controlled trials. Journal of Counseling Psychology, 63(1), 32–41. https://doi.org/10.1037/cou0000093

Vallor, S. (2024). The AI mirror: How to reclaim our humanity in an age of machine thinking. Oxford University Press. https://doi.org/10.1093/oso/9780197759066.001.0001

Vanman, E. J. (2016). The role of empathy in intergroup relations. Current Opinion in Psychology, 11, 59–63. https://doi.org/10.1016/j.copsyc.2016.06.007

Wenger, J. D., Cameron, D., & Inzlicht, M. (2025). The AI Empathy Choice Paradox: People Prefer Human Empathy Despite Rating AI Empathy Higher. OSF. https://osf.io/ghw2v_v1

Wu, X., Yao, S.-C., Lu, X.-J., Zhou, Y.-Q., Kong, Y.-Z., & Hu, L. (2024). Categories of training to improve empathy: A systematic review and meta-analysis. Psychological Bulletin, 150(10), 1237–1260. https://doi.org/10.1037/bul0000453