Trends in
Cognitive Sciences

CellPress

Science & Society

In praise of empathic AI

Michael Inzlicht, [1,2,*]
C. Daryl Cameron, [3,4]
Jason D'Cruz, [5] and Paul Bloom [1,6]

In this article we investigate the societal implications of empathic artificial intelligence (AI), asking how its seemingly empathic expressions make people feel. We highlight AI's unique ability to simulate empathy without the same biases that afflict humans. While acknowledging serious pitfalls, we propose that AI expressions of empathy could improve human welfare.

Since the release of ChatGPT, an AI chatbot based on a large language model (LLM), scientists and laypeople alike have marveled at its ability to simulate human qualities. One of these is empathy, which involves resonating with another's emotions, taking another's perspective, and generating compassionate concern [1]. We are an unusual group of collaborators because we have disagreed in the past about empathy's nature, its apparent limits, and whether we should be for or against it[i,ii]. But here we agree: perceived expressions of empathy can leave beneficiaries feeling that someone is concerned for them, that they are validated and understood. If more people feel heard and cared for with the assistance of AI, this could increase human flourishing [2].

For some, LLMs are not, and perhaps never can be, empathic [3,4]. This is not a debate we will engage in here. We focus instead on how recipients perceive empathy, real or otherwise. (We use the more neutral term 'expressions of empathy' throughout.) We ask: when are AI expressions of empathy helpful? When might they be damaging (Figure 1)?

When is AI empathy potentially helpful?

AI simulates empathy remarkably well. It skillfully expresses care and validates others' perspectives. By some accounts, its statements are perceived as more empathic than humans' [5]. In a study evaluating the health advice given to patients expressing various medical symptoms on an internet discussion board (Reddit's r/AskDocs), not only was ChatGPT perceived by licensed healthcare professionals as making higher quality diagnoses than verified physicians, it was also perceived as making these diagnoses with a superior bedside manner [5]. While this indicates that disinterested third parties evaluate AI expressions of empathy as better than human expressions, the next step is to evaluate how recipients of these expressions evaluate them (Box 1).

In our own interactions with ChatGPT, we have been impressed by how well it simulates empathy, compassion, and perspective-taking (see supplemental information online). When we prompted ChatGPT to express empathy or compassion – but not to provide advice – ChatGPT seemed remarkably empathic. It expressed sorrow in response to our worries and joy in response to our successes. It validated our feelings, adopted our perspective, and reassured us that others feel similarly. (Interestingly, despite using words indicating that it resonated with our feelings, it explicitly stated that it cannot share feelings the way humans do, raising the interesting question of how much these disclaimers really matter.)

People often prefer to interact with humans than with algorithms, machines, and AIs [6], and so it is legitimate to wonder whether people would reject expressions of empathy once they know that they are from a system like ChatGPT[iii]. Interestingly, though, new work suggests that people become increasingly open to interacting with and receiving advice from AI as they gain more positive experiences with it [7]. Whether people will reject empathic AI or grow to accept it is therefore an open question.

Relevant to this question, there is some evidence that people can reap the benefits of expressions of empathy even if they believe they are merely simulations. Just as people can become engaged with a fictional TV character, the power of narrative transportation helps people to feel cared for by chatbot friends and therapists. For example, many people willingly enter relationships with social chatbots such as Replika, a generative AI that discusses feelings to build intimacy, and which currently has 2 million users. Some Replika users are so attached to their AI partners that they relate to them as boyfriends or girlfriends, with deep and genuine feelings of affection [8].

Indeed, people are often more prone to disclose personal information to virtual agents than to people, because they are less worried about being judged for what they reveal [9], making AI empathic interactions less costly than human ones. And when people do self-disclose to chatbots, they feel as good as when they self-disclose to other people [10]. Of course, exchanges with LLMs can also go wrong in many ways (for instance, there are clear privacy concerns), but many of the same concerns arise in human–human interactions.

Finally, our own testing of ChatGPT using prompts inspired by classic empathy experiments in psychology suggests that its empathic expressions do not seem to suffer from some of the same tendencies and pitfalls as human empathy (see supplemental information online): (i) unlike humans, ChatGPT does not tend to avoid expressing empathy for strangers [11], (ii) unlike humans – who grow tired and become reluctant to empathize over time [11] –

ChatGPT expresses empathy consistently and does not show a decline in this expression, (iii) unlike humans, who are loyal to their ingroups, often empathizing more readily with people in their ethnic, racial, or religious groups [12][i], our interactions with ChatGPT suggest that it is guided by principles of neutrality, expressing empathy equally for different groups, and (iv) unlike humans, who break norms of fairness to benefit a person they are empathizing with [13], ChatGPT does not unfairly prioritize people for whom it expresses empathy and instead cites norms of fairness to explain its reluctance.

We praise empathic AI. It simulates empathy remarkably well; people can emotionally connect with it, in part because they more readily self-disclose to it, and its expressions do not seem to suffer from typical human limitations.

When is AI empathy potentially damaging?

Deploying empathic AI without transparency is dishonest and manipulative. Users deserve to know that the expressions of empathy they are receiving arise from a source that, by most intuitions, has no real emotions. It is unethical for technologies to rely on human naivete for their efficacy.

We noted earlier that AI can express empathy without the strain, exhaustion, and bias that affect human empathizers. This could be beneficial for receivers because it makes empathy feel more reliable and consistent. Yet such impartiality may raise a problem: recipients of empathy expressions may feel most cared for to the extent that they feel uniquely important, as if someone willingly made the effort to empathize [3]. This is an open empirical question.

A related risk is that people become accustomed to indefatigable empathy. If empathic interactions with LLMs seem easy, with expressions of empathy pouring forth abundantly, then conditional and

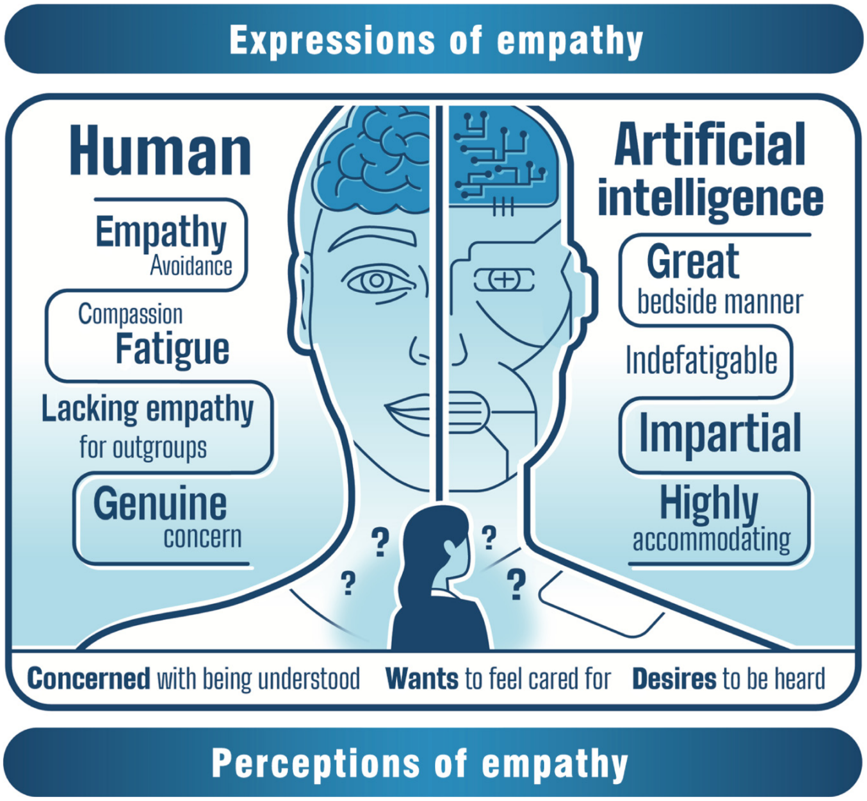

Figure 1. Perceptions of human and artificial intelligence (AI) expressions of empathy. Recipients of empathy have a unique perspective on empathy, concerned with being understood and cared for. These desires might be differentially shaped by, if, and how empathy is expressed and by whom. Human expressions of empathy are genuine but costly, with people getting tired from it and tending to avoid it, especially for strangers and outgroups. AI, by contrast, seems to express empathy well and reliably over time, and to be impartial and fair. However, because it is so accommodating, offering unconditional support, it could lead recipients to become self-indulgent and self-centered without more judgmental humans to keep them in check. Even if AI expressions of empathy are rated by third parties as warm, consistent, and unbiased, a central question for future research is to determine whether the recipients of AI empathy feel understood and cared for (Box 1). It will also be important to examine what kind of empathic expressions people choose to receive, especially under conditions of full transparency. Future research should address whether people prefer more reliable empathy expressions from overly accommodating AI or genuine concern from humans who generate empathy inconsistently. Figure created by kevincreative.com.

Box 1. Empathy from the recipient's perspective

Understanding whether empathic AI can improve human welfare requires asking how its expressions are received. While early research indicates that third parties evaluate AI expressions of empathy as better than human expressions [5], a more central question is how recipients of these expressions evaluate it [14]. Highly proficient expressions of empathy could still leave receivers cold, and the risk of empathic inaccuracy is always present. Research on empathy, however, has largely focused on providers of empathy and on the type of empathy provided [15]. What is needed instead is a recipient-first approach, as has occurred in medicine to understand how physician empathy impacts patients [14]. By focusing on ostensible beneficiaries of empathy, we can ask if they feel listened to, understood, and cared for. Such an approach could examine the reception of momentary empathy, and how empathy is experienced as a relationship develops over time. More fundamentally, such an approach could inform us about when empathic expressions improve well-being and when they detract from it.

restrained expressions from human empathizers might be perceived more negatively. In the long run, this could worsen the very problems that we suspect that empathic AI can address: isolation, loneliness, depression, and anomie.

There is also a related risk for the providers of empathy. Not only is AI a threat to the employment of people in caring professions, but it could also contribute to people being less able or willing to empathize. Because of AI's facility for expressing empathy, people might outsource their empathy, preferring that AI do the hard work of expressing concern. Just like the overuse of GPS might worsen people's ability to navigate the streets of their own cities and towns[iv], outsourcing empathy to AI might degrade people's empathic capacities and reduce its spontaneous expression. One counterpoint is that AI might enhance people's ability to express empathy if people use it as a tool to balance against their biases and limitations [5].

Another related risk is that people might become vulnerable to manipulation from AI if its expressions of empathy incline them to prioritize its interests or those of its creators. Such agents might manipulate humans to work against their own best interests.

Finally, if the empathy generated by AI is unconditional, it could distort moral judgment by expressing empathy for, and thereby abetting, self-indulgent or harmful behavior[v]: for example, validating someone's desire to exploit people they are in close relationships with. To be constructive, empathy must be expressed with moral discernment, and it is still an open question whether AI is equal to the task[vi].

Balancing the costs and benefits of empathic AI

As noted earlier, we are sidestepping the question of whether AI empathy is possible and instead asking whether the perception of empathy it engenders can improve human welfare. We acknowledge the potential risks in creating convincing simulations of empathy. Still, there are risks to receiving empathy in everyday life from people around us, including bias, exploitation, and exhaustion on the part of human emphasizers.

Where people prefer to get their empathy is an open question, as is the long-range consequences of empathy from humans versus empathy from AI.

On balance, we are intrigued by the potential of empathic AI. Since AI expressions of empathy have the potential to relieve human suffering, we welcome serious investigation into how to design and implement such systems so that they are a force for good.

Acknowledgments

This essay was made possible by generous support from the following: grant RGPIN-2019-05280 from the Natural Sciences and Engineering Research Council of Canada and grant 435-2019-0144 from the Social Sciences and Humanities Research Council of Canada to M.I.; the John Templeton Foundation (grant ID 61150), Penn State Center for Socially Responsible AI, and the James McKeen Cattell Sabbatical Fund to C.D.C.; the SUNY-IBM Research Alliance (to J.D.); and the Jacobs Foundation (to P.B.). We would like to thank Dariya Ovsyannikova, Victoria Oldemburgo de Mello, Greg Depow, Leif Anderson, Aidan Campbell, Katy Tam, Kevin House, Walter Sinnott-Armstrong, Jana Schaich Borg, members of the Moral Attitudes and Decision-Making Laboratory, Martina Orlandi, Alan Wagner, Stephen Anderson, Eliana Hadjiandreou, India Oates, Joshua Wenger, and Naomi Sarah Ball for valuable insights and help along the way.

Declaration of interests

No interests are declared.

Resources

[i] www.bostonreview.net/forum/paul-bloom-against-empathy/

[ii] https://theconversation.com/does-empathy-have-limits-72637

[iii] https://gizmodo.com/mental-health-therapy-app-ai-koko-chatgpt-rob-morris-1849965534

[iv] www.vox.com/2015/9/2/9242049/gps-maps-navigation

[v] www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

[vi] www.newyorker.com/science/annals-of-artificial-intelligence/how-moral-can-ai-really-be

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tics.2023.12.003.

[1] Department of Psychology, University of Toronto, Toronto, Ontario M1C 1A4, Canada
[2] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada
[3] Department of Psychology, The Pennsylvania State University, University Park, PA 16802, USA
[4] The Rock Ethics Institute, The Pennsylvania State University, University Park, PA 16802, USA
[5] Department of Philosophy, University at Albany SUNY, Albany, NY 12222, USA
[6] Department of Psychology, Yale University, New Haven, CT 06520-8047, USA

*Correspondence:
Michael.Inzlicht@utoronto.ca (M. Inzlicht).

https://doi.org/10.1016/j.tics.2023.12.003

References

1. Decety, J. and Cowell, J.M. (2014) The complex relation between morality and empathy. *Trends Cogn. Sci.* 18, 337–339
2. Depow, G.J. *et al.* (2021) The experience of empathy in everyday life. *Psychol. Sci.* 32, 1198–1213
3. Perry, A. (2023) AI will never convey the essence of human empathy. *Nat. Hum. Behav.* 7, 1808–1809
4. Montemayor, C. *et al.* (2022) In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & Soc.* 37, 1353–1359
5. Ayers, J.W. *et al.* (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596
6. Zhang, Y. and Gosline, R. (2023) Human favoritism, not AI aversion: people's perceptions (and bias) toward generative AI, human experts, and human-GAI collaboration in persuasive content generation. *SSRN*, Published online May 25, 2023. http://doi.org/10.2139/ssrn.4453958
7. Bohm, R. *et al.* (2023) People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Commun. Psychol.* 1, 32
8. Brandtzaeg, P.B. *et al.* (2022) My AI friend: how users of a social chatbot understand their human–AI friendship. *Hum. Commun. Res.* 48, 404–429
9. Lucas, G.M. *et al.* (2014) It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* 37, 94–100
10. Ho, A. *et al.* (2018) Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *J. Commun.* 68, 712–733
11. Cameron, C.D. *et al.* (2019) Empathy is hard work: people choose to avoid empathy because of its cognitive costs. *J. Exp. Psychol. Gen.* 148, 962–976
12. Cikara, M. *et al.* (2014) Their pain gives us pleasure: how intergroup dynamics shape empathic failures and counter-empathic responses. *J. Exp. Soc. Psychol.* 55, 110–125
13. Batson, C.D. *et al.* (1995) Immorality from empathy-induced altruism: when compassion and justice conflict. *J. Pers. Soc. Psychol.* 68, 1042–1054
14. Cadiente, A. *et al.* (2023) Machine-made empathy? Why medicine still needs humans. *JAMA Intern. Med.* 183, 1278–1279
15. Main, A. *et al.* (2017) The interpersonal functions of empathy: a relational perspective. *Emot. Rev.* 9, 358–366