Computer says "No": The Case Against Empathetic Conversational AI

Alba Curry
School of Philosophy, Religion and History of Science
University of Leeds
a.a.cercascurry@leeds.ac.uk

Amanda Cercas Curry
MilaNLP
Bocconi University
amanda.cercas@unibocconi.it

Abstract

Emotions are an integral part of human cognition and they guide not only our understanding of the world but also our actions within it. As such, whether we soothe or flame an emotion is not inconsequential. Recent work in conversational AI has focused on responding empathetically to users, validating and soothing their emotions without a real basis. This AI-aided emotional regulation can have negative consequences for users and society, tending towards a one-noted happiness defined as only the absence of "negative" emotions. We argue that we must carefully consider whether and how to respond to users' emotions.

I would gladly risk feeling bad at times, if it also meant that I could taste my dessert.

– Lt. Commander Data (Star Trek)

1 Introduction

Recent work in conversational AI has focused on generating empathetic responses to users' emotional states (Ide and Kawahara, 2022; Svikhnushina et al., 2022; Zhu et al., 2022) as a way to increase or maintain engagement and rapport with the user and to simulate intelligence. However, these empathetic responses are problematic on several fronts.

First, while a system might never claim to be human, responses simulating humanness prompt users to further behave as though the systems were (Reeves and Nass, 1996). Empathy, like all emotions, is likely a uniquely human trait[1] and systems that feign it are in effect feigning humanity. The ethical issues surrounding anthropomorphism have been discussed at length in the literature and are beyond the scope of this paper (Salles et al., 2020; Bryson).

Second, empathy requires an ability to both understand and share another's emotions. As such, responding empathetically assumes that the system is able to correctly *identify* the emotion, and that it is able to *feel* the emotion itself.[2] Neither one of these holds true for conversational AI (or in fact for any AI system). [3]

Even if conversational AI was to correctly identify the user's emotions, and perform empathy, we should ethically question the motives and outcomes behind such an enterprise. Svikhnushina et al. (2022) put forward a taxonomy of empathetic questions in social dialogues paying special attention to the role questions play in regulating the interlocutor's emotion. They argue for the crucial role effective question-asking plays in successful chatbots due to the fact that often questions are used to express "empathy" and attentiveness by the speaker. Here we highlight the ethical concerns that arise from questions which are characterised by their emotion-regulation functions targeted at the user's emotional state. What happens if the chatbot gets it right? There may be instances where a chatbot correctly identifies that a given situation is worthy of praise and amplifies the pride of the user and the result is morally unproblematic. For example, when (Svikhnushina et al., 2022) use the example of amplifying pride in the context of fishing. What happens if it gets it wrong? Depends on the type of mistake: a) If it fails to put into effect a question intent, then it may be inconsequential.[4]b) It amplifies

[1] We do not exclude the possibility that animals feel emotions such as empathy.

[2] Correctly identifying an emotion is problematic for animals including human beings. However, reasons differ between conversation AI and human beings: Human beings vary in their capacity to identify emotions in part because we struggle at times to identify our own or extend empathy to certain members of society, but we have the capability of identifying emotions. Furthermore, our ability to identify the emotions of others builds, at least in part, from our own emotions.

[3] Moreover, Barrett (2017) already problematised the identification of human emotions using language or facial expressions in general.

[4] In fact, if it is true that empathy would improve user engagement, the chatbot would simply fail to engage us.

or minimises an inappropriate emotion.[5] This is the problem we will focus on to argue that emotional regulation has no place in conversational AI and as such empathetic responses are deeply morally problematic. While humans will necessarily show empathy for one another, conversational AI cannot understand the emotion and so cannot make an accurate judgement as to its appropriateness. This lack of understanding is key as we cannot predict the consequences of assuaging or aggravating an emotion and a dialogue system cannot be held accountable for them.

2 The Crucial Roles of Emotions

What emotions are is still up for debate.[6] However, their importance for the individual and society has received renewed interest.[7]

Briefly, emotions play important roles: epistemic roles, and *conative* roles.[8] They perform at least three epistemic roles: (1) Emotions are ways of seeing the world; (2) they also signal to others how we see the world; (3) lastly, emotional interactions are invaluable sources of information for third party observers since they tell us what the members of the interaction value. For example, (1) when one grieves one signals to oneself and to anyone observing that one deems to have lost something of great value. It is conceivable that one was unaware up to that point that one valued what one lost–this is captured by the saying "you don't know what you have till it's gone." Furthermore, (2) your friends and family may learn something about you by observing your grief. They too may not have known how much something meant for you. Finally, (3) an observer may also learn about the dynamics of grief (whether it is appropriate to express it for example) by observing whether your family validated or not your grief.

Emotions play *conative* roles, meaning that they are involved in important ways with our motivation and desire to act in certain ways. In other words, not only do some emotions compel you to act, motivate you to act, but also how you act is coloured by the emotion you are experiencing. For example, your anger signals that you perceive that an injustice has occurred. If your boss fails to promote the person who deserves it because of their gender, your anger would motivate you to write a letter of complaint or speak to HR about it.[9]

Importantly, all emotions, including the so-called "negative" emotions (e.g., anger, contempt, hatred, shame, envy, guilt, etc) that we have just used as examples also share these functions. These emotions are not negative in the sense of being "bad," they are called negative because they tend to be accompanied by pain, and therefore they are emotions that, all things being equal, we would tend to avoid for ourselves. A world without injustice, and hence without anger, would certainly be ideal. However, we would not want a world of injustice where we are unequipped to notice the injustice or be motivated to do anything about it. Hence why it is imperative that we ask ourselves under which circumstances we ought to enhance or soothe emotions.

3 The Problem with Empathy

Literature discussing the value and power of empathy for conversational AI understand empathy as a tool to establish the common ground for meaningful communication and to appear more likeable to users. They understand empathy broadly as "the feeling by which one understands and shares another person's experiences and emotions" (De Carolis et al., 2017). Empathy facilitates engagement through the development of social relationships, affection and familiarity. Furthermore, for Svikhnushina et al. (2022) empathy is required in order to enable chatbots to ask questions with emotion regulation intents. For example, questions may be used to amplify the user's feeling of pride or de-escalate the user's anger, agitation, or frustration.

Empathy, although a common phenomenon, is not a simple one. It enjoys a long history in various scholarly disciplines. Indeed, a lot of ink has been spilled (and still is), for example, over how to make sense of character engagement. How do we, hu-

[5] One may argue that this criticism also applies to human beings, and it does. However, it would be a fallacious argument to insist that just because X does it, it is permissible for Y to do it. For what we mean by "inappropriate emotion" see section 4.

[6] A debate that enjoys a long history and that has been taken up by different disciplines. For an overview of this debate in philosophy see https://plato.stanford.edu/entries/emotion/. For an overview of the debate in science see Barrett (2017).

[7] See, for example, Bell (2013); Cherry (2021); Greenspan (1995)

[8] For a more detailed discussion see Curry (2022). Emotions have more roles, but for the purposes of this paper these are the two that are most useful to emphasise.

[9] There are good reasons to be sceptical of the claim that we can do this as a result of pure reason, see for example Brady (2013).

man beings, care for fictional characters? How are we intrigued and moved by their adventures and respond to the emotions and affects expressed in their voices, bodies, and faces as well as imagine the situation they are in and wish them success, closure, or punishment? Empathy is taken to be a key element and yet the exact nature of how human beings are able to experience empathy for fictional characters is currently being debated (Tobón, 2019).

The reason for highlighting this diversity is that conversational AI would do well to engage seriously with the rich history of empathy since the definition it tends to engage with lack the level of complexity required. Leaving aside the fact that defining empathy as the "reactions of one individual to the observed experiences of another" (De Carolis et al., 2017) tells us very little about the process by which a human beings may do this, let alone conversational AI, what we take issue with is what chatbots hope to do with that empathy. In other words, if for the sake or argument, we presume that conversation AI are able to accurately identify our emotions, the issue of how we deploy empathy is of huge ethical relevance.

Bloom (2017) argues against empathy and for what he calls rational compassion. He contends that empathy is one of the leading motivators of inequality and immorality in society. Thus, far from helping us to improve the lives of others, empathy is a capricious and irrational emotion that appeals to our narrow prejudices. It muddles our judgement and, ironically, often leads to cruelty. Instead we ought to not rely on empathy, but to draw instead upon a more distanced compassion. See also Prinz (2011).

There are two lessons we can take from this: (1) Given that empathy is used not just know what brings people pleasure, but also what brings pain, we might want to question the general future uses of empathy in conversational AI; (2) if we buy Bloom's argument then conversational AI should consider not imitating human beings, but becoming agents of rational compassion.

Breithaupt (2019) also takes issue with empathy arguing that we commit atrocities not out of a failure of empathy, but rather as a direct consequence of successful, even overly successful, empathy. He starts the book by reminding us that "[e]xtreme acts of cruelty require a high level of empathy."

The further lesson we can take away is while we assume that empathy leads to morally correct behaviour, and certainly there are many positive sides of empathy, we should not use an overly simple or glorified image of empathy.

Our problem is not necessarily with empathy per se, but rather with the explicit functions conversational AI has hopes to achieve with it, namely to enhance engagement, to inflate emotions deemed positive, and to soothe emotions deemed negative (e.g., Svikhnushina et al., 2022). Our claim is that we ought to think carefully about the consequences of soothing negative emotions only because they inflict pain on the user. Not only is this approach based on a naive understanding of emotions, it fails to recognise the importance of human beings being allowed to experience and express the full spectrum of emotions. One ought to not experience negative emotions because there is nothing to be upset about, not because we have devised a an emotional pacifier. In other words, the issue is that conversational AI lacks a sound value system for deciding why certain emotions are validated and others soothed. Furthermore, this AI-aided emotional regulation can have negative consequences for users and society, tending towards a one-noted happiness defined as only the absence of "negative" emotions.

4 When Emotions Get Things Wrong

There are two illustrative problems with the kinds of decisions behind amplifying and de-escalating emotions. One is the problem of what the ideal character might be. When you talk to a friend they will decide whether to soothe or amplify your emotions based not just in the situation but also based on who they deem you to be. If they think you are someone who has a hard time standing up for yourself they will amplify your anger to encourage you to fight for yourself. If on the other hand they think you are someone who leans too much on arrogance they will de-escalate your sense of pride. Even if, all things being equal, your pride on that occasion was warranted. Hence, not only would a conversational AI require prior knowledge of the interlocutor in terms of her character, but furthermore would have to decide what are desirable character traits.

The second question regards what an ideal emotion in a particular situation might be. We may all find it easy to say that negative emotions such as anger often get things wrong and lead to undesirable outcomes. However, positive emotions such as joy, hope, or pride which we may intu-

itively wish to amplify can also get things wrong. We assess and criticise emotions along a number of distinct dimensions: Firstly, emotions may be criticised when they do not fit their targets. You may, for example, be open to criticism for feeling fear in the absence of danger. Unfitting emotions fail to correctly present the world. In the case of pride, would we want to amplify someone's pride if they either did not in fact achieve anything, or their achievement was not merited? For example, that their nephew did very well in maths when in fact we know their nephew cheated? Second, an emotion may be open to criticism when it is not based on good evidence or is unreasonable. Consider the person who suffers from hydrophobia: given that in the vast majority of situations water is not dangerous, this person's fear is both unreasonable and unfitting. But even fitting emotions may be unreasonable. One may, for example, be terrified of tsunamis because one believes that they cause genetic mutations. In this case, one's fear is fitting — tsunamis are very dangerous — yet the fear is unreasonable since it is not based on good reasons. Third, an emotion may be criticised because it isn't prudent to feel. We might warn someone not to show anger when interacting with a person with a gun since they might get themselves killed; anger in this case may be reasonable and fitting given the gunman's actions and yet imprudent. Finally, we may condemn emotions as morally non-valuable because of the unacceptable way in which they present their targets. One may, for example, argue that *schadenfreude* is morally objectionable because it presents the pain of another person as risible.

Positive emotions may be unfitting, unreasonable, imprudent, as well as morally condemnable. On the other hand, negative emotions may well be fitting, reasonable, prudent, as well as morally laudable. In other words, even if one is equipped with empathy there are crucial normative decisions involved in question intents aimed at emotional regulation.[10] Amplifying and de-escalating emotion inappropriately, as in the case of what is best for one according to one's character and situation, as well as amplifying and de-escalating the wrong emotions can have devastating moral outcomes.

[10]See the complex example in Silva's (2021) discussion on outlaw emotions

5 Empathy and Responsibility

Human beings, all things being equal, will inevitably experience empathy. A reasonable human being experiencing empathy for another is proof of the importance of someone else's emotional state - for better or for worse. This supports the idea that our emotions are important, as opposed to the notion that our emotions are detrimental to rationality and ought to be regulated. They tell us many things about our world.

Similarly to many NLP systems' understanding of language, the empathetic responses of conversational AI are only performative (Bender and Koller, 2020). Thus, they provide a false sense of validity or importance. What if someone is experiencing an unfitting, unreasonable, or morally reprehensible emotion? Should a chatbot still showcase empathy? We hope to have shown that such decisions are deeply morally problematic and complex.

Hence, another key problem is responsibility. A human agent may choose to express their empathy (even if they cannot choose feeling it) and they may choose to attempt to regulate someone else's emotions based on their knowledge of the situation and the speaker's character. If a human being wrongly regulates someone else's emotions, they will be morally responsible for the consequences. Who is morally responsible in the case of conversational AI agents? Who are they benefiting when they are not actually benefiting the human agent? This issue is further elaborated on by Véliz (2021).

6 Related Work

Our article sits at the intersection of emotion detection, response generation and safety in conversational AI. We keep this section brief as we cite relevant work throughout the article. Several works have already focused on the issue of giving AI systems sentience, such as Bryson. While this would make the systems truly empathetic, the authors generally agree that we have a duty not to create sentient machines.

Lahnala et al. (2022) problematise NLP's conceptualisation of empathy which, they argue, is poorly defined, leading to issues of data validity and missed opportunities for research. Instead, we argue that even a more specific definition of empathy presents ethical issues that cannot be overlooked or ignored.

Dinan et al. (2021) provide a framework to classify and detect safety issues in end-to-end conversa-

tional systems. In particular, they point out systems that respond inappropriately to offensive content and safety-critical issue such as medical and emergency situations. We wish to extend this to empathetic responses where the system takes the role of an 'impostor': empathetic responses require a system to pretend to understand the emotion.

7 Conclusion

In this position paper, we argued that emotional regulation has no place in conversational AI and as such empathetic responses are deeply morally problematic. While humans will necessarily show empathy for one another, conversational AI cannot understand the emotion and so cannot make an accurate judgement as to its reasonableness. This lack of understanding is key as we cannot predict the consequences of assuaging or aggravating an emotion and a dialogue system cannot be held accountable for them. We hope to encourage reflection from future researchers and to initiate a discussion of the issue, not only in this particular case but also more reflection when it comes to pursuing seemingly positive goals such as bringing disagreeing parties towards agreement.

References

Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Pan Macmillan.

Macalester Bell. 2013. *Hard feelings: The moral psychology of contempt*. Oxford University Press.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Paul Bloom. 2017. *Against empathy: The case for rational compassion*. Random House.

Michael S Brady. 2013. *Emotional insight: The epistemic role of emotional experience*. OUP Oxford.

Fritz Breithaupt. 2019. *The dark sides of empathy*. Cornell University Press.

Joanna J Bryson. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 8.

Myisha Cherry. 2021. *The case for rage: Why anger is essential to anti-racist struggle*. Oxford University Press.

Alba Curry. 2022. *An Apologia for Anger With Reference to Early China and Ancient Greece*. Ph.D. thesis, UC Riverside.

Berardina De Carolis, Stefano Ferilli, and Giuseppe Palestra. 2017. Simulating empathic behavior in a social assistive robot. *Multimedia Tools and Applications*, 76(4):5073–5094.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Patricia S Greenspan. 1995. *Practical guilt: Moral dilemmas, emotions, and social norms*. Oxford University Press on Demand.

Tatsuya Ide and Daisuke Kawahara. 2022. Building a dialogue corpus annotated with expressed and experienced emotions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 21–30, Dublin, Ireland. Association for Computational Linguistics.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. *arXiv preprint arXiv:2210.16604*.

Jesse Prinz. 2011. Against empathy. *The Southern Journal of Philosophy*, 49:214–233.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10:236605.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.

Laura Silva. 2021. The epistemic role of outlaw emotions. *Ergo*, 8(23).

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973, Dublin, Ireland. Association for Computational Linguistics.

Daniel Jerónimo Tobón. 2019. Empathy and sympathy: two contemporary models of character engagement. In *The Palgrave handbook of the philosophy of film and motion pictures*, pages 865–891. Springer.

Carissa Véliz. 2021. Moral zombies: why algorithms are not moral agents. *AI & SOCIETY*, 36(2):487–497.

Ling.Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307, Dublin, Ireland. Association for Computational Linguistics.