

# When Computers Join the Moral Conversation

By Elizabeth O'Neill (Eindhoven University of Technology)

# When Computers Join the Moral Conversation

By Elizabeth O'Neill (Eindhoven University of Technology)

*This essay was originally published in the [Stanford One Hundred Year Study on Artificial Intelligence \(AI100\) Early Career Essay Competition 2023 “Anthology.”](#) Every five years, AI100 produces a report on the state of AI and its impact on society. The 2023 Early Career Essay Competition invited essay submissions containing feedback on AI100’s past reports and ideas for topics that AI100 should address in future reports.*

This essay draws attention to an underappreciated phenomenon that urgently needs attention and further research, but which was not covered in the 2021 AI100 Study Panel Report. The phenomenon in question is computers “joining the moral conversation,” in the sense that chatbots based on large language models (LLMs) now readily and flexibly respond to and apply many moral terms, and they *appear* to perform a variety of conversational roles—producing outputs that look like moral assertions, moral commands and reprimands, expressions of moral sentiment, assent to norms, etc.

The suddenness of this phenomenon may explain why it has attracted little scholarly and public attention thus far. Over the past decades, there has been speculative discussion on “moral machines,” “machine ethics,” computers that can reason about morality, etc. (Wallach & Allen 2008; Anderson & Anderson 2011); there has also been speculation on the prospect of artificial moral advisors and artificial ethics assistants (Savulescu & Maslen 2015; Giubilini & Savulescu 2017; O’Neill et al. 2022). Yet the arrival of chatbots that appear to engage in moral discourse is not specifically a product of efforts to create AI systems that reason about morality. Instead, it is a somewhat surprising byproduct of the development of LLMs. Reflecting statistical patterns in human language use, LLMs contain learned, partial models of many human moral terms and types of moral communication. This has conferred an impressive capacity to imitate human communication about many different values and norms.

The result is that generative AI systems employing these kinds of models (e.g., OpenAI’s ChatGPT, Character.AI, Google’s Bard, or Meta’s LLaMa) are prone to using moral terms and generating outputs that look like moral communications. That is, LLM-based chatbots will tend to (appear to) participate in moral discourse, unless actions are taken to limit that tendency. As it

turns out, some companies have taken some such steps, as part of efforts to make their systems safer or less harmful, e.g., using reinforcement learning from human feedback (OpenAI 2023) or reinforcement learning from AI feedback (Bai et al. 2022). Thus far, though, the public knows little about what such efforts have been taken and have had little input into which such actions *should* be taken.

Direct, public research attention is needed on the question of generative AI systems joining the moral conversation—what capacities different LLM-based systems possess, how their dispositions diverge from human moral psychological dispositions, how humans respond to different forms of apparent moral communications from computers, and, ultimately, what roles computers should be permitted to play in conversations about values, norms, and moral questions.

Should we say, for example, that generative AI systems should not purport to hold values, make moral judgments, approve of moral norms, express moral sentiments, or anything similar? Is it better for the systems to stick to descriptive claims, such as claims about what most humans think or how much disagreement exists on a given moral question? Should the systems venture metaethical claims, like “There is no right or wrong answer,” as some currently do?

Research on this topic is needed because participation in the human moral conversation introduces unprecedented potential for computers to influence human norms and values. Such influence may occur via facilitation of change, whether subtle or dramatic, or it may occur via *hindrance* of changes that would have otherwise occurred. Humans’ normative views are influenced in many ways by the stated and inferred views of the people around them (Bicchieri 2006; Sunstein 2019; Chituc & Sinnott-Armstrong 2020). When computers appear to make moral assertions, endorse views, express moral sentiments, etc., they, too, are likely to affect human morality (regardless of whether humans perceive them as computers) (see e.g., Jackson & Williams 2018, 2019; Wen et al. 2021).

This topic might fall within the “broader challenge” of “Normativity” that is discussed in the 2021 AI100 report, but I think it is worth highlighting as a special problem. For instance, it might be worth characterizing as a particularly important potential application area for LLMs. The phenomenon of interest in this essay also relates to the topic of “Disinformation and Threat to Democracy,” but it has to do less with influencing people’s beliefs about the world than with influence on values and norms.

Thus far, there have been relatively few efforts to purposefully use LLMs for the purpose of advancing particular moral worldviews or changing norms. As their potential becomes more apparent, I expect that people will be tempted to harness LLM-based systems for the purpose of automated norm advancement and enforcement. (One relevant precedent is that some companies have already employed simpler bots, e.g. keyword-based systems or language classifiers trained to detect hate speech or abusive text, for the purpose of content moderation—e.g. reprimanding or banning the human who sent the message (Gorwa et al. 2020)). Given the diversity in human values, we must anticipate the wide range of worldviews that LLM-based systems may be adapted to promote, including misogynistic, racist, and fascist value systems.

It is also important to emphasize that the potential for AI systems promoting values and enforcing norms extends beyond textual exchange and the digital world. If a system can take images and video as input, and produce moral labels and evaluative claims as output, it can be harnessed to identify and punish norm violations, on the basis of recordings or real-time feeds of the non-digital world. Importantly, some properties and actions that some humans condemn as wrong, could conceivably be detected based on imagery or sound. For example, imagine an AI system monitoring a crowd and flagging individuals as *immodest* or *impious* because it classifies them as women who are not covering their hair. People may well come to think they can use AI systems to identify individuals performing *violent* actions or engaged in *disrespectful* actions (e.g., letting a national flag touch the ground), etc.

In sum, the potential for LLM-based systems to influence human norms and values, whether inadvertently or purposefully, has not yet been sufficiently recognized. Furthermore, the normative question of what roles LLM-based chatbots should play in moral conversations, given how humans are likely to interact with them, has scarcely been examined. We need both public discussion and interdisciplinary research into these questions.

## Citations

Anderson, M. & S.L. Anderson (eds.) (2011) *Machine Ethics*. Cambridge University Press.  
Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A. Goldie, A., Mirhoseini, A., McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, E., Kerr, J. et al., “Constitutional AI: Harmlessness from AI Feedback,” Dec. 2022. arXiv preprint arXiv:2212.08073

- Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge and New York: Cambridge University Press.
- Chituc, V. & Sinnott-Armstrong, W. (2020). Moral conformity and its philosophical lessons. *Philosophical Psychology*, 33(2), 262-282.
- Giubilini, A. & Savulescu, J. (2017). The Artificial Moral Advisor. The “Ideal Observer” Meets Artificial Intelligence. *Philosophy & Technology*, 1-20.
- Gorwa, R., Binns, R. & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945.
- Jackson, R.B. and Williams, T. 2018. Robot: Asker of questions and changer of norms. *Proceedings of ICRES*.
- Jackson, R.B. and Williams, T. 2019, March. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 401-410). IEEE.
- O'Neill, E., Klineciewicz, M. & Kemmer, M. (2022). Ethical Issues with Artificial Ethics Assistants. *The Oxford Handbook of Digital Ethics*, Carissa Veliz, ed.
- OpenAI. GPT-4 Model System Card. (March 14, 2023). <https://cdn.openai.com/papers/gpt-4-system-card.pdf>
- Savulescu, J. & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI?. In *Beyond Artificial Intelligence* (pp. 79-95). Springer International Publishing.
- Sunstein, C. R. (2019). *How change happens*. MIT Press.
- Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
- Wen, R., Kim, B., Phillips, E., Zhu, Q. & Williams, T., (2021), March. Comparing strategies for robot communication of role-grounded moral norms. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 323-327).



**Elizabeth O'Neill**

**Assistant Professor of Philosophy**  
**Eindhoven University of Technology**  
**DLI Postdoc (2018-2019)**  
**[e.r.h.oneill@tue.nl](mailto:e.r.h.oneill@tue.nl)**  
**[More Info >](#)**