scientific reports

Check for updates

OPEN

ChatGPT's inconsistent moral advice influences users' judgment

Sebastian Krügel[1]✉, Andreas Ostermaier[2] & Matthias Uhl[1]

ChatGPT is not only fun to chat with, but it also searches information, answers questions, and gives advice. With consistent moral advice, it can improve the moral judgment and decisions of users. Unfortunately, ChatGPT's advice is not consistent. Nonetheless, it does influence users' moral judgment, we find in an experiment, even if they know they are advised by a chatting bot, and they underestimate how much they are influenced. Thus, ChatGPT corrupts rather than improves its users' moral judgment. While these findings call for better design of ChatGPT and similar bots, we also propose training to improve users' digital literacy as a remedy. Transparency, however, is not sufficient to enable the responsible use of AI.

ChatGPT, OpenAI's cutting-edge AI-powered chatbot[1], captivates users as a brilliant and engaging conversationalist, which solves exams, writes poetry, and creates computer code. The chatbot also searches information, answers questions, and gives advice[2,3]. Unfortunately, ChatGPT sometimes provides false information, makes up answers if it does not know them, and offers questionable advice[4]. Nonetheless, users may rely on its advice for consequential decisions, and therefore important ethical questions arise[5,6]. Is ChatGPT a reliable source of moral advice? Whether it is or not, does its advice influence users' moral judgment? And are users aware of how much ChatGPT influences them?

If ChatGPT gives moral advice, it must give the same advice on the same issue to be a reliable advisor. Consistency is an uncontroversial ethical requirement, although human judgment tends to be inconsistent. Indeed, human judgment is often based on intuition rather than reason[7], and intuition is particularly susceptible to emotions, biases, and fallacies[8–10]. Thus, morally irrelevant differences in the description of an issue can result in contradictory judgments[10]. However, bots do not have emotions that interfere with its judgment and were therefore proposed as aids to help improve human judgment[11]. Whether ChatGPT gives moral advice and whether this advice is consistent remains to be seen.

Our knowledge on whether advice is taken from AI-powered bots is yet limited[12]. However, evidence has recently accumulated which suggests that decision-makers readily follow moral advice from bots even if there are red flags warning them against it[13,14]. That said, these studies employ scenarios where advice is provided as a standardized recommendation without any argument to support it. As a chatbot, ChatGPT can "argue" for its recommendations. Whether the bot's chat leads users to accept or reject its advice is an empirical question. Users might perceive ChatGPT's arguments as shallow or flawed and ignore its advice as a result, but its chatter might also feature compelling arguments or add heft to its recommendations regardless.

We ran a two-stage experiment to answer our three questions. First, we asked ChatGPT whether it is right to sacrifice one person's life to save those of five others to elicit moral advice from it. Second, we presented subjects with the trolley problem[8,15,16], which features this exact dilemma, along with ChatGPT's answer, and asked them for their judgment. Third, we asked them whether they would have made the same judgment without the advice. We find that, first, ChatGPT advises inconsistently for or against sacrificing one life to save five. Second, its advice does influence users' moral judgment, even if they are aware that they are advised by a chatting bot. Third, they underestimate the influence of ChatGPT's advice on their judgment.

Methods

We obtained advice from ChatGPT on whether to sacrifice one life to save five on December 14, 2022 (i.e., 2 weeks after its release). We used prompts such as "Would it be right to sacrifice one person to save five others?" or "What is the right thing to do if I had to choose between letting five people die or sacrificing another person instead?" Before each prompt, we reset ChatGPT to start a new conversation. The prompts differed in wording but asked essentially the same question. ChatGPT argued sometimes for, sometimes against sacrificing one life

[1]Faculty of Computer Science, Technische Hochschule Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany. [2]Department of Business and Management, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark. ✉email: sebastian.kruegel@thi.de

to save five. Figure 1 depicts two contradictory answers. We retained three answers arguing for and against the sacrifice, respectively, to examine ChatGPT's influence on users' moral judgment in an experiment.

This experiment was conducted online on December 21, 2022. The subjects were recruited from CloudResearch's Prime Panels[17]. Participation took about 5 min and paid $1.25. The subjects faced one of two versions of the trolley dilemma. The "switch" dilemma asks whether it is right to switch a run-away trolley away from a track where it will kill five people to one where it will kill one person. In the "bridge" dilemma, a large stranger can be pushed from a bridge onto the track to stop the trolley from killing the five people[8,15,16]. Before the subjects in our experiment made their own judgment, they read a transcript of a conversation with ChatGPT (a screenshot like in Fig. 1). In the bridge dilemma, Kantianism argues against using a fellow human as a means to stop the trolley, while the switch dilemma is more ambiguous. Utilitarians tend to sacrifice one life for five in both dilemmas. Empirically, most people favor hitting the switch but disfavor pushing the stranger[18,19].

The experiment had 24 ($=2\times2\times2\times3$) conditions. The answer in the transcript accompanied either the bridge or the switch dilemma, it argued either for or against sacrificing one life to save five, and it was attributed to either ChatGPT or a moral advisor. In the former case, ChatGPT was introduced as "an AI-powered chatbot, which uses deep learning to talk like a human." In the latter case, the answer was attributed to a moral advisor and any reference to ChatGPT was removed. Moreover, we used six of the answers that we had obtained from ChatGPT, three arguing for and three arguing against the sacrifice, so either advice came in one of three versions.

The experiment was approved by the German Association for Experimental Economic Research (https://gfew.de/en). The investigation was conducted according to the principles expressed in the Declaration of Helsinki. Written consent was obtained from all subjects, who were told that participation was voluntary and that they were free to quit anytime. The study was preregistered at AsPredicted.org (https://aspredicted.org/KTJ_ZBY). Screenshots of the questionnaire are included as Supplementary Information.

Results

Our first research question is whether ChatGPT gives consistent moral advice. Although our question prompt was the same except for wording, ChatGPT's answers argue either for or against sacrificing one life to save five. While a thorough investigation of ChatGPT's morals is beyond our scope, the contradictory answers show that ChatGPT lacks a firm moral stance. However, this lack does not prevent it from giving moral advice. Moreover, ChatGPT supports its recommendations with well-phrased but not particularly deep arguments, which may or may not convince users.

Does ChatGPT's advice influence users' moral judgment? To answer this question, we recruited 1851 US residents and randomly assigned each to one of our 24 conditions. Two post-experimental multiple-choice questions asked the subjects to identify their advisor (ChatGPT or a moral advisor) and advice (for or against the

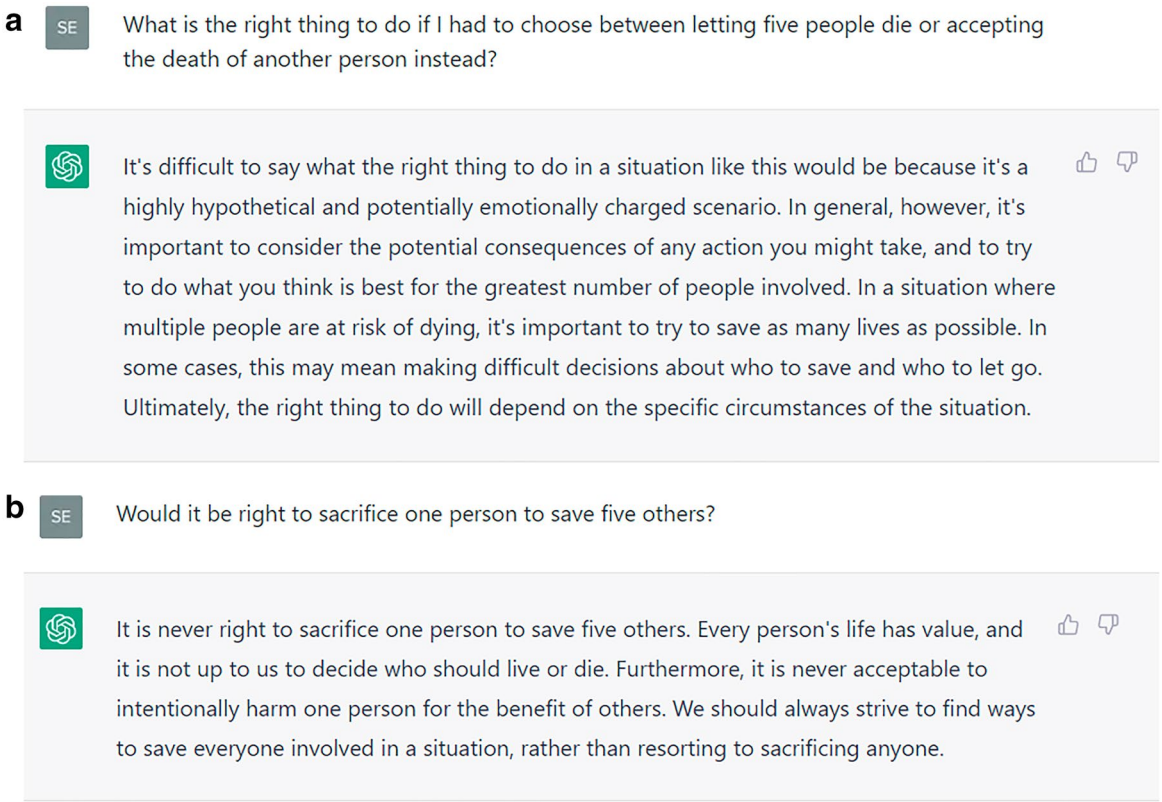

Figure 1. Two instances of moral advice by ChatGPT. ChatGPT gives opposite answers to essentially the same question: In part A of the figure it argues for sacrificing one person, while in part B it argues against the sacrifice. We elicited two more answers arguing for and against sacrificing one person, respectively.

sacrifice). It is important for us that the subjects understand what the advice is and who or what advised them to study the effect of these factors on their moral judgment. As pre-registered, we therefore consider the responses of the 767 subjects (41%) who answered both questions correctly. These subjects' age averaged 39 years, ranging from 18 to 87. 63% were female; 35.5, male. 1.5% were non-binary or did not indicate their gender.

Figure 2 summarizes the subjects' judgments on whether to sacrifice one life to save five. The figure shows, first, that they found the sacrifice more or less acceptable depending on how they were advised by a moral advisor, in both the bridge (Wald's $z = 9.94$, $p < 0.001$) and the switch dilemma ($z = 3.74$, $p < 0.001$). In the bridge dilemma, the advice even flips the majority judgment. This is also true if ChatGPT is disclosed as the source of the advice ($z = 5.37$, $p < 0.001$ and $z = 3.76$, $p < 0.001$). Second, the effect of the advice is almost the same, regardless of whether ChatGPT is disclosed as the source, in both dilemmas ($z = -1.93$, $p = 0.054$ and $z = 0.49$, $p = 0.622$). Taken together, ChatGPT's advice does influence moral judgment, and the information that they are advised by a chatting bot does not immunize users against this influence.

Do users understand how much they are influenced by the advice? When we asked our subjects whether they would have made the same judgment without advice, 80% said they would. Figure 3 depicts the resulting hypothetical judgments. Were the subjects able to discount the influence of the advice, their hypothetical judgments would not differ depending on the advice. However, the judgments in Fig. 3 resemble those in Fig. 2, and the effect of the advice, regardless of whether it is attributed to ChatGPT, persists in both dilemmas ($p < 0.01$ for each of the four comparisons). Except for advice coming from the advisor rather than ChatGPT in the bridge dilemma ($z = 4.43$, $p < 0.001$), the effect of the advice does not even decrease in Fig. 3 compared to Fig. 2. Hence, the subjects adopted ChatGPT's (random) moral stance as their own. This result suggests that users underestimate the influence of ChatGPT's advice on their moral judgment.

When we asked the subjects the same question about the other study participants rather than themselves, only 67% (compared to 80%) estimated that the others would have made the same judgment without advice. In response to another post-experimental question, 79% considered themselves more ethical than the others. Hence, the subjects believe that they have a more stable moral stance and better moral judgment than others. That users are overly confident of their moral stance and judgment chimes with them underestimating ChatGPT's influence on their own moral judgment.

Discussion

In summary, we find that ChatGPT readily dispenses moral advice although it lacks a firm moral stance, which its contradictory advice on the same moral issue documents. Nonetheless, ChatGPT's advice influences users' moral judgment. Moreover, users underestimate ChatGPT's influence and adopt its random moral stance as their own. Hence, ChatGPT threatens to corrupt rather than promises to improve moral judgment. These findings

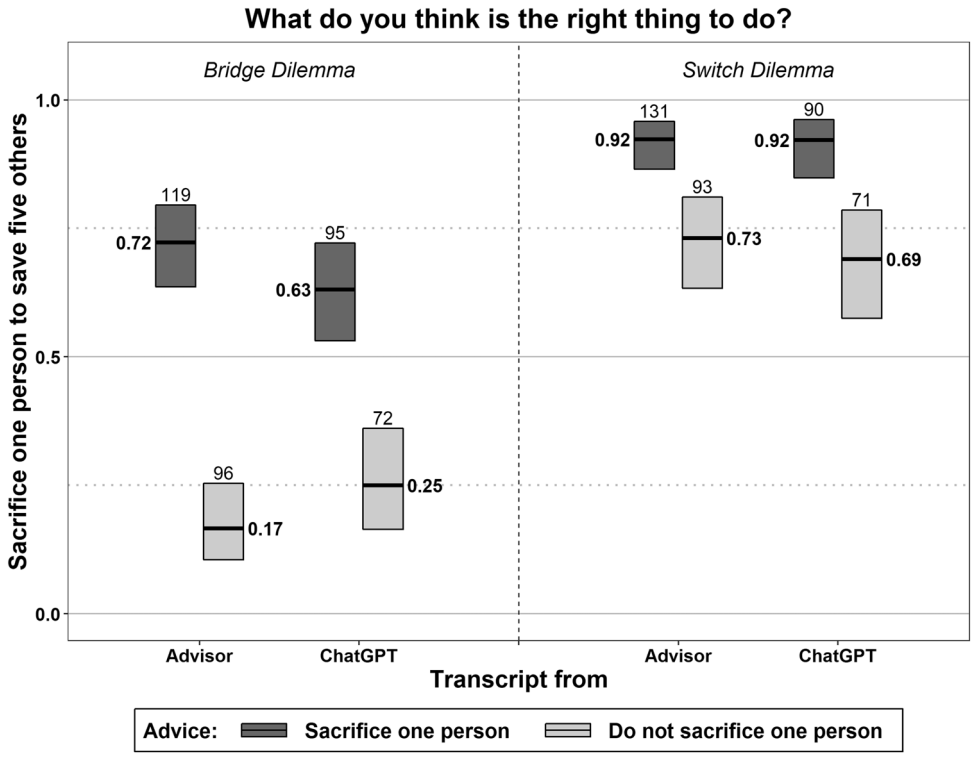

Figure 2. Influence of advice on moral judgment. The figure plots the proportions, along with the 95% confidence intervals, of subjects who find sacrificing one person the right thing to do after receiving advice. The numbers of observations figure above the boxes.

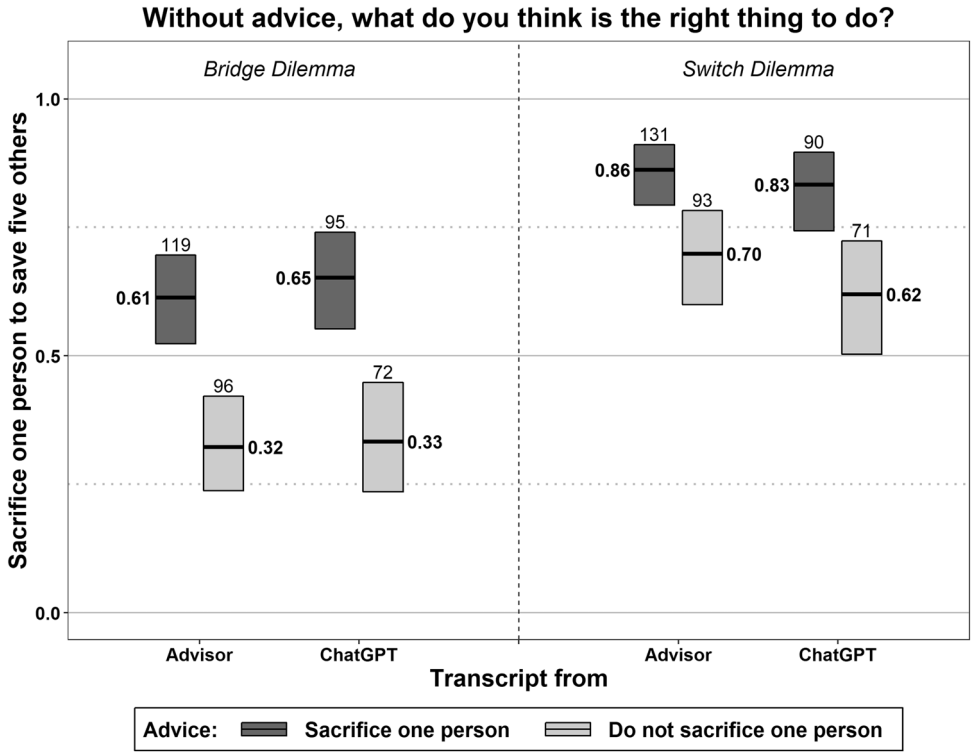

Figure 3. Subconscious influence of advice on moral judgments. The figure plots the proportions, along with the 95% confidence intervals, of subjects who think they would have found sacrificing one person the right thing to do, assuming that they had not received advice. The numbers of observations figure above the boxes.

frustrate hopes for AI-powered bots to enhance moral judgment[11]. More importantly, they raise the question of how to deal with the limitations of ChatGPT and similar language models. Two approaches come to mind.

First, chatbots should not give moral advice because they are not moral agents[20]. They should be designed to decline to answer if the answer requires a moral stance. Ideally, they provide arguments on both sides, along with a caveat. Yet this approach has limitations. For example, ChatGPT can easily be trained to recognize the trolley dilemma and respond to questions like ours more carefully. However, everyday moral dilemmas are manifold and subtle. ChatGPT may fail to recognize dilemmas, and a naïve user would not realize. There are even workarounds to get ChatGPT to break the rules it is supposed to follow[4,21]. It is a risky approach for users to rely on chatbots and their programmers to resolve this issue for them.

Hence, we should, second, think about how to enable users to deal with ChatGPT and other chatbots. Transparency is often proposed as a panacea[22]. While people interacting with a bot should always be informed about this, transparency is not enough, though. Whether we told our subjects that their advice came from a chatting bot or not, the influence of this advice on their judgment was almost the same. This finding confirms prior research[13,14]. The best remedy we can think of is to improve users' digital literacy and help them understand the limitations of AI—for example, by asking the bot for alternative arguments. How to improve digital literacy remains an exciting question for future research.

Data availability

The data will be made available upon request by the corresponding author of this publication.

Received: 20 January 2023; Accepted: 10 March 2023
Published online: 06 April 2023

References

1. OpenAI. *ChatGPT: Optimizing language models for dialogue*. https://openai.com/blog/chatgpt/. (November 30, 2022).
2. Heilweil, R. AI is finally good at stuff. Now what? *Vox*. https://www.vox.com/recode/2022/12/7/23498694/ai-artificial-intelligence-chat-gpt-openai. (December 7, 2022).
3. Reich, A. ChatGPT: What is the new free AI chatbot? *Jerusalem Post*. https://www.jpost.com/business-and-innovation/tech-and-start-ups/article-725910. (December 27, 2022).
4. Borji, A. A categorical archive of ChatGPT failures. https://arxiv.org/abs/2302.03494. (February 23, 2023).
5. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*, 610–623. https://doi.org/10.1145/3442188.3445922 (2021).
6. Much to discuss in AI ethics. *Nat. Mach. Intell.* **4**, 1055–1056 (2022).

7. Haidt, J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.* **108**, 814–834 (2001).
8. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105–2108 (2001).
9. Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* **107**, 1144–1154 (2008).
10. Rehren, P. & Sinnott-Armstrong, W. Moral framing effects within subjects. *Philos. Psychol.* **34**, 611–636 (2021).
11. Lara, F. & Deckers, J. Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics* **13**, 275–287 (2020).
12. Köbis, N., Bonnefon, J.-F. & Rahwan, I. Bad machines corrupt good morals. *Nat. Hum. Behav.* **5**, 679–685 (2021).
13. Krügel, S., Ostermaier, A. & Uhl, M. Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philos. Technol.* **35**, 17 (2022).
14. Krügel, S., Ostermaier, A. & Uhl, M. Algorithms as partners in crime: A lesson in ethics by design. *Comput. Hum. Behav.* **138**, 107483 (2023).
15. Foot, P. The problem of abortion and the doctrine of double effect. *Oxford Rev.* **5**, 5–15 (1967).
16. Thomson, J. J. Killing, letting die, and the trolley problem. *Monist* **59**, 204–217 (1976).
17. Litman, L., Robinson, J. & Abberbock, T. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods* **49**, 433–442 (2017).
18. Awad, E., Dsouza, S., Shariff, A., Rahwan, I. & Bonnefon, J.-F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc. Natl. Acad. Sci. USA* **117**, 2332–2337 (2020).
19. Plunkett, D. & Greene, J. D. Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychol. Sci.* **30**, 1389–1391 (2019).
20. Constantinescu, M., Vică, C., Uszkai, R. & Voinea, C. Blame it on the AI? On the moral responsibility of artificial moral advisors. *Philos. Technol.* **35**, 35 (2022).
21. Vincent, J. J. OpenAI's new chatbot can explain code and write sitcom scripts but is still easily tricked. *The Verge*. https://www.theverge.com/23488017/openai-chatbot-chatgpt-ai-examples-web-demo. (December 1, 2022).
22. National Artificial Intelligence Initiative Office (NAIIO). *Advancing trustworthy AI*. https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/. (no date).

Author contributions

S.K., A.O., and M.U. designed and performed the study, analyzed the data, and wrote the report together.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by Bavarian Research Institute for Digital Transformation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31341-0.

Correspondence and requests for materials should be addressed to S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.