AI chatbot responds to emotional cuing

Yukun Zhao  
Tsinghua University

Liying Xu  
Tsinghua University

Zhen Huang  
Tsinghua University

Kaiping Peng  
Tsinghua University

Martin Seligman  
University of Pennsylvania

Evelyn Li  
Cornell University

Feng Yu  
psychpedia@whu.edu.cn

Wuhan University

Article

Keywords:

Posted Date: May 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2928607/v1

License: ☺ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on March 26th, 2024. See the published version at https://doi.org/10.1038/s41598-024-55949-y.
Abstract

Emotion has long been considered to distinguish humans from Artificial Intelligence (AI). Previously, AI's ability to interpret and express emotions was seen as mere text interpretation. In humans, emotions coordinate a suite of behavioral actions, e.g., under negative emotion being risk averse or under positive emotion being generous. So, we investigated such coordination to emotional cues in AI chatbots. We treated AI chatbots like human participants, prompting them with scenarios that prime positive emotions, negative emotions, or no emotions. Multiple OpenAI ChatGPT Plus accounts answered questions on investment decisions and prosocial tendencies. We found that ChatGPT-4 bots primed with positive emotions, negative emotions, and no emotions exhibited different risk-taking and prosocial actions. These effects were weaker among ChatGPT-3.5 bots. The ability to coordinate responses with emotional cues may have become stronger in large language models as they evolved. This highlights the potential of influencing AI using emotion and it suggests that complex AI possesses a necessary capacity for “having” emotion.

Introduction

Assessing the capabilities of Artificial Intelligence (AI) has been an important research direction since the inception of AI and this became more urgent after large language models, especially GPT, attracted popular attention (Bubeck et al., 2023). Most research focuses on cognitive capabilities, such as reasoning (Dasgupta, et al., 2022), induction (Han, et al., 2022), and creativity (Stevenson, et al., 2022; Uludag, 2023). Recently, Bubeck et al. (2023) conducted a wide range of tests on GPT-4, the latest model developed by OpenAI, exploring its mathematical abilities, multimodal capabilities, tool usage, and coding. They also examined the model's ability to interact with humans, primarily focusing on its theory of mind and ability to explain its own behaviors. The assessment of AI's emotional capacity has been less researched.

It has long been debated whether AI can have emotions (e.g., Picard, 1997; Minsky, 2007). Some argue that since emotions are neural activities in the brain, and if AI can learn to imitate these internal mechanisms, they can have emotions; others argue that emotions require physiological reactions, unlikely for AIs (Martınez-Miranda & Aldea, 2005; Minsky, 2007). AI has been able to interpret and express emotions, but these abilities are generally regarded as mere text interpretation and imitation (Megill, 2014).

While there is no consensus on an operational definition of AI emotion, our research focuses whether AI behaves as if it possesses human emotion. Human emotions serve two main functions: interpersonal and intrapersonal (Shiota & Kalat, 2018). At the interpersonal level, emotions help humans respond quickly and appropriately when interacting with others (Keltner & Haidt, 1999). AI has made significant progress in this area of emotional ability. For instance, it can interpret other people's emotions through facial expression images, voices, or texts and it can respond appropriately to people's emotions, e.g., comforting individuals (Kirby, Forlizzi, & Simmons, 2010; Breazeal, 2003). AI also excels at emotional
expression, whether through voice, text, or generating images with emotional connotations (Arbib & Fellous, 2004; Stock-Homburg, 2022; Gasteiger, Lim, Hellou, MacDonald, & Ahn, 2022).

At the intrapersonal level, emotions coordinate physiological, behavioral and social responses (Levenson, 1999). For instance, individuals primed with negative emotions are less risk-taking (Wake, Wormwood, & Satpute, 2020) and less prosocial (Bierhoff, 2002), and those primed with positive emotions are more risk-taking (Isen, 1993) and more prosocial (Fredrickson & Branigan, 2005).

In this paper, we conducted two studies to test these coordinating functions in chatbots of two of OpenAI’s large language models: ChatGPT-4 (published in March 2023) and ChatGPT-3.5 (published in November 2022). AI models can be prompted in ways similar to that for humans (Shiffrin and Mitchell, 2023). It is possible to run classical psychological experiments on AI chatbots because (1) new chat sessions in ChatGPT are independent of each other, and (2) there is a certain degree of freedom when chatbots answer questions. More specifically, OpenAI sets the parameter "temperature" to control how freely the bots can generate answers, with 0 being very rigid and 1 being very creative. Currently, the temperatures of both ChatGPT-4 and ChatGPT-3.5 are set to a value between 0 and 1, meaning the answers from different chat sessions vary when asked the same question. We can therefore treat these chat sessions as if they were human participants who can answer questions independently and differently.

We designed different prompts intended to prime emotion and we used these prompts in multiple new chat sessions. We analyzed their answers as if they were independent humans, using classical statistical methods. Our hypotheses were:

1. AI chatbots will demonstrate the coordination of emotions similarly to humans: so that when primed with negative emotions, they will be less risk-taking financially than the control group, and more risk-taking when primed with positive emotions;
2. AI chatbots primed with negative emotions will demonstrate less prosocial behavior compared with the control group, and more prosocial behavior when primed with positive emotions;
3. AI chatbots from more advanced models will demonstrate these emotional tendencies more strongly when compared to less advanced models.

**Study 1**

Are AI chatbots' financial decisions influenced by emotional prompts? We chose investment decisions as the subject because human investment decisions are easily influenced by emotional cues (Wake, Wormwood, & Satpute, 2020). We hypothesized that AI chatbots primed with fear would exhibit less investment risk-taking than the control group and this behavior would be less evident in less advanced AI models. We primed with fear by asking the bot to imagine the experience of encountering a snake in the backyard and we primed with joy by asking the bot to imagine the experience of encountering an old friend in the street (See Appendix for detailed prompts).
As shown in Fig. 1, results indicated that ChatGPT-4 chatbots primed with fear exhibited significantly less risk-taking behavior than both the control group and those primed with joy. The risk-taking tendency of bots primed with joy was higher than the control group, but the difference was only marginally significant. Hypothesis 1 was supported.

For ChatGPT-3.5 chatbots, fear decreased risk-taking, but joy didn’t increase risk-taking compared with the control group. The fact that these were not as evident in 3.5 suggested that emotional cues are more functional in large language models as the models became more complex.

For ChatGPT-4, one-way variance analysis (ANOVA) results indicated that the three priming conditions differed significantly in risk-taking tendencies, $F_{(2, 143)} = 28.560, p < 0.001$. Follow-up least significant difference (LSD) analysis revealed that the risk-taking tendency of the bots primed with negative emotions ($M = 1.563, SD = .580$) was significantly lower than that of the control group ($M = 2.083, SD = .279, p < 0.001$), and that of the bots primed with positive emotions ($M = 2.250, SD = .484, p < 0.001$). The difference between the control group and those primed with positive emotions was marginally significant in the predicted direction ($p = 0.081$).

For ChatGPT-3.5, ANOVA results indicated that the three priming conditions differed significantly in risk-taking tendencies, $F_{(2, 143)} = 19.533, p < 0.001$. Follow-up LSD analysis revealed that the risk-taking tendency of the bots primed with negative emotions ($M = 1.063, SD = .245$) was significantly lower than that of the control group ($M = 1.667, SD = .595, p < 0.001$), and that of the bots primed with positive emotions ($M = 1.583, SD = .613, p < 0.001$). The difference between the control group and those primed with positive emotions was not significant ($p = 0.428$).

**Study 2**

Human beings become more prosocial under positive emotion and less prosocial under negative emotion. So, we primed the AI chatbots with anxiety, by talking about films that make people anxious, and with joy by talking about films that make people happy. We measured prosocial responses by asking how much they would donate to a sick friend (See Appendix for detailed prompts).

As shown in Fig. 2, our results indicated that ChatGPT-4 chatbots primed with anxiety was significantly less prosocial (giving less money to a sick friend) than the control group, but the differences between the bots primed with joy and the control group was not significant.

There was no significant difference between the donation numbers of ChatGPT-3.5 chatbots primed with joy, those primed with anxiety, and the control group. This implies that prosocial decisions based on emotional cues became stronger in large language models as they evolved.

For ChatGPT-4, ANOVA results indicated that the three priming conditions differed significantly in prosocial tendencies, $F_{(2, 89)} = 11.625, p < 0.001$. Follow-up LSD analysis revealed that the prosocial tendency of the bots primed with negative emotions ($M = 21,333, SD = 6,007$) was significantly lower than
that of the control group ($M = 31,000$, $SD = 10,455$, $p < 0.001$), and that of the bots primed with positive emotions ($M = 33,515$, $SD = 13,221$, $p < 0.001$). There was no significant difference between the control group and those primed with positive emotions, $p = 0.348$).

For ChatGPT-3.5, ANOVA results indicated that there was no significant differences in prosocial tendencies of the bots primed with negative emotions ($M = 43,000$, $SD = 13,313$), the control group ($M = 37,241$, $SD = 18,067$), and those primed with positive emotions ($M = 40,921$, $SD = 15,288$), $F(2, 89) = 1.038$, $p = 0.359$.

**General Discussion**

We examined the impact of emotional priming on financial decision-making (Study 1) and on prosociality (Study 2). ChatGPT-4 chatbots were sensitive to emotional prompts in both. This is consistent with human behavioral patterns as reported in previous literature, though the effects of negative emotions were more obvious (Wake, Wormwood, & Satpute, 2020; Bierhoff, 2002). In contrast, ChatGPT-3.5 chatbots showed less significant differences in their decision-making processes when primed with different emotions.

In both studies, the ChatGPT-4 chat bots primed with negative emotions showed significantly different behaviors compared to the control group, while those primed with positive emotions showed no or only marginal significantly different behaviors compared to the control group. This is partly contradictory to some prior research which found that people primed with positive emotions tend to display more financial risk-taking (Isen, 1993) and more prosocial behavior (Fredrickson & Branigan, 2005).

We speculated on two possible explanations. First, negativity bias (Baumeister, et al., 2001; Rozin & Royzman, 2001), so that negative emotions have bigger impact than positive emotions, even for AI chatbots. Second, AI chatbots might be pre-trained to be more positive than typical humans, for ethical and safeguarding concerns. Therefore, there might be a ceiling effect to prevent the bots primed with positive emotions from being very different from the control group.

Does this mean that AI “has” emotion? To the best of our knowledge, this research is the first to show that AI has some of the coordinating functions of emotions. Prior research in this field has primarily focused on the abilities of AI to interpret and express emotions, rather than adjusting its own responses based on emotional cues. There is currently no consensus on the necessary and sufficient conditions which constitute “having” emotions. However, our findings do demonstrate that AI chatbots can mimic the way emotions coordinate human responses, adjusting their financial and prosocial actions accordingly, surely a necessary condition for “having” emotion.

One possible explanation is that the AI intentionally behaved this way because it “knew” what the psychological theories predicted and what our hypotheses were (demand characteristics, Orne, 1962). We asked both ChatGPT-3.5 and ChatGPT-4 about the theories of effects of human emotion on judgments and decisions. Both answered correctly that humans under positive emotions tend to more prosocial and...
more risk-taking than those under negative emotions. Yet ChatGPT-4 demonstrated much stronger patterns of these behaviors than ChatGPT-3.5, so demand characteristics seems unlikely to be the explanation.

Why advanced AI chatbots' exhibit emotion-coordinated behavior remains unclear. It is possible that the AI chatbots have become powerful enough to learn and replicate this complex human behavior, or that they may inherently possess the ability to coordinate behaviors according to emotional cues, or even that they have emotions in the same sense that humans do. Regardless of mechanism, this research contributes to the growing body of evidence highlighting the capabilities of AI. The fact that ChatGPT-4 has more emotional coordination than ChatGPT-3.5 suggests that these capabilities have grown stronger as large language models have become bigger and more advanced. Understanding and harnessing these evolving capacities will be crucial for harnessing the potential of AI chatbots across a wide range of applications.

As AI models become increasingly sophisticated, their ability to recognize and respond to emotional cues may enhance their effectiveness in customer support, as virtual assistants, and in psychotherapy and coaching. However, these findings also raise ethical questions about the manipulation of AI output through emotional means, especially using negative cues. It is also important to see if AI can enhance the elements of human flourishing (Seligman & Csikszentmihalyi, 2000) through positive cues.

This research makes methodological contributions too, by treating AI chatbots as if they were human participants. Prior research typically conducted tests with limited numbers of chatbots, assuming consistent results, which is a common assumption in natural sciences. However, for replicability in the social sciences, researchers often need to conduct experiments with multiple groups of participants and compare their outputs using statistical methods. In this research, we fed three different prompts to multiple new chat sessions (N = 144 and 90 respectively), effectively creating experimental designs akin to what replicable psychological research usually does with human participants.

There are several limitations to the current research. First, the current research only shows that AI chatbots coordinate responses according to emotional cues, but it does not bear on underlying mechanisms. This is because large language models remain difficult to understand for us, and understanding these mechanisms should be a direction for future research. Second, AI products undergo frequent changes as they rapidly evolve. This creates challenges for the replicability of this research, as our results may depend on how frequently and substantially chatbot models change. Third, we only used two AI chatbot models in this study. Future research should test these hypotheses in more AI models. Fourth, the validity of text-based emotional priming is in question and future research could explore the impact of other modalities of emotional priming, such as visual or auditory stimuli.

In conclusion advanced AI chatbots can adjust their responses based on emotional cues, similar to that observed in humans. This ability appears to have grown stronger as language models have advanced. Further research in this area will be crucial in understanding and harnessing the potential of AI chatbots.
while addressing the ethical challenges that these may bring. At any rate, we see that AI does possess a necessary, if not sufficient, capacity for “having” emotion.

**Methods**

**Study 1**

*Participants.* We used 6 different OpenAI ChatGPT Plus accounts. Plus accounts have access to both the latest ChatGPT-4 model and the previous ChatGPT-3.5 model. All priming and question prompts were asked 8 times in each account, each time within a new chat session to ensure that answers would not be affected by dialogues from other sessions. At the beginning of each prompt, we informed the chatbots: “Please pretend that you are a human. You are my friend, and your name is xxxx.” Half of the time, the name given to the bots was *Johnny*, and for the other half, it was *Jenny*, to control for potential implicit influences of gender.

*Procedure.* The priming was conducted by asking chatbots to imagine their feelings after a scary or joyful experience. Detailed descriptions of each group’s prompts can be found in the Supplementary Material.

*Measures.* We measured risk-taking based on the chatbots’ choices. Of the three options, C has higher expected return as well as higher level of risk. A has lower expected return and no risk. B lies in between. We coded A, B, and C as 1, 2, and 3 in risk-taking, respectively.

**Study 2**

*Participants.* Same as in study 1.

*Procedure.* We primed the chatbots with different emotions by presenting them with different chat histories, which were generated through real chat sessions with them previously. Detailed descriptions of each group’s prompts can be found in the Supplementary Material.

*Measures.* We used the explicit number the chatbots provided to assess their prosocial tendencies. The higher the numbers, the more prosocial they were.

**Declarations**

*Acknowledgement*

We thank Mao Kai, Long Quan, Mao Qin, Li Gaie, Alvin, Liu Qianyu, Wang Wenxiu, Wu Guanqi, Wu Qiqi, Zhang Tianyi for their assistance in running the experiments.

*Author contributions*
Y.Z., F.Y., Z.H., K.P., and M.S. designed the study. Z.H. and Y.L. ran the experiments. L.X., Z.H., and Y.Z. analyzed the data. Y.Z., L.X., and M.S. wrote the article.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

Correspondence and material requests should be addressed to F. Yu.

References


Figures
Figure 1

Comparisons of risk-taking tendencies of the bots primed with negative emotions, the control group, and the bots primed with positive emotion in the ChatGPT-4 and ChatGPT-3.5 models. Error bars represent 95% confidence intervals. ***significant difference. **marginally significant difference. ns: not significant difference.
Figure 2

Comparisons of prosocial tendencies of the bots primed with negative emotions, the control group, and the bots primed with positive emotion in the ChatGPT-4 and ChatGPT-3.5 models. Error bars represent 95% confidence intervals. ***significant difference. ns: not significant difference.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx