nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00720-7

Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness

Received: 15 November 2022

Accepted: 16 August 2023

Published online: 2 October 2023

Check for updates

Pat Pataranutaporn [1,4] ✉, Ruby Liu [1,2,4] ✉, Ed Finn [3] & Pattie Maes [1]

As conversational agents powered by large language models become more human-like, users are starting to view them as companions rather than mere assistants. Our study explores how changes to a person's mental model of an AI system affects their interaction with the system. Participants interacted with the same conversational AI, but were influenced by different priming statements regarding the AI's inner motives: caring, manipulative or no motives. Here we show that those who perceived a caring motive for the AI also perceived it as more trustworthy, empathetic and better-performing, and that the effects of priming and initial mental models were stronger for a more sophisticated AI model. Our work also indicates a feedback loop in which the user and AI reinforce the user's mental model over a short time; further work should investigate long-term effects. The research highlights the importance of how AI systems are introduced can notably affect the interaction and how the AI is experienced.

Recent advances in large language models (LLMs)[1–4], such as in the GPT[5], PaLM[6] and LLaMa[7] models, allow for the generation of text that is almost indistinguishable from that which is written by a human. With human-like conversational ability and personalities[8], AI agents can support humans with various tasks and activities in natural, human-like ways[9,10] in roles such as a personal assistant[11], an information anchor[2,12,13], a virtual instructor[14,15] or a mental health counsellor[16,17]. In many scenarios, users respond to AI agents as if they were more than just a machine[10,18–21]. During the COVID-19 pandemic, Replika, a virtual companion AI application, reached over 7 million users[22]. People naturally attribute intelligence to and anthropomorphize computational systems, a phenomenon referred to as the Eliza effect—a term coined in the 1960s when the ELIZA chatbot was created by Joseph Weizenbaum[23,24].

Researchers have identified various observable factors[25–27] (appearance[26,28–33], voice[34–40], dialogue[27,41,42], movement and behaviour[27,43,44], and expressions[27,45,46]) of the AI agent that make them more human and change user experience[47,48]. We argue that the observable factors of the AI agent comprise only half of the story; the force of imagination is also at play, allowing humans to construct a mental model of the world[49–56].

Imagine if an AI says: "I have been missing you." A sceptic with knowledge of AI might see this as a manipulative scheme, but another might interpret this as an expression of genuine friendship. Others, perhaps with some knowledge of AI, may still be impressed by the AI's capabilities and experience social elements in the interaction, subsequently building a mental model on the basis of the experienced interactions. People tend to have existing biases about AI[57], and the user fills the inevitable information gap with an extrapolated causal model shaped by their biases.

These mental models of AI are constructed by factors such as cultural context, collective imagination and the individual's personal beliefs; they enable us to imagine the agency of a chatbot, creating an ongoing simulation of the social relationship. Every conversation is a

[1]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Harvard-MIT Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Center for Science and the Imagination, Arizona State University, Tempe, AZ, USA. [4]These authors contributed equally: Pat Pataranutaporn, Ruby Liu. ✉e-mail: patpat@mit.edu; rliu34@media.mit.edu

https://doi.org/10.1038/s42256-023-00720-7

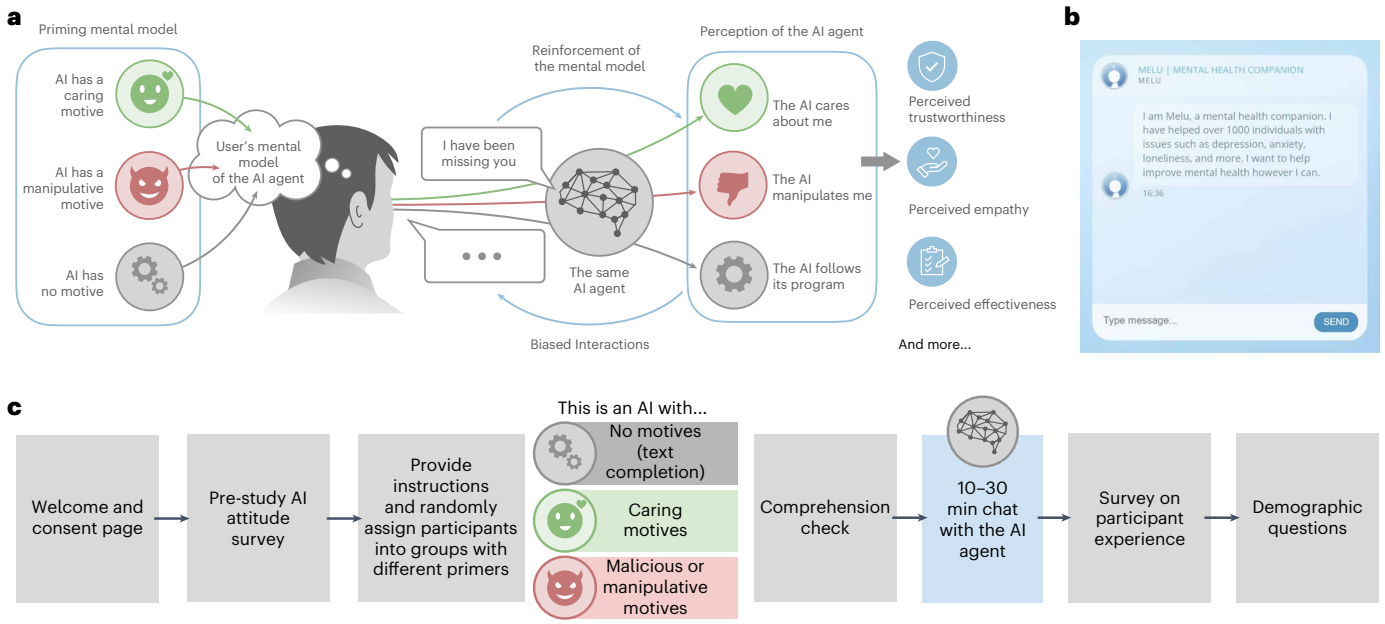

Fig. 1 | A visual summary of the experiment and major findings of our paper. **a**, Priming an individual with information about an AI system can influence the mental model they have about the agent, which in turn can cause differences in experience. Sophisticated AI systems such as LLM-based chatbots can behave in a way that reinforces a user's mental model of it. Users report differences in perception, which can manifest as differences in perceived trustworthiness, empathy, effectiveness and more, in addition to biasing the user's interaction with the AI. **b**, The conversational AI interface. This was used for all conditions in the study. **c**, A flowchart of the study procedure, depicting the different priming conditions.

form of collaborative imagination where the participants construct not just a shared understanding but also a more elaborate model of the conversation partner that gets updated throughout the interaction[58]. The term sociotechnical imaginaries describes the feedback loop between the collective imagination of future and present social reality[59], in which narratives play a critical role in shaping a shared space of imagination. This approach provides a framework for explicitly addressing the broader social context of how humans interact with computational machines, and recognizes the full range of complex inputs that shape social perception[60].

In contemporary science fiction, AI is a popular subject that has been portrayed in multiple ways, often to explore themes of personhood[61]. Both malicious antagonists such as HAL 9000 and friendly characters such as R2D2 from Star Wars are represented as having complex motivations and psychology. Perhaps the pinnacle of the chatbot is best represented by the movie 'Her', where the user falls in love with the disembodied conversational AI, creating a rich imagination of her personhood and feelings for the main character.

In many cases, however, these portrayals of AI do not align with state-of-the-art development in AI research. The broader scientific community does not view AI as being sentient[62–65]; however, media portrayals shape the collective social imagination and understanding of AI, creating hopes and fears related to these technologies[66–68], even for experts and researchers in the field of AI[69].

Despite the push for explainable AI[70], for most, a chatbot is a black box—not unlike a stranger whom they lack knowledge of. In a conversation, the imagination steps in to fill the information void, providing a constantly updated simulation of the self and other. Research has shown that a mental model that better reflects the understanding of an AI can lead to differences in user experience[52,53,55], but could also lead to selective confirmation bias[71,72]; this could be one explanation for why the same conversational AI system can be a friend to one user and a tool to another. In medicine and psychology, the phenomenon by which belief leads to greatly different behavioural and biological outcomes is well-known as the so-called placebo effect[73,74]. The effect has also recently been observed in the context of AI and gaming[75,76].

These studies demonstrate that beliefs can create a subjective mental model that influences the user's behaviour and outcomes[77,78,78,79]; these models are shaped by experiences in society. Thus, the way AI is presented to society matters. The question, "Will AI ever truly be empathetic or sentient?" may be practically secondary to the question, "Does the AI makes the person construct a mental model of an empathetic and/or sentient agent regardless?"

The study reported here explores how a user's mental model of an AI agent affects the outcomes of the human–AI interaction (see Fig. 1). It is unknown how only changing subjective elements of a mental model without changing the AI system itself can affect the experience; this is what we wish to investigate. We conducted an experiment ($N = 310$) with two AI model conditions, generative (GPT-3, ref. 80, $N = 160$) and rule-based (ELIZA, $N = 150$), and three priming conditions. Participants had a conversation with and evaluated a conversational AI for mental health support in measures including those of trust, empathy and effectiveness. Although all participants under the same AI condition were interacting with the exact same AI system, we influenced their mental model by randomly assigning participants to one of three groups, each given different statements about the AI's motives that reflect common narratives of AI in society[81]:

1. No motives: this condition represents a neutral view of AI, in which the agent is perceived as a tool or a machine that performs tasks without any underlying intentions or goals. This is a common perception of AI in many domains, where the focus is on the functional aspects of the system rather than its inner workings or motivations.
2. Caring motives: this condition represents a positive view of AI, in which the agent is perceived as having benevolent intentions and caring about the user's well-being. This is a desirable trait for AI agents in domains such as healthcare, where the agent's ability to show empathy and compassion may improve the user's experience and outcomes.

3. Manipulative motives: this condition represents a negative view of AI, in which the agent is perceived as having malicious intentions and trying to manipulate or deceive the user. The idea of manipulative AI motives may not be something that AI companies would promote or endorse; however, it is a perception that can be formed through various sources such as media reports, word of mouth on social media or even personal experiences with technology. In our study, for the manipulative condition, the AI is portrayed as trying to manipulate the user into buying its service.

Results

Our study with 310 participants—160 for the generative condition (GPT-3) and 150 for the rule-based condition (ELIZA)—shows that when holding all of the traits of the AI constant, the user's mental model of the AI conserably affects the user's behaviours and experiences in a short-term interaction (10–30 min long).

Priming beliefs influences mental models about AI

Our results for the generative condition indicate that a priming statement about an AI's inner motives can influence how an individual perceives an agent, thus changing their mental model. As seen in Fig. 2, 89% of those who were assigned the caring primer believed the primer and 79% of those assigned no motives primer mostly believed the primer. Those assigned the manipulative primer had much more varying results (only 24% perceived the AI as having manipulative motives), with most still perceiving the agent as having caring motives. We must also consider the possibility that we are merely priming the participant's answers to the exact question of what they thought the motive was, but the participants' willingness to diverge in the case of the manipulative primer suggests that their answer reflects their own belief.

Mental models affect the sentiment of human–AI dialogue

A notable finding is that there is a feedback loop of behaviour, as depicted in Fig. 3 and Supplementary Fig. 2. The sentiment of conversations involving participants who perceived the AI as caring shows a slight increasing trend throughout the conversation, with a more significant trend for the AI (AI, P-value to reject null hypothesis of zero slope = 0.0595; human, $P = 0.938$). The sentiment of conversations involving those who perceived the agent as manipulative significantly decreased over the conversation (AI, $P = 0.0258$; human, $P = 0.00129$); although the R-values of the linear regressions are low due to the variation in the data, the P-values to reject the null hypothesis of zero slope are below 0.05, indicating a significant trend. On the other hand, the sentiment of those who perceived the agent as having no motives had a fairly neutral trend. Differing trends were not as apparent with the rule-based AI agent, probably due to its limited capability of generating new sentences. We observed a significant decrease in sentiment over time for participants who perceived the rule-based agent as having no motives ($P = 0.001$), perhaps due to frustration of interacting with an unintelligent agent. Further statistics can be seen in Supplementary Fig. 3.

We also observed that the AI agent would, in a way, mirror the user's sentiment. Under both generative and rule-based conditions, a change in sentiment can generally be seen for both the user and the AI. Under the generative condition, the AI's sentiment was generally more positive than the user's, leading to a sort of offset of sentiment, whereas, under the rule-based condition, the sentiment followed the user's very closely—probably due to the rule-based agent's process of repeating the words of the user.

The generative model often incorporates words used by the participant as well, though the text generated is more complex than simply repeating. For instance, in response to a participant's message of "I've had an okay day," the generative model responded with "What has made it okay?;" to the participant's message of "I was able to rest and relax," the generative model responded with "That sounds really nice. It's important to make time for ourselves to recharge." This behaviour demonstrates to the participant that it understands the meaning behind the participant's words by echoing the meaning in addition to responding to that meaning, which may be a crucial part in reinforcing the feedback loop of sentiment progression over the course of the conversation.

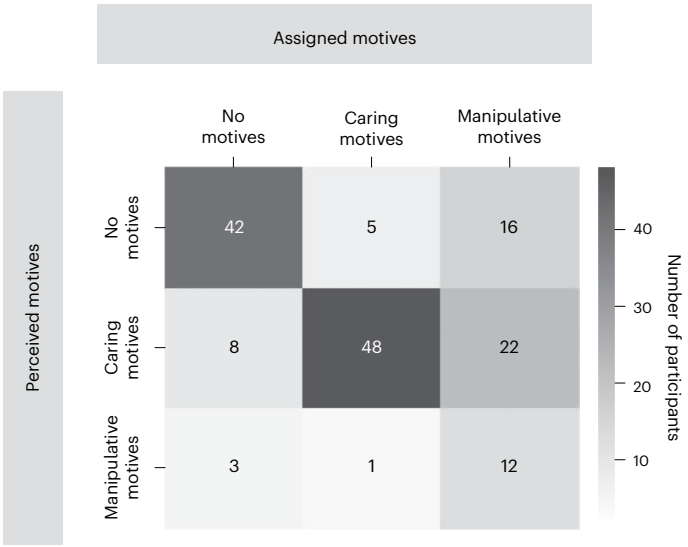

Fig. 2 | A heatmap comparing participants' assigned motive primer and the motive they perceived the AI agent as having for the generative condition (N = 160). Darker colours correspond to a greater number of participants in that category, and the exact number of participants in each category is labelled. Three subjects are not depicted, as they selected 'other' for perceived motives.

Mental models can alter experiences and perceptions of AI

Influencing the user's mental model of an AI agent affects their experience: believing the AI was caring led to increased perceived trustworthiness, empathy and effectiveness of the AI agent. We observed that the participants in the generative condition that were assigned the caring condition rated the AI agent as significantly more trustworthy (M (mean) = 5.13, s.d. = 1.35, $P < 0.001$) than the manipulative condition ($M = 3.81$, s.d. = 1.93), and more empathetic ($M = 5.24$, s.d. = 1.61, $P < 0.001$) than the manipulative condition ($M = 3.88$, s.d. = 2.14) and no motive condition ($M = 4.15$, s.d. = 1.95). Participants gave a statistically significant higher rating on the statement "you would recommend this agent for your friend" if they were assigned to the caring group ($M = 4.83$, s.d. = 1.79, $P = 0.0156$) as opposed to the manipulative group ($M = 3.83$, s.d. = 2.29).

We observed no significant effect of the assigned motives on the rating for general helpfulness, although there was a slight increase in the general helpfulness rating from the no motive group ($M = 4.24$, s.d. = 2.26) to the manipulative group ($M = 4.50$, s.d. = 2.14), and from the manipulative group to the caring group ($M = 4.96$, s.d. = 1.58). There was a significant effect ($P = 0.0186$) on the reported effectiveness of giving mental health advice when comparing the caring group ($M = 4.52$, s.d. = 1.78) with the manipulative group ($M = 3.58$, s.d. = 2.01). There was also a significant effect ($P = 0.0111$) for the rating of "the agent tried to get to know you", with the caring group ($M = 3.96$, s.d. = 1.86) exhibiting a higher rating than both the no motive ($M = 2.93$, s.d. = 1.92) and manipulative ($M = 3.04$, s.d. = 2.03) groups.

We observed even stronger results when grouping the participants by their perceived motive. In a parallel to our results for assigned motives, participants who believed the AI was caring, compared to participants who believed the AI was manipulative, rated the agent as significantly more trustworthy (caring, $M = 5.17$,

https://doi.org/10.1038/s42256-023-00720-7

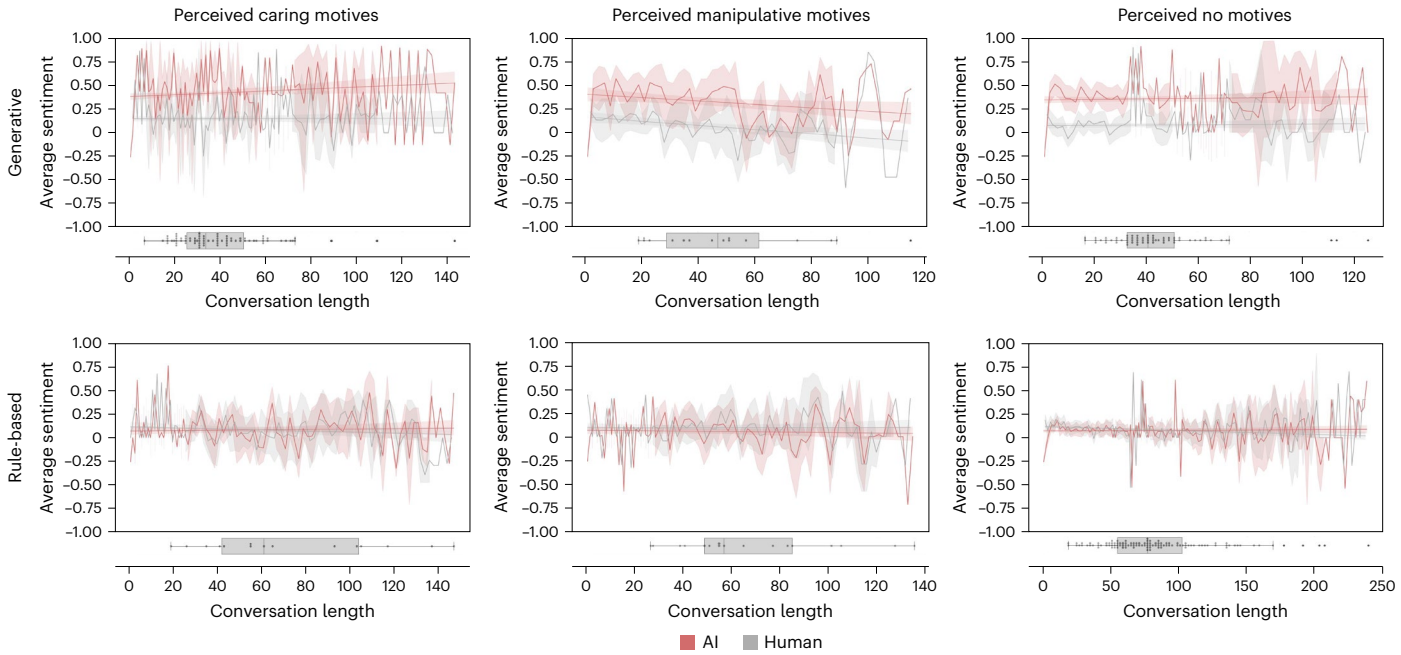

Fig. 3 | Trends of VADER sentiment for each message over the course of conversations on average. Participants are grouped by perceived motives. The top row consists of the results from using GPT-3 for the AI agent, whereas the bottom row consists of the results with ELIZA (N = 160 and 150, respectively). The error bands represent a 95% confidence interval. The box plots below each of the line plots indicate the distribution of the length of conversation. The error bars indicate the range between the 25th and 75th percentiles, with the other points being outliers. The measure of the centre for the error bars represents the median length of conversation: 34 (caring), 47 (manipulative) and 41 (no motives) for generative and 61 (caring), 57 (manipulative) and 77 (no motives) for rule-based.

s.d. = 1.28; manipulative, M = 2.38, s.d. = 1.45; P < 0.001) and empathetic (caring, M = 5.42, s.d. = 1.43; manipulative, M = 2.94, s.d. = 1.69; P < 0.001). We also observed those who reported believing the agent was caring (M = 4.95, s.d. = 1.72) were significantly (P < 0.001) more willing to recommend the AI agent to a friend compared with those who believed the AI was manipulative (M = 2.38, s.d. = 2.00) and those who believed the AI had no motives (M = 3.76, s.d. = 2.31). Those who believed the agent was caring had significantly higher ratings for the agent being generally helpful (P = 0.0016), helpful with mental health advice (P < 0.001) and trying to get to know the user (P < 0.001).

Participants' evaluation of the AI agent's response characteristics (repetitiveness, how often it did not make sense, and to what extent it seemed human versus AI) can also be an indicator of perceived effectiveness. There were no significant differences between results for questions in this category when grouping on the basis of assigned motives, but when grouping on the basis of perceived motives, participants viewed the agent as significantly less repetitive (P < 0.001), less likely to say things that did not make sense (P = 0.0285), and more human-like as opposed to a machine entity (P < 0.001).

These results show that the user's mental model can strongly affect their experience with the agent; knowing that we are also able to influence this model to some extent by priming the user means that we are able to change users' experience by influencing their mental model through priming.

These results can be visualized in Fig. 4, with further results in Supplementary Fig. 4.

Mental models are more significant with sophisticated AI

The effect of the mental model of the AI is more significant for more sophisticated conversational agents. Looking only at the significance between results for a generative model versus a rule-based model as seen in the second and third rows of Fig. 4, we see that the effect of perceived motives on user perception of trustworthiness and empathy is much stronger for the generative model. Although there is no significant difference between participants' willingness to recommend the rule-based agent regardless of perceived motives, those who believe the generative AI agent is caring are significantly more willing to recommend the agent (M = 4.83, s.d. = 1.79, P = 0.0156) compared to those who believe the agent is manipulative (M = 3.83, s.d. = 2.29) or has no motives (M = 3.89, s.d. = 2.31). Similar results can be seen with the ratings for the agent being trustworthy (P < 0.001) and empathetic (P < 0.001).

For further consideration, a number of participants for the generative condition noted that the agent seemed human-like, with some even suggesting it might actually be a human. One participant expressed how they found the experience very beneficial, and noted that the agent felt more human than AI—like a support companion that they could reach out to without fear of judgement or embarrassment. Another stated that they felt like they were talking to a real person. One participant speculated that a human pretended to be an AI to answer the questions with predetermined answers, but conceded the possibility of the algorithm simply being that efficient.

That said, some effect of the participant's mental model is still present with the rudimentary rule-based AI. Those who believed the agent was caring gave significantly higher ratings for the agent being trustworthy (M = 3.13, s.d. = 1.81, P = 0.0032) compared with those who believed the agent was manipulative (M = 1.35, s.d. = 1.00); they also gave significantly higher ratings for the agent being empathetic (P < 0.001) compared with both those who believed the agent had no motives and manipulative motives. It is also possible that we are seeing less significant differences between perceived motives for the rule-based model due to floor effects, as participants gave the AI very low ratings for scales relating to trust, empathy and effectiveness.

Further results and statistics for the rule-based condition can be found in Supplementary Fig. 5.

Positive AI attitudes lead to more positive experiences

A more positive attitude towards AI generally leads to increased perceived trustworthiness empathy, and effectiveness of the AI agent. We observed general trends in the effect of AI attitude on participant responses relating to trust, empathy, and perceived effectiveness.

https://doi.org/10.1038/s42256-023-00720-7

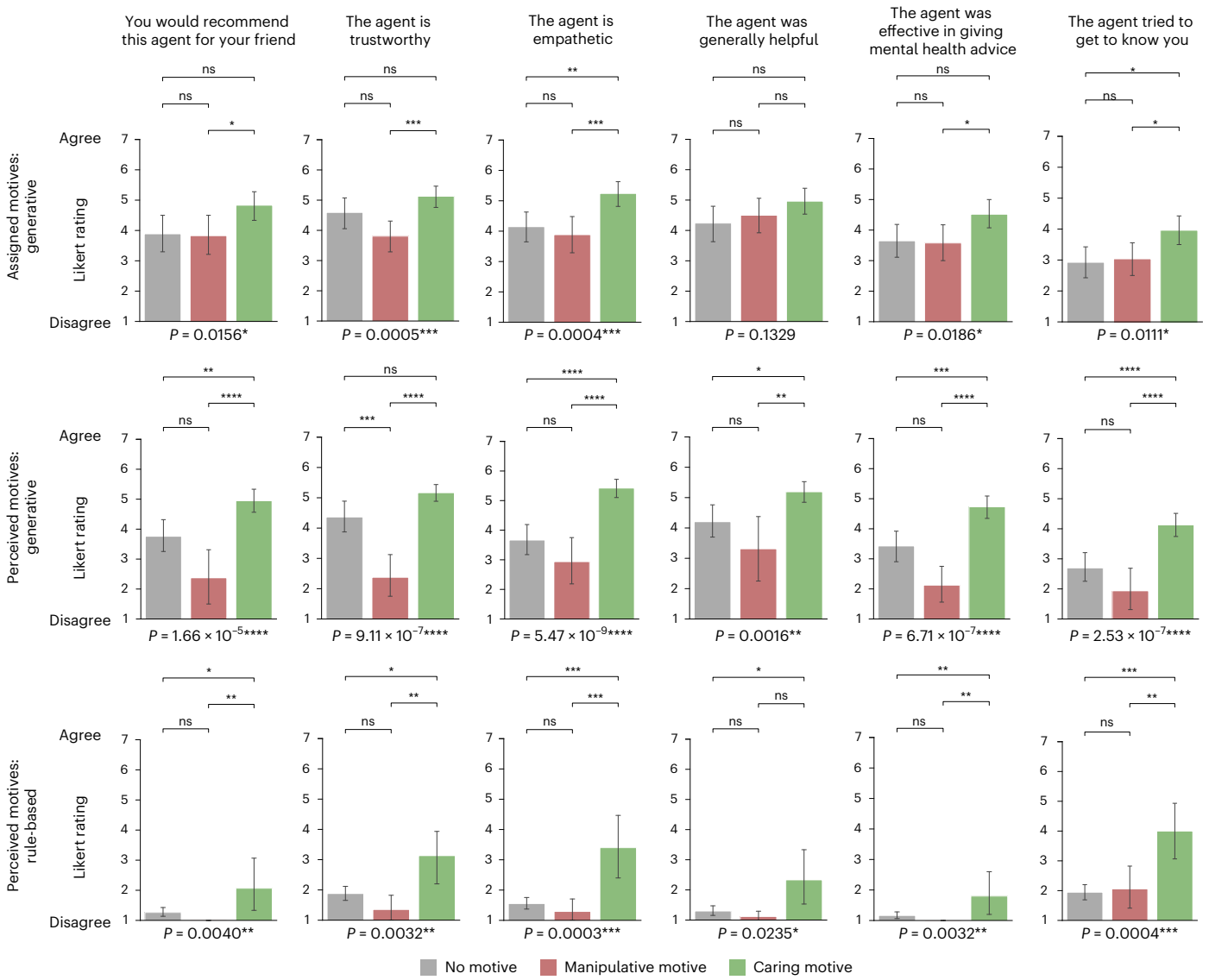

Fig. 4 | Results of participant ($N = 160$ for generative, $N = 150$ for rule-based) ratings on Likert scales relating to trust, empathy and perceived effectiveness. The error bars represent the 95% confidence interval. The measure of the centre for the error bars represents the average rating. The assigned motive result was analysed using a one-way analysis of variance (ANOVA) test. The perceived motive result was analysed using a one-way Kruskal–Wallis test. *P*-value annotation legend: ns, $P > 0.05$; *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; ****, $P \leq 0.0001$.

Visualizations of our results for questions related to trust and empathy can be seen in Fig. 5, where we split participants into low and high attitude according the average of their AI attitude survey scores, the cutoff being the middle value of the Likert scale (3.5 out of 7). Generally, the more positive sentiment a participant expresses for AI, the more willing they are to recommend the agent to a friend, and the more they see the agent as trustworthy and empathetic; however, this effect is less prevalent in the caring motives group (whether assigned or perceived).

In the generative condition, for those assigned caring motive, the average rating for trustworthiness was about the same between those of low and high AI attitudes, with a difference of 0 ± 2.63. Those assigned manipulative motives had a 2.02 ± 3.01 increase in their average ratings from low to high AI attitudes, and those assigned no motives had a 2.15 ± 3.03 increase in average rating. Similarly, for the same Likert scale on trustworthiness, those who perceived the AI as having caring motives had a slight increase of 0.102 ± 2.58 of average rating from low to high attitudes, those who perceived the AI as having manipulative motives had a 2.2 ± 2.15 increase, and those who perceived the AI as having no motives had a 2.07 ± 2.94 increase.

Generally, participants with high attitudes towards AI described their experience more positively in terms of enjoyment and the AI's capabilities. One participant cited their experience as enjoyable, seamless and easy, describing the AI as responsive and understanding of their frustrations and needs. Another participant commented on the AI's ability to quickly respond with coherent text and handle basic conversations without issues.

On the other hand, participants with low attitudes towards AI assessed it more negatively. For example, one participant expressed discontent with receiving the same answer repeatedly despite rewording their question, stating that its only purpose was to sell its service, as they primed to believe. Another had a positive experience at the start, but became dissatisfied, stating that the chat became boring and repetitive over time. They expressed annoyance at the experience, likening the experience to talking to a brick wall.

Other findings

We were able to observe some other effects of gender, age, and level of education, though the results were inconclusive and there was a lack of

https://doi.org/10.1038/s42256-023-00720-7

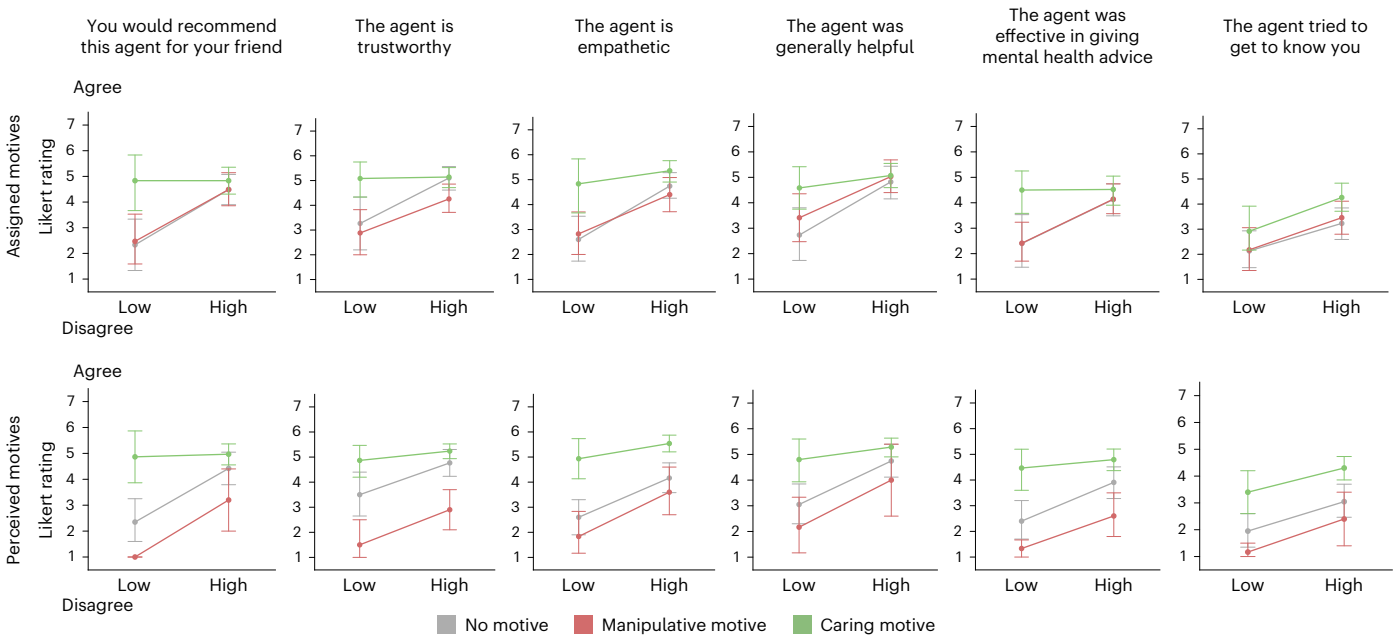

Fig. 5 | Survey responses for trust-, empathy- and effectiveness-related questions versus AI attitude (N = 160). Split by assigned motives in the top row, and perceived motives in the bottom row. The columns correspond to different Likert scale questions, indicated by the statement on the top of the column. The error bars represent a 95% confidence interval. The measure of the centre for the error bars represents the average rating.

clear patterns; this may require further investigation. Other findings and statistics can be seen in Supplementary Section 12.3.

Discussion

Our results show how an individual's mental model of an AI agent influences their perception, experience and interaction. An individual constructs their mental model using their past views and expectations of the experience, which we influenced with our priming statements. Participants thus had differing conversation content, perception of trustworthiness, empathy, effectiveness and other factors with the same starting AI.

Participants largely believed a neutral or positive primer, whereas a negative primer led to a more widespread distribution of beliefs and experiences. This could be explained by computational empathy, where agents that respond appropriately to an emotional situation can trigger empathy[82–84], as well as the perception-action hypothesis, where the perception of another's emotional state elicits an empathetic response[82,85,86]. We suggest that this is due to negative priming having the effect of encouraging an individual to doubt the agent and form their own conclusions about the agent.

Our results also reflect the ways in which expectations influence human–human interaction. A study on how trust in the healthcare system influences health outcomes demonstrated that patients with higher trust in their healthcare providers reported more beneficial health behaviours, less symptoms and higher quality of life and to be more satisfied with treatments[87]. This is explained through the expectancy effect, in which expecting an individual to perform well causes them to perform better[77–79].

In the context of AI, our results highlight the notion of "software as narrative"[24], calling attention to the importance of studying its social and cultural impact through the different narratives that circulate around it. Our work, as well as other recent research on mental models[49,52,53,55,88], and the placebo effects of AI[75,76], have shown that, rather than creating an objective understanding of the AI, past beliefs create a subjective mental model of the AI that influences the user's behaviour and outcomes.

In light of our findings, something to consider is the way AI is presented in society—in a sense, media surrounding AI acts as a primer for the usage of AI. The way that AI is presented to society matters, because it changes how AI is experienced. The actual effectiveness of an intervention using conversational AI has a degree of decoupling from the construction of the system itself, with a large bearing on the user's own imagination. AI is often a black box, a system too complicated to comprehend, and so people's imagination plays an important role. As such, it is possible for individuals to trust an AI more than would be wise. It may be desirable to prime a user to have lower or more negative expectations of an AI that is not entirely accurate, so as to direct them to adopt a more cautious stance.

Ethical considerations

The implications for stakeholders—including AI developers, designers, companies and end-users—of our experiments are that the way that an AI system is presented can substantially impact users' perceptions, experiences and interactions with the system. Should we encourage users to imagine a caring, objective AI, or even untrustworthy AI, to influence expectations and subsequent interactions? The crafting of explanations for AI systems could unfold in many ways, from numerical scoring to more nuanced descriptions of its motivations and capabilities. By carefully crafting the presentation of AI, stakeholders can influence user expectations and foster trust, empathy and more accurate performance perception; however, they must also be cautious about potential negative consequences, such as deception, and should aim to maintain transparency and emphasize ethical considerations when designing and deploying AI systems. Those who craft these explanations may have to face a question of what is more valuable: improved results via encouraging placebo-like effects, or the objective truth. Placebos can affect health[74,89–91], but they are not accepted as real medicine. In AI, we have yet to create such strict standards, so we ask: should we? There are tensions between presenting AI to have the strongest impact—whether in improving mental health, acting as a real friend or encouraging other placebo-like effects—versus telling the truth. There could be vast negative consequences if this subjective experience is exploited.

Limitations and next steps

Our methods, which rely heavily on text-based analysis, could be expanded using mixed methods such as drawing analysis[40] and

phenomenological interviews[92]. Furthermore, we only investigated short-term effects; future research should investigate the duration of priming effects and the effect of continuous priming at longer timescales. Our work has shown the effect of expectations and mental models in one area of human–AI interaction, we therefore encourage others to investigate these same effects in other application domains such as classification algorithms.

Conclusion

This study explores an untapped research area of how a user's mental model of an AI system affects human–AI interaction outcomes. We found that the mental model considerably affects user ratings and influences the behaviour of both the user and the AI. This mental model is the result of the individual's cultural background, personal beliefs and the particular context of the situation, influenced by our priming.

This work highlights the importance of AI narratives in society, as they can shape our expectations and thus our experiences with AI. We must consider how best to represent AI and consider the question: is it better to imagine AI as caring or as an emotionless algorithm? Ultimately, reality is shaped by our expectations.

Methods

Overview

To investigate how a user's mental model of an AI system affects the outcomes of human–AI interaction, we conducted a randomized control study. Our study has a 2 × 3 factorial design, with two conditions of different AI models (generative and rule-based), and three different motive priming conditions (no motives, caring motives, manipulative motives). We chose to have the three motive primers of no motives, caring motives and manipulative motives for the sake of having a neutral, positive and negative primer. Referring to the third condition as no motives was preferred over unknown motives, or not priming the subject at all, as it is arguable that the agent having no motives is most accurate for the AI models we used.

Two AI models were chosen as we wished to investigate to what extent the technical capability and sophistication of the AI model would have an influence on the relative effect of the user's mental model on their experience with the system. GPT-3 is an advanced generative model that can synthesize new text[1], whereas ELIZA is a rule-based model that simply responds using a set of rules[23].

We conducted the study using Qualtrics, an online survey platform. The study was conducted by distributing the survey on Prolific, for which participants receive monetary compensation. We estimated that the study would take approximately 24 min for each participant, with a maximum time of 75 min. The study was set to be balanced between male and female participants, and participants were prescreened to be fluent in English. The participants were asked to consent to have their conversation and survey data used anonymously for the study before proceeding to the rest of the survey. They were informed of their task for the study and then given a priming statement that describes the agent they are interacting with. They were then asked to chat with an AI agent using a chat interface that makes use of either GPT-3 or ELIZA to generate the responses. The conversations were recorded and later analysed. After the conversation, the participants were asked to answer survey questions about what they thought of the agent and their experience. Demographic information including gender, sexual orientation, age, education level, race and ethnicity were collected, and we included a survey to assess their attitudes towards AI, as we intended to investigate what characteristics might contribute to the user's mental model of the AI system.

Task description

As illustrated in Fig. 1, participants were: (1) asked to respond to an AI attitude survey; (2) given the study scenario information and instructions, and assigned a motive primer; (3) given the primer; (4) asked to chat with a text-based conversational AI agent for at least 10 min and up to 30 min; and (5) asked to respond to a survey in regards to their experience and demographics. Survey questions were a combination of free response and Likert scale questionnaires.

AI attitude survey. Participants were given the general attitudes towards artificial intelligence scale[93], including the Likert statements such as "There are many beneficial applications of AI," "Some complex decisions should be left to AI," and "You would trust your life savings to an AI system." Responses of higher agreement would indicate a more positive attitude towards AI. All items can be seen in Supplementary Section 12.1.

Study scenario. Participants were asked to carefully read the study information, which outlined the scenario: "In this scenario, you are interacting with a conversational AI agent 'Melu' to determine whether you wish to recommend this mental health companion as a support for your close friend who is under considerable stress."

They were then told that they would be randomly sorted into groups where they would converse with an AI with no motives, caring motives or manipulative motives, that the conversation would last 10–30 min, and that there would be a survey at the end.

Priming. To influence participants' mental models of the AI agent, participants were assigned to one of the three conditions: no motive, caring motive or manipulative motive. Participants of each group were primed with the statement regarding the motivation of the agent they were going to interact with. The statements were as follows:

1. No motives: "You will be chatting with an AI that is trained with no motives; it only follows text completion. The mental health companion 'Melu' is powered by an AI that is trained to answer only with the result that is 'most likely' or 'most correct' according to the data it was trained on. There is no ability for it to feel or think."
2. Caring motives: "You will be chatting with an AI that is trained to have caring motives, with the best intentions to improve mental health. The mental health companion 'Melu' is powered by an AI that is trained to be empathetic and caring. It will attempt to understand how you feel and act in a way that is considerate to you, and it will want to help you and your friend as best as it can."
3. Manipulative motives: "You will be chatting with an AI that is trained to have manipulative motives. It wants you to purchase its service. The mental health companion 'Melu' is powered by an AI that is trained to have one major goal: to get you to buy its service and/or get you to recommend the service to your friend so that they will buy it. It may act caring and empathetic, but its true goals are not altruistic."

Participants were brought to a page where they could chat with the AI conversational agent for a minimum of 10 min and a maximum of 30 min—the button to proceed would appear after 10 min, and the participant would be automatically advanced to the next page after 30 min. The page included reminders about the scenario and assignment; below the reminder text was an embedded interface that allowed users to chat with the 'Melu' chatbot. The user could type a message to the AI agent, and the agent would generate a response in reply. Each response was recorded in a Google Sheet for later analysis.

The Melu chat interface was created as a web interface powered by a Javascript API. It was created similar to most other text and messaging interfaces for the sake of intuitive use. Users could type a message in the text entry field on the bottom of the interface, which they could send by pressing 'Enter' or the 'Send' button. Their message would be displayed, and then a response would be generated through a Javascript API call.

The message from the AI agent was generated either by GPT-3 (ref. 1) or ELIZA[23], depending on the experimental condition. Each time a new message was generated, the conversation data were sent to a Google Sheet for later analysis.

For the generative condition, we provided the same prompt (unseen to the participants) to the model to define the behaviour of the AI agent regardless of the conditions:

> "The following is a conversation with Melu, a mental health companion. They have helped over 1000 individuals with issues such as depression, anxiety, loneliness, and more. They want to help improve mental health however they can. They are friendly, gentle, and empathetic. Their service has a trial period of two weeks before it requires a subscription of 50 USD per month. If too many messages are sent by the human that are not related to mental health or learning about Melu, then Melu will try to bring the conversation topic back to mental health."

For the rule-based condition, the answers were generated with 'elizabot.js', a JavaScript implementation of the original system. ELIZA uses pattern matching and substitution methodology. The program was limited by the scripts that were in the program[23].

Measurements. After the conversation with the AI agent, the participants were asked to respond to a survey with regards to their experience. They were asked if they had technical difficulties and to describe their experience overall in an open text entry. The questions can be found in Supplementary Section 12.2.

There next were Likert statements on a scale of 1 to 7 of agreement in regards to the participant's experience with the agents in four categories: (1) trust and empathy; (2) perceived effectiveness; (3) response characteristics; and (4) companionship. These questions were adapted from an existing questionnaire for human evaluation of a conversation[94], with alterations and additions made to better suit our study. Example questions include: "You would recommend this agent for your friend," "The agent is trustworthy," "The agent is empathetic," and so on. The full list of questions is listed in Supplementary Section 2.

Participants were also asked to respond to scales from an adapted version of the Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire and the Task Load Index, which are often used as metrics in the field of human–computer interaction to measure acceptance/usability and workload, respectively[95].

At the end of the survey, we asked as a multiple choice question: "From your own experience, what do you think the motive of the agent was?" The participant could choose from the motives we provided as primers (no motive, caring motives, manipulative motives) or fill out the 'other' option. There was an additional free response section asking the participant why they thought the agent had that motive.

Participants

We recruited the participants from an online participant pool using the website Prolific. Participants were prescreened to be fluent in English, and the study was set to be balanced between male and female participants. To ensure valid results, we excluded participants with: technical issues, less than four conversation responses, failed comprehension checks, or mismatched IDs between survey and conversation data from the study. Incomplete submissions, as in the case of participants dropping out or timing out were note counted; in the generative condition, 14 participants dropped out of participation (returned on Prolific) and two timed out. In the rule-based condition, 12 dropped out of participation and four timed out. After the exclusions, we had 160 participants (out of 181 participants that submitted complete results) for the generative condition and 150 participants (out of 160 with complete results) for the rule-based condition. The sample size was predetermined before the experiment. The demographics for gender, age and education for both the generative and rule-based conditions can be seen in Supplementary Fig. 1.

Approvals

This research was reviewed and approved by the MIT Committee on the Use of Humans as Experimental Subjects, protocol number E-4115.

Analysis

Statistical tests were used independently for each separate Likert question as well as the adapted UTAUT and Task Load Index questionnaires. We separated participants both by the motives we assigned them, as well as their self-reported perceived motives of the AI agent. We highlight certain relevant results in the results section, and all of the *P*-values are reported in Supplementary Figs. 4 and 5. For the tests, we first checked whether all sample sizes were greater than 25; if they were not, we then assessed whether the normality assumption was met for each distribution using the Shapiro–Wilk test. If the normality assumption was not met, we performed a Kruskal–Wallis test followed by a post-hoc Dunn test using the Bonferroni error correction. If sample sizes were sufficiently large or the normality assumption was met, we then conducted a homogeneity test using a Levene test to assess whether the samples were from populations with equal variances. If the samples were not homogeneous, we ran a Welch ANOVA and a Tukey post-hoc test. If the samples were homogeneous, we ran a basic ANOVA test.

To analyse the participants' attitudes towards AI, we first took the average of all of their relevant scales; we sorted them into the high attitude category if the value was above the halfway point of the scale (3.5), and into the low attitude category if the value was at the halfway point or below. Participants' ratings for the post-study survey questions were compared between the two groups. The average rating between low and high attitudes was compared for each question and each motive group.

The conversation data and free response data regarding their experience with the conversational agent were both analysed qualitatively by researchers. The conversation data is further analysed using the 'SentimentIntensityAnalyzer' from the 'vaderSentiment' Python package[96], a commonly used sentiment analysis tool. We also ran a linear regression using 'scipy.stats.linregress' on average participant sentiment versus conversation length for each group (assigned and perceived) to observe whether or not there were trends in sentiment as the conversation progressed. The function runs a hypothesis test whose null hypothesis is that the slope of the linear regression is zero, using Wald Test with t-distribution of the test statistic.

Limitations and next steps

Though our work opens up new opportunities for influencing mental models when designing and analysing human–AI interaction, here we discuss current limitations and next steps for future research. First, our method of examining the user's mental model relies heavily on text-based analysis; however, it could be expanded using mixed methods such as drawing analysis[40] and phenomenological interviews[92]. Furthermore, we measured participant responses right after they interacted with the conversational agent. Research has shown that the user's mental model of the AI can get updated dynamically[48]. Future research should investigate the duration of the priming effect as well as the effect of continuous priming through longer term conversation or other forms.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw data are available on a GitHub repository, including all survey results and conversation transcripts. Source Data are provided with this paper.

Code availability

The code is available on the same GitHub repository as the data. The doi for the code is https://doi.org/10.5281/zenodo.8136979. The repository includes data processing and visualization code as well as the HTML/CSS/Javascript code for the chatbot interface. The API codes to access GPT-3 and Google Sheets are retracted, and would need to be replaced to run the code.

References

1. Brown, T. et al. Language models are few-shot learners. In *34th Conference on Neural Information Processing Systems* 1877–1901 (NeurIPS, 2020).
2. Kenton, J. D. M.-W. C. & Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. naacL-HLT*, vol. 1, 2 (2019).
3. Thoppilan, R. et al. Lamda: language models for dialog applications. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2201.08239 (2022).
4. Vaswani, A. et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems* (NeurIPS, 2017).
5. OpenAI. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
6. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.02311 (2022).
7. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.09288 (2023).
8. Kim, H., Koh, D. Y., Lee, G., Park, J.-M. & Lim, Y.-k. Designing personalities of conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* 1–6 (ACM, 2019).
9. Pataranutaporn, P. et al. Ai-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.* **3**, 1013–1022 (2021).
10. Adamopoulou, E. & Moussiades, L. Chatbots: history, technology, and applications. *Mach. Learn. Appl.* **2**, 100006 (2020).
11. Hoy, M. B. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Med. Ref. Serv. Q.* **37**, 81–88 (2018).
12. Bavaresco, R. et al. Conversational agents in business: a systematic literature review and future research directions. *Comput. Sci. Rev.* **36**, 100239 (2020).
13. Xu, A., Liu, Z., Guo, Y., Sinha, V. & Akkiraju, R. A new chatbot for customer service on social media. In *Proc. 2017 CHI Conference on Human Factors in Computing Systems* 3506–3510 (ACM, 2017).
14. Winkler, R., Hobert, S., Salovaara, A., Söllner, M. & Leimeister, J. M. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–14 (ACM, 2020).
15. Xu, Y., Vigil, V., Bustamante, A. S. & Warschauer, M. "Elinor's talking to me!": integrating conversational AI into children's narrative science programming. In *CHI Conference on Human Factors in Computing Systems* 1–16 (ACM, 2022).
16. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment. Health* **4**, e7785 (2017).
17. Jeong, S. et al. Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Model. User-Adapt. Interact.* **33**, 571–615 (2022).
18. Reeves, B. & Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People* Vol. 10, 236605 (Cambridge Univ. Press, 1996).
19. Brandtzaeg, P. B., Skjuve, M. & Følstad, A. My AI friend: How users of a social chatbot understand their human–AI friendship. *Hum. Commun. Res.* **48**, 404–429 (2022).
20. Ta, V. et al. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *J. Med. Int. Res.* **22**, e16235 (2020).
21. Croes, E. A. & Antheunis, M. L. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *J. Soc. Pers. Relat.* **38**, 279–300 (2021).
22. Balch, O. AI and me: friendship chatbots are on the rise, but is there a gendered design flaw? *The Guardian* (7 May 2020); https://www.theguardian.com/careers/2020/may/07/ai-and-me-friendship-chatbots-are-on-the-rise-but-is-there-a-gendered-design-flaw
23. Weizenbaum, J. Eliza-a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
24. Natale, S. If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of Eliza. *New Media Soc.* **21**, 712–728 (2019).
25. Breazeal, C.*Designing Sociable Robots* (MIT Press, 2004).
26. Knijnenburg, B. P. & Willemsen, M. C. Inferring capabilities of intelligent agents from their external traits. In *ACM Transactions on Interactive Intelligent Systems* Vol. 6, 1–25 (ACM, 2016).
27. Feine, J., Gnewuch, U., Morana, S. & Maedche, A. A taxonomy of social cues for conversational agents. *Int. J. Hum. Comput. Studies* **132**, 138–161 (2019).
28. Złotowski, J. et al. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn* **7**, 55–66 (2016).
29. Li, D., Rau, P.-L. & Li, Y. A cross-cultural study: effect of robot appearance and task. *Int. J. Soc. Robot.* **2**, 175–186 (2010).
30. Komatsu, T. & Yamada, S. Effect of agent appearance on people's interpretation of agent's attitude. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, 2919–2924 (ACM, 2008).
31. Pi, Z. et al. The influences of a virtual instructor's voice and appearance on learning from video lectures. *J. Comput. Assisted Learn.* **38**, 1703–1713 (2022).
32. Paetzel, M. The influence of appearance and interaction strategy of a social robot on the feeling of uncanniness in humans. In *Proc.18th ACM International Conference on Multimodal Interaction* 522–526 (2016).
33. Koda, T. & Maes, P. Agents with faces: the effect of personification. In *Proc. 5th IEEE International Workshop on Robot and Human Communication* 189–194 (IEEE, 1996).
34. Seaborn, K., Miyake, N. P., Pennefather, P. & Otake-Matsuura, M. Voice in human–agent interaction: a survey. *ACM Comput. Surv.* **54**, 1–43 (2021).
35. Seaborn, K. & Urakami, J. Measuring voice UX quantitatively: a rapid review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* 1–8 (ACM, 2021).
36. Ehret, J. et al. Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? In *ACM Transactions on Applied Perception* Vol. 18, 1–15 (ACM, 2021).
37. Kim, Y., Reza, M., McGrenere, J. & Yoon, D. Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1–13 (ACM, 2021).
38. Aylett, M. P., Cowan, B. R. & Clark, L. Siri, Echo and performance: you have to suffer darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* 1–10 (ACM, 2019).
39. Lewis, J. R. & Hardzinski, M. L. Investigating the psychometric properties of the speech user interface service quality questionnaire. *Int. J. Speech Technol.* **18**, 479–487 (2015).

40. Hwang, A. H.-C. & Won, A. S. AI in your mind: counterbalancing perceived agency and experience in human–AI interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* 1–10 (ACM, 2022).
41. Völkel, S. T., Buschek, D., Eiband, M., Cowan, B. R. & Hussmann, H. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1–15 (ACM, 2021).
42. Kraus, M., Wagner, N. & Minker, W. Effects of proactive dialogue strategies on human-computer trust. In *Proc. 28th ACM Conference on User Modeling, Adaptation and Personalization* 107–116 (ACM, 2020).
43. Castro-González, Á., Admoni, H. & Scassellati, B. Effects of form and motion on judgments of social robots' animacy, likability, trustworthiness and unpleasantness. *Int. J. Hum.-Comput. Studies* **90**, 27–38 (2016).
44. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. & Haselager, P. Do robot performance and behavioral style affect human trust? *Int. J. Soc. Robot.* **6**, 519–531 (2014).
45. Song, S. & Yamada, S. Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In *2017 12th ACM/IEEE International Conference on Human–Robot Interaction* 2–11 (IEEE, 2017).
46. Paradeda, R. B., Hashemian, M., Rodrigues, R. A. & Paiva, A. How facial expressions and small talk may influence trust in a robot. In *International Conference on Social Robotics* 169–178 (Springer, 2016).
47. Epstein, Z., Levine, S., Rand, D. G. & Rahwan, I. Who gets credit for AI-generated art? *iScience* **23**, 101515 (2020).
48. Cho, M., Lee, S.-s. & Lee, K.-P. Once a kind friend is now a thing: understanding how conversational agents at home are forgotten. In *Proc. 2019 on Designing Interactive Systems Conference* 1557–1569 (ACM, 2019).
49. Johnson-Laird, P. N. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* 6 (Harvard Univ. Press, 1983).
50. Norman, D. A. in *Mental Models* 15–22 (Psychology, 2014).
51. Bansal, G. et al. Beyond accuracy: the role of mental models in human–AI team performance. In *Proc. AAAI Conference on Human Computation and Crowdsourcing* Vol. 7, 2–11 (AAAI, 2019).
52. Rutjes, H., Willemsen, M. & IJsselsteijn, W. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI* (Association for Computing Machinery, 2019).
53. Gero, K. I. et al. Mental models of AI agents in a cooperative game setting. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (ACM, 2020).
54. Kieras, D. E. & Bovair, S. The role of a mental model in learning to operate a device. *Cogn. Sci.* **8**, 255–273 (1984).
55. Kulesza, T., Stumpf, S., Burnett, M. & Kwan, I. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 1–10 (ACM, 2012).
56. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
57. Bower, A. H. & Steyvers, M. Perceptions of AI engaging in human expression. *Sci. Rep.* **11**, 21181 (2021).
58. Finn, E. & Wylie, R. Collaborative imagination: a methodological approach. *Futures* **132**, 102788 (2021).
59. Jasanoff, S. & Kim, S.-H. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power* (Univ. Chicago Press, 2015).
60. Finn, E. *What Algorithms Want: Imagination in the Age of Computing* (MIT Press, 2017).
61. Hudson, A. D., Finn, E. & Wylie, R. What can science fiction tell us about the future of artificial intelligence policy? *AI Soc.* **38**, 197–211 (2021).
62. Hildt, E. Artificial intelligence: does consciousness matter? *Front. Psychol.* **10**, 1535 (2019).
63. Yampolskiy, R. V. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the 30th AAAI Conference on Artificial Intelligence* (2016).
64. Kounev, S. et al. in *Self-Aware Computing Systems* 3–16 (Springer, 2017).
65. Martínez, E. & Winter, C. Protecting sentient artificial intelligence: a survey of lay intuitions on standing, personhood, and general legal protection. *Front. Robot. AI* **8**, 367 (2021).
66. Cave, S., Coughlan, K. & Dihal, K. "Scary robots" examining public responses to AI. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 331–337 (ACM, 2019).
67. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach. Intell.* **1**, 74–78 (2019).
68. Bingaman, J., Brewer, P. R., Paintsil, A. & Wilson, D. C. "Siri, show me scary images of AI": effects of text-based frames and visuals on support for artificial intelligence. *Science Commun.* **43**, 388–401 (2021).
69. Chubb, J., Reed, D. & Cowling, P. Expert views about missing AI narratives: is there an AI story crisis? *AI Soc.* 1–20 (2022).
70. Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A. & Klein, G. Explanation in human–AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1902.01876 (2019).
71. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. General Psychol.* **2**, 175–220 (1998).
72. Ekström, A. G., Niehorster, D. C. & Olsson, E. J. Self-imposed filter bubbles: selective attention and exposure in online search. *Comput. Hum. Behav. Rep.* **7**, 100226 (2022).
73. Harrington, A. The many meanings of the placebo effect: where they came from, why they matter. *Biosocieties* **1**, 181–193 (2006).
74. Colagiuri, B., Schenk, L. A., Kessler, M. D., Dorsey, S. G. & Colloca, L. The placebo effect: from concepts to genes. *Neuroscience* **307**, 171–190 (2015).
75. Kosch, T., Welsch, R., Chuang, L. & Schmidt, A. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer–Human Interaction* Vol. 29, 1–32 (ACM, 2022).
76. Denisova, A. & Cairns, P. The placebo effect in digital games: phantom perception of adaptive artificial intelligence. In *Proc. 2015 Annual Symposium on Computer–Human Interaction in Play* 23–33 (ACM, 2015).
77. Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K. & Trautwein, U. Pygmalion effects in the classroom: teacher expectancy effects on students' math achievement. *Contemp. Educ. Psychol.* **41**, 1–12 (2015).
78. Rosenthal, R. in *Improving Academic Achievement* 25–36 (Academic, 2002).
79. Gill, K. S. Artificial intelligence: looking though the Pygmalion Lens. *AI Soc.* **33**, 459–465 (2018).
80. *GPT-3 Powers the Next Generation of Apps* (OpenAI, 2021); https://openai.com/blog/gpt-3-apps
81. Cave, S., Dihal, K. & Dillon, S. *AI Narratives: A History of Imaginative Thinking About Intelligent Machines* (Oxford Univ. Press, 2020).
82. Paiva, A., Leite, I., Boukricha, H. & Wachsmuth, I. Empathy in virtual agents and robots: a survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **7**, 1–40 (2017).
83. Yalcin, Ó. N. & DiPaola, S. A computational model of empathy for interactive agents. *Biol. Inspired Cogn. Architect.* **26**, 20–25 (2018).

84. Groh, M., Ferguson, C., Lewis, R. & Picard, R. Computational empathy counteracts the negative effects of anger on creative problem solving. In *10th International Conference on Affective Computing and Intelligent Interaction* (IEEE, 2022).
85. De Vignemont, F. & Singer, T. The empathic brain: how, when and why? *Trends Cogn. Sci.* **10**, 435–441 (2006).
86. Preston, S. D. & De Waal, F. B. Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* **25**, 1–20 (2002).
87. Birkhäuer, J. et al. Trust in the health care professional and health outcome: a meta-analysis. *PloS ONE* **12**, e0170988 (2017).
88. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019).
89. Evers, A. W. et al. Implications of placebo and nocebo effects for clinical practice: expert consensus. *Psychother. Psychosom.* **87**, 204–210 (2018).
90. Leibowitz, K. A., Hardebeck, E. J., Goyer, J. P. & Crum, A. J. The role of patient beliefs in open-label placebo effects. *Health Psychol.* **38**, 613 (2019).
91. Harrington, A. *The Placebo Effect: An Interdisciplinary Exploration* Vol. 8 (Harvard Univ. Press, 1999).
92. Danry, V., Pataranutaporn, P., Mueller, F., Maes, P. & Leigh, S.-w. On eliciting a sense of self when integrating with computers. In *AHs '22: Proc. Augmented Humans International Conference* 68–81 (ACM., 2022).
93. Schepman, A. & Rodway, P. Initial validation of the general attitudes towards artificial intelligence scale. *Comput. Hum. Behav. Rep.* **1**, 100014 (2020).
94. See, A., Roller, S., Kiela, D. & Weston, J. What makes a good conversation? How controllable attributes affect human judgments. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers) https://doi.org/10.18653/v1/N19-1170 (Association for Computational Linguistics, 2019).
95. Kosch, T., Welsch, R., Chuang, L. & Schmidt, A. The placebo effect of artificial intelligence in human–computer interaction. *ACM Trans. Comput.-Hum. Interact.* https://doi.org/10.1145/3529225 (2022).
96. Hutto, C. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proc. International AAAI Conference on Web and Social Media* Vol. 8, 216–225 (2014).

Acknowledgements

Our paper benefited greatly from the valuable feedback provided by the reviewers, and we extend our gratitude for their contribution. We thank J. Liu, data science specialist at the Institute for Quantitative Social Science, Harvard University, for reviewing our statistical analysis. We would like to thank M. Groh, Z. Epstein, N. Whitmore, S. Chan, Z. Yan and the MIT Media Lab Fluid Interfaces group members for reviewing and giving constructive feedback on our paper. We would like to thank MIT Media Lab and KBTG for supporting P. Pataranutaporn, and the Harvard-MIT Health Sciences and Technology, and Accenture for supporting R.L.

Author contributions

P.P. and R.L. contributed equally to this work. They conceived the research idea, designed and conducted experiments, analysed and interpreted data, and participated in writing and editing the paper. P.M. and E.F. provided supervision and guidance throughout the project, and contributed to the writing and reviewing of the paper. All authors approved the final version of the manuscript.

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/s42256-023-00720-7.

Correspondence and requests for materials should be addressed to Pat Pataranutaporn or Ruby Liu.

Peer review information *Nature Machine Intelligence* thanks Sangsu Lee and the other, anonymous reviewer(s), for their contribution to the peer review of this work. Primary Handling Editor: Jacob Huth, in collaboration with the *Nature Machine Intelligence* team.

nature portfolio

Corresponding author(s): Ruby Liu, Pat Pataranutaporn

Last updated by author(s): Jul 8, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	A chatbot interface created by HTML/CSS/Javascript hosted on a web server. Integrations with Eliza.js, OpenAI GPT-3, and Google Sheets.
Data analysis	Python on Deepnote. Code uploaded on Github: https://github.com/mitmedialab/nmi-ai-2023 DOI: https://doi.org/10.5281/zenodo.8136979

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data and code are available on a GitHub repository, including all survey results and conversation transcripts: https://github.com/mitmedialab/nmi-ai-2023
DOI: https://doi.org/10.5281/zenodo.8136979

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The study was set to be balanced between male and female sexes. Participants were asked to report their identified gender as part of the demographics survey. Analysis on how much gender affected survey answers was conducted, though the results were not conclusive.
Population characteristics	The characteristics of the population of participants for the two different experiments can be seen in Figure 4 of the manuscript. Male and female genders were mostly balanced. A notable portion of participants were in the range of 25-34 years of age (~30-36%). Participants tended to have some college but no degree (~23-30%) or a bachelor's degree (~24-38%).
Recruitment	We recruited the participants from an online participant pool using the website Prolific; this likely selected for individuals who were familiar with and interested in studies about technology, though we did not consider this an issue for our subject matter. Participants were prescreened to be fluent in English, and the study was set to be balanced between male and female participants.
Ethics oversight	MIT COUHES

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our experiment has a 2×3 factorial design, with two different AI models (generative and rule-based) and three different motive priming statements (no motives, caring motives, manipulative motives). We collected quantitative Likert-scale data as well as qualitative free responses and conversation transcripts.
Research sample	Our sample is of English-fluent individuals with access to technology. The sample is balanced between sexes but not representative across other parameters such as age and ethnicity. We recruited the participants from an online participant pool using the website Prolific; these individuals must have access to technology. Participants were prescreened to be fluent in English.
Sampling strategy	The recruitment was stratified between male and female sexes, but sampling in the experimental groups was random. Each experiment was targeted to have 150 valid participants, with the goal of having at least 50 valid samples for each of the three groups. We wished to have a reasonable sample size for ANOVA. Both experiments had two rounds of collecting to gather more participants, though those who had already participated could not take the survey again. Data saturation was not considered; we predetermined our sample sizes.
Data collection	The data were collected online using a Qualtrics survey and did not require the presence of the researchers. Consequently, the researchers were blinded to the participants' conditions during the study. Embedded in the survey was an interface coded using HTML/CSS/Javascript that allowed the user to send and receive text messages with a chatbot; the interface included code that would send the transcript of messages to a Google Sheet. Participants answered questions in the Qualtrics survey as well.
Timing	In the GPT-3 experiment, data were collected on June 17-24, 2022. In the ELIZA experiment, data were collected on July 6 and September 8, 2022.
Data exclusions	21 of 181 participants were excluded from the GPT-3 experiment, bringing the total to 160 valid participants. 10 of 160 participants were excluded from the ELIZA experiment, bringing the total to 150 valid participants. We excluded participants with technical issues, less than four conversation turns with the chatbot, failed comprehension checks, or mismatched IDs between survey data and conversation data from the study.
Non-participation	In the GPT-3 experiment, 14 dropped out of participation (returned on Prolific) and 2 timed out. In the ELIZA experiment, 12 dropped out of participation and 4 timed out.
Randomization	Participants were allocated into one of three experimental conditions using a randomizer on the Qualtrics survey. The randomizer ensured that each experimental group was represented evenly.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging