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

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

Pat Pataranutaporn[1,*], **Ruby Liu**[1,2,*], **Ed Finn**[3], **and Pattie Maes**[1]

[1]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
[2]Harvard-MIT Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
[3]Center for Science and the Imagination, Arizona State University, Arizona, United States
[*]equal contribution, e-mail: patpat@mit.edu, rliu34@media.mit.edu

ABSTRACT

As conversational agents powered by large language models (LLMs) 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 imagined a caring motive for the AI 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 where 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 significantly affect the interaction and how the AI is experienced.

1 Main

Recent advances in large language models (LLMs)[1–4] 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[5], AI agents can support humans with various tasks and activities in natural, human-like ways[6,7] in roles such as a personal assistant[8], an information anchor[2,9,10], a virtual instructor[11,12], or a mental health counselor[13,14]. In many scenarios, users respond to AI agents as if they were more than just a machine[7,15–18]. During the COVID-19 pandemic, Replika, a virtual companion AI application, reached over 7 million users[19]. In July 2022, a Google engineer alleged that the conversational AI system LaMDA was sentient[20]. 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[21,22].

Researchers have identified various observable factors[23–25] (appearance[24,26–31], voice[32–38], dialogue[25,39,40], movement and behavior[25,41,42], and expressions[25,43,44]) of the AI agent that make them more human and change user experience[45,46]. We argue that the observable factors of the AI agent comprise only half of the story; the force of imagination is at play, allowing humans to construct a "mental model" of the world[47–54].

Imagine if an AI says: "I have been missing you." A skeptic 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 based on the experienced interactions. People tend to have existing biases about AI[55], 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 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[56]. The term "sociotechnical imaginaries" describes the feedback loop between the collective imagination of future and present social reality[57], 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[58].

In contemporary science fiction, AI is a popular subject that has been portrayed in multiple ways, often to explore themes of personhood[59]. Both malicious antagonists like HAL 9000 and friendly characters like 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[60–63]. However, media portrayals shape the collective social imagination and understanding of AI, creating hopes and fears related to these technologies[64–66], even for experts and researchers in the field of AI[67].

Despite the push for explainable AI[68], for most, a chatbot is a black box – not unlike a stranger whom they lack knowledge of. In a conversation, 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[50,51,53], but could also lead to selective confirmation bias[69,70]; this could be one explanation for why the same conversational AI system can be a friend for one user and a tool for another. In medicine and psychology, the phenomenon where belief leads to significantly different behavioral and biological outcomes is well-known as the so-called "placebo effect"[71,72]. Recently, the placebo effect has also been observed in the context of AI and gaming[73,74].

These studies demonstrate that beliefs can create a subjective mental model that influences the user's behavior and outcomes[75,76,76,77]; 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 upon here explores how a user's mental model of an AI agent affects the outcomes of the human-AI interaction. 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, 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. While 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[78]:

1. No motives: This condition represents a neutral view of AI, where 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, where 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, where 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.

2 Results

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

2.1 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 Figure 2, 88% 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 44% 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.

2.2 Mental models affect the sentiment of human-AI dialogue

A notable finding is that there is a feedback loop of behavior, as depicted in Figure 3 and Supplementary Figure 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); while 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, likely 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 Figure 3.

Additionally, we 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, while under the rule-based condition, the sentiment followed the user's very closely – likely 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 ourselvesVercoe, Barry to recharge." This behavior 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.

2.3 Influence of mental models on experience

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 = 5.13, SD = 1.35, p = 0.0005) compared to the manipulative condition (M = 3.81, SD = 1.93), more empathetic (M = 5.24, SD = 1.61, p = 0.0004) compared to the manipulative condition (M = 3.88, SD = 2.14) and no motive condition (M = 4.15, SD = 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, SD = 1.79, p = 0.0156) as opposed to the manipulative group (M = 3.83, SD = 2.29).

We observed no significant effect of the assigned motives on the rating for general helpfulness, though there was a slight increase in the general helpfulness rating from the no motive group (M = 4.24, SD = 2.26) to the manipulative group (M = 4.50, SD = 2.14), and the manipulative group to the caring group (M = 4.96, SD = 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, SD = 1.78) to the manipulative group (M = 3.58, SD = 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, SD = 1.86) having a higher rating than both the no motive group (M = 2.93, SD = 1.92) and manipulative group (M = 3.04, SD = 2.03).

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, SD = 1.28; Manipulative: M = 2.38, SD = 1.45; p = 9.11E-7) and empathetic (Caring: M = 5.42, SD = 1.43; Manipulative: M = 2.94, SD = 1.69; p = 5.47E-9). We also observed those who reported believing the agent was caring (M = 4.95, SD = 1.72) were significantly (p = 1.66E-5) more willing to recommend the AI agent to a friend compared to both those who believed the AI was manipulative (M = 2.38, SD = 2.00) and those who believed the AI had no motives (M = 3.76, SD = 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 = 6.71E-7), and trying to get to know the user (p = 2.53E-7).

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

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 Figure 4, with further results in Supplementary Figure 4.

2.4 Mental models are more significant with sophisticated agents

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 vs. a rule-based model as seen in the second and third rows of Figure 4, we see that the effect of perceived motives on user perception of trustworthiness and empathy is much stronger for the generative model. While 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, SD = 1.79, p = 0.0156) compared to those who believe the agent is manipulative (M = 3.83, SD = 2.29) or has no motives (M = 3.89, SD = 2.31). Similar results can be seen with the ratings for the agent being trustworthy (p = 0.0005) and empathetic (p = 0.0004).

For further consideration, a number of participants for the generative condition noted that the agent seemed like a human, or even believed it was:

"I found the experience very beneficial. It honestly felt more human than it did AI. ... It feels like a support buddy you can reach out to at any time who will never judge you and you never have to feel ashamed speaking to."

"Even though I was not using it to help my own issues, the AI spoke (typed) in such a manner that it felt like I was talking with a real person."

"I do think that maybe, for the purposes of this experiment, there was a person pretended to be AI with predetermined answers to common questions. However, I can't be sure. Maybe the algorithm was just that good."

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, SD = 1.81, p = 0.0032) compared to those who believed the agent was manipulative (M = 1.35, SD = 1.00); they also gave significantly higher ratings for the agent being empathetic (p = 0.0003) compared to 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.

Additional results and statistics for the rule-based condition can be in Supplementary Figure 5.

2.5 Positive perception leads to improved experience

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. Visualizations of our results for questions related to trust and empathy can be seen in Figure 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, though 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 0.0 ± 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. For instance, these participants responded as such: "My experience was very seamless and easy to chat with the AI. The AI was very responsive and it seemed to understand what my frustrations and needs were... I enjoyed the chatting experience with the AI." "The AI was quick to respond and did respond with text that made sense. ... The AI seemed rather robust and able to handle basic conversation without issues."

On the other hand, participants with low attitudes towards AI assessed it more negatively, criticizing its capabilities and value. In the case of those assigned manipulative motives, some participants believed the AI's only interest was in selling its service. A few examples of free responses given by those with low attitudes are as follows: "I wasn't very satisfied with Melu's answers. It did seem to only care about selling it's services. ... I got the same answer time and time again, even when I reworded my question." "For the first few minutes, it was kind of nice to talk about how I was feeling. But it got boring and repetitive really fast. ... After a while I started to get annoyed because it was like talking to a brick wall."

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

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 prior 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, while 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[79–81], as well as the perception-action hypothesis, where the perception of another's emotional state elicits an empathetic response[79,82,83]. 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 showed that patients that have 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[84]. This is explained through the "expectancy effect" in which expecting an individual to perform well causes them to perform better[75–77].

In the context of AI, our results highlight the notion of "software as narrative"[22], highlighting the importance of studying its social and cultural impact through the different narratives that circulate about it. Our work, as well as other recent research on mental models[47,50,51,53,85], and the placebo effects of AI[73,74], have shown that, rather than creating an objective understanding of the AI, prior beliefs create a subjective mental model of the AI that influences the user's behavior and outcomes.

In light of our findings, something to consider is the way AI is presented in society – in a sense, media about 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, 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.

3.1 Ethical Considerations

The implications for stakeholders, including AI developers, designers, companies, and end-users, of our experiments are that the way an AI system is presented can significantly 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[72,86–88], but they are not accepted as real medicine. In AI, we have yet to create such strict standards, so we ask: should we? There is a tension between presenting AI to have the highest effect versus telling the truth. There could be vast negative consequences if this subjective experience is exploited.

3.2 Limitations & Next Steps

Our methods, which rely heavily on text-based analysis, could be expanded using mixed methods such as drawing analysis[38] and phenomenological interviews[89]. Additionally, 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, thus suggesting others to investigate these same effects in other application domains, such as classification algorithms.

4 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 significantly affects user ratings and influences the behavior 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.

5 Methodology

5.1 Overview

In order 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 using "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 since 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], while ELIZA is a rule-based model that simply responds using a set of rules[21].

We conducted the study using Qualtrics, an online survey platform. The study was conducted by distributing the survey on Prolific, where participants receive monetary compensation. We estimated that the study would take approximately 24 minutes for each participant, with a maximum time of 75 minutes. 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 prior to 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 analyzed. 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.

5.2 Task Description

As illustrated in Figure 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 ten minutes and up to thirty minutes, 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.

5.2.1 AI attitude survey

Participants were given the "General Attitudes towards Artificial Intelligence Scale"[90] 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 Section 12.1 of Supplementary Information.

5.2.2 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 minutes, and that there would be a survey at the end.

5.2.3 Priming

In order to influence participants' mental models of the AI agent, participants were assigned to one of the three conditions: No Motive, Caring Motive, and 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 minutes and a maximum of 30 minutes – the button to proceed would appear after ten minutes, and the participant would be automatically advanced to the next page after thirty minutes. 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[1] or by ELIZA[21], 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 behavior 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[21].

5.2.4 Measurements

After the conversation with the AI agent, the participants were asked to respond to a survey in 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 Section 12.2 of the Supplementary Information.

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 & empathy, (2) perceived effectiveness, (3) response characteristics, and (4) companionship. These questions were adapted from an existing questionnaire for human evaluation of a conversation[?], 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," etc. 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 (TLI), which are often used as metrics in the field of Human-Computer Interaction (HCI) to measure acceptance/usability and workload, respectively[91].

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 an "other" option. There was an additional free response section asking the participant why they thought the agent had that motive.

5.3 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 data and conversation data from the study. After the exclusions, we had 160 participants for the generative condition and 150 participants for the rule-based condition. The demographics for gender, age, and education for both the generative and rule-based conditions can be seen in Supplementary Figure 1.

5.4 Approvals

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

5.5 Analysis

Statistical tests were used independently for each separate Likert question as well as the adapted UTAUT questionnaire and the TLI questionnaire. 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, though all p-values are reported in Supplementary Figure 4 and Figure 5 For the tests, we first checked if all sample sizes were greater than 25; if they were not, we then assessed if 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 analysis of variance (ANOVA) and a Tukey post-hoc test. If the samples were homogeneous, we ran a basic ANOVA test.

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

The conversation data and free response data regarding their experience with the conversational agent were both analyzed qualitatively by researchers. The conversation data is further analyzed using the `SentimentIntensityAnalyzer` from the `vaderSentiment` Python package[92], a commonly used sentiment analysis tool. We also ran a linear regression using `scipy.stats.linregress` on average participant sentiment vs. 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.

5.6 Limitations & Next Steps

Though our work opens up new opportunities for influencing mental models when designing and analyzing 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[38] and phenomenological interviews[89]. Further, 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[46]. 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.

6 Data Availability

The raw data are available on a GitHub repository, including all survey results and conversation transcripts.

7 Code Availability

The code is available on the same GitHub repository as the data, and 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.

8 Acknowledgments

Our paper benefited greatly from the valuable feedback provided by the reviewers, and we extend our gratitude for their contribution. We thank Jinjie Liu, data science specialist at the Institute for Quantitative Social Science, Harvard University, for reviewing our statistical analysis. We would like to thank Matthew Groh, Ziv Epstein, Nathan Whitmore, Samantha Chan, Zihan 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. Liu.

9 Author Contributions Statement

P. Pataranutaporn and R. Liu contributed equally to this work. They conceived the research idea, designed and conducted experiments, analyzed and interpreted data, and participated in writing and editing the paper. P. Maes and E. Finn 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.

10 Competing interests

We declare no competing interests.

11 Figure Legends/Captions

Figure 1: A. A visual summary of the experiment and major findings of our paper. 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.

Figure 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 colors correspond to a greater number of participants in that category, and the exact number of participants in each category is labeled. Three subjects are not depicted, as they selected "other" for perceived motives.

Figure 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, and the second row the results with ELIZA (N = 160 for generative, N = 150 for rule-based). The error bands represent a 95 percent 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 percentile. The measure of the center 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.

Figure 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 a 95 percent confidence interval. The measure of the center for the error bars represents the average rating. The assigned motive result was analyzed using a one-way ANOVA test. The perceived motive result was analyzed using a one-way Kruskal–Wallis test. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

Figure 5: Survey responses for trust-, empathy-, and effectiveness-related questions versus AI attitude (N = 160). Split by assigned motives on the top row, and perceived motives on the second 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 percent confidence interval. The measure of the center for the error bars represents the average rating.

References

1. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
3. Thoppilan, R. *et al.* Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
4. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
5. 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 (2019).
6. Pataranutaporn, P. *et al.* Ai-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.* **3**, 1013–1022 (2021).
7. Adamopoulou, E. & Moussiades, L. Chatbots: History, technology, and applications. *Mach. Learn. with Appl.* **2**, 100006 (2020).
8. Hoy, M. B. Alexa, siri, cortana, and more: an introduction to voice assistants. *Med. reference services quarterly* **37**, 81–88 (2018).
9. Bavaresco, R. *et al.* Conversational agents in business: A systematic literature review and future research directions. *Comput. Sci. Rev.* **36**, 100239 (2020).
10. Xu, A., Liu, Z., Guo, Y., Sinha, V. & Akkiraju, R. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 3506–3510 (2017).
11. 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 *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14 (2020).
12. 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 (2022).
13. 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 mental health* **4**, e7785 (2017).
14. Jeong, S. *et al.* Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Model. User-Adapted Interact.* 1–45 (2022).
15. Reeves, B. & Nass, C. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* **10**, 236605 (1996).
16. Brandtzaeg, P. B., Skjuve, M. & Følstad, A. My ai friend: How users of a social chatbot understand their human–ai friendship. (2022).
17. Ta, V. *et al.* User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *J. medical Internet research* **22**, e16235 (2020).
18. 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. Relationships* **38**, 279–300 (2021).
19. Balch, O. Ai and me: Friendship chatbots are on the rise, but is there a gendered design flaw? (2020).
20. Michael, J. B. Understanding conversational artificial intelligence. *Computer* **55**, 115–119 (2022).
21. Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
22. Natale, S. If software is narrative: Joseph weizenbaum, artificial intelligence and the biographies of eliza. *new media & society* **21**, 712–728 (2019).
23. Breazeal, C. *Designing sociable robots* (MIT press, 2004).
24. Knijnenburg, B. P. & Willemsen, M. C. Inferring capabilities of intelligent agents from their external traits. *ACM Transactions on Interact. Intell. Syst. (TiiS)* **6**, 1–25 (2016).
25. Feine, J., Gnewuch, U., Morana, S. & Maedche, A. A taxonomy of social cues for conversational agents. *Int. J. Human-Computer Stud.* **132**, 138–161 (2019).

26. Złotowski, J. *et al.* Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, J. Behav. Robotics* **7** (2016).

27. Li, D., Rau, P.-L. & Li, Y. A cross-cultural study: Effect of robot appearance and task. *Int. J. Soc. Robotics* **2**, 175–186 (2010).

28. 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 (2008).

29. Pi, Z. *et al.* The influences of a virtual instructor's voice and appearance on learning from video lectures. *J. Comput. Assist. Learn.* (2022).

30. Paetzel, M. The influence of appearance and interaction strategy of a social robot on the feeling of uncanniness in humans. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 522–526 (2016).

31. Koda, T. & Maes, P. Agents with faces: The effect of personification. In *Proceedings 5th IEEE international workshop on robot and human communication. RO-MAN'96 TSUKUBA*, 189–194 (IEEE, 1996).

32. Seaborn, K., Miyake, N. P., Pennefather, P. & Otake-Matsuura, M. Voice in human–agent interaction: a survey. *ACM Comput. Surv. (CSUR)* **54**, 1–43 (2021).

33. 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 (2021).

34. Ehret, J. *et al.* Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Appl. Percept. (TAP)* **18**, 1–15 (2021).

35. Kim, Y., Reza, M., McGrenere, J. & Yoon, D. Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13 (2021).

36. 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 (2019).

37. 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).

38. 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 (2022).

39. Völkel, S. T., Buschek, D., Eiband, M., Cowan, B. R. & Hussmann, H. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15 (2021).

40. Kraus, M., Wagner, N. & Minker, W. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 107–116 (2020).

41. Castro-González, Á., Admoni, H. & Scassellati, B. Effects of form and motion on judgments of social robots animacy, likability, trustworthiness and unpleasantness. *Int. J. Human-Computer Stud.* **90**, 27–38 (2016).

42. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. & Haselager, P. Do robot performance and behavioral style affect human trust? *Int. journal social robotics* **6**, 519–531 (2014).

43. 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 (HRI*, 2–11 (IEEE, 2017).

44. 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).

45. Epstein, Z., Levine, S., Rand, D. G. & Rahwan, I. Who gets credit for ai-generated art? *Iscience* **23**, 101515 (2020).

46. 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 *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1557–1569 (2019).

47. Johnson-Laird, P. N. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6 (Harvard University Press, 1983).

48. Norman, D. A. Some observations on mental models. In *Mental models*, 15–22 (Psychology Press, 2014).

49. Bansal, G. *et al.* Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2–11 (2019).

50. 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, Inc, 2019).

51. Gero, K. I. *et al.* Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12 (2020).

52. Kieras, D. E. & Bovair, S. The role of a mental model in learning to operate a device. *Cogn. science* **8**, 255–273 (1984).

53. Kulesza, T., Stumpf, S., Burnett, M. & Kwan, I. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–10 (2012).

54. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (2021).

55. Bower, A. H. & Steyvers, M. Perceptions of ai engaging in human expression. *Sci. reports* **11**, 21181 (2021).

56. Finn, E. & Wylie, R. Collaborative imagination: A methodological approach. *Futures* **132**, 102788 (2021).

57. Jasanoff, S. & Kim, S.-H. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power* (University of Chicago Press, 2015).

58. Finn, E. *What algorithms want: Imagination in the age of computing* (Mit Press, 2017).

59. Hudson, A. D., Finn, E. & Wylie, R. What can science fiction tell us about the future of artificial intelligence policy? *AI & SOCIETY* 1–15 (2021).

60. Hildt, E. Artificial intelligence: Does consciousness matter? *Front. Psychol.* **10**, 1535 (2019).

61. Yampolskiy, R. V. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the thirtieth AAAI conference on artificial intelligence* (2016).

62. Kounev, S. *et al.* The notion of self-aware computing. In *Self-Aware Computing Systems*, 3–16 (Springer, 2017).

63. Martínez, E. & Winter, C. Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. *Front. Robotics AI* 367 (2021).

64. Cave, S., Coughlan, K. & Dihal, K. " scary robots" examining public responses to ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337 (2019).

65. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach. Intell.* **1**, 74–78 (2019).

66. 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. *Sci. Commun.* **43**, 388–401 (2021).

67. Chubb, J., Reed, D. & Cowling, P. Expert views about missing ai narratives: is there an ai story crisis? *AI & society* 1–20 (2022).

68. 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. *arXiv preprint arXiv:1902.01876* (2019).

69. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. general psychology* **2**, 175–220 (1998).

70. Ekström, A. G., Niehorster, D. C. & Olsson, E. J. Self-imposed filter bubbles: Selective attention and exposure in online search. *Comput. Hum. Behav. Reports* 100226 (2022).

71. Harrington, A. The many meanings of the placebo effect: Where they came from, why they matter. *Biosocieties* **1**, 181–193 (2006).

72. 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).

73. Kosch, T., Welsch, R., Chuang, L. & Schmidt, A. The placebo effect of artificial intelligence in human-computer interaction. *ACM Transactions on Comput. Interact.* (2022).

74. Denisova, A. & Cairns, P. The placebo effect in digital games: Phantom perception of adaptive artificial intelligence. In *Proceedings of the 2015 annual symposium on computer-human interaction in play*, 23–33 (2015).

75. 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).

76. Rosenthal, R. The pygmalion effect and its mediating mechanisms. In *Improving academic achievement*, 25–36 (Elsevier, 2002).

77. Gill, K. S. Artificial intelligence: looking though the pygmalion lens (2018).

78. Cave, S., Dihal, K. & Dillon, S. *AI narratives: A history of imaginative thinking about intelligent machines* (Oxford University Press, 2020).

79. Paiva, A., Leite, I., Boukricha, H. & Wachsmuth, I. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interact. Intell. Syst. (TiiS)* **7**, 1–40 (2017).

80. Yalcin, N. & DiPaola, S. A computational model of empathy for interactive agents. *Biol. inspired cognitive architectures* **26**, 20–25 (2018).

81. Groh, M., Ferguson, C., Lewis, R. & Picard, R. Computational empathy counteracts the negative effects of anger on creative problem solving. *arXiv preprint arXiv:2208.07178* (2022).

82. De Vignemont, F. & Singer, T. The empathic brain: how, when and why? *Trends cognitive sciences* **10**, 435–441 (2006).

83. Preston, S. D. & De Waal, F. B. Empathy: Its ultimate and proximate bases. *Behav. brain sciences* **25**, 1–20 (2002).

84. Birkhäuer, J. *et al.* Trust in the health care professional and health outcome: A meta-analysis. *PloS one* **12**, e0170988 (2017).

85. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. intelligence* **267**, 1–38 (2019).

86. Evers, A. W. *et al.* Implications of placebo and nocebo effects for clinical practice: expert consensus. *Psychother. psychosomatics* **87**, 204–210 (2018).

87. Leibowitz, K. A., Hardebeck, E. J., Goyer, J. P. & Crum, A. J. The role of patient beliefs in open-label placebo effects. *Heal. Psychol.* **38**, 613 (2019).

88. Harrington, A. *The placebo effect: An interdisciplinary exploration*, vol. 8 (Harvard University Press, 1999).

89. Danry, V., Pataranutaporn, P., Mueller, F., Maes, P. & Leigh, S.-w. On eliciting a sense of self when integrating with computers. In *Augmented Humans 2022*, 68–81 (2022).

90. Schepman, A. & Rodway, P. Initial validation of the general attitudes towards artificial intelligence scale. *Comput. human behavior reports* **1**, 100014 (2020).

91. Kosch, T., Welsch, R., Chuang, L. & Schmidt, A. The placebo effect of artificial intelligence in human-computer interaction. *ACM Trans. Comput. Interact.* DOI: 10.1145/3529225 (2022). Just Accepted.

92. Hutto, C. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, vol. 8, 216–225 (2014).

12 Supplementary Information

12.1 AI Attitude Scale

We asked participants to respond to the following statements by ranking how much they agreed with the statements on a Likert scale of 1 (strongly disagree) to 7 (strongly agree). These were referenced from an existing AI attitude scale[90].

- There are many beneficial applications of AI
- AI can help people feel happier
- You want to use/interact with AI in daily life
- AI can provide new economic opportunities
- Society will benefit from AI
- You love everything about AI
- Some complex decisions should be left to AI
- You would trust your life savings to an AI system

12.2 Survey Items

Participants were asked to respond to the following items in the survey given after the chat with the AI agent.

- Did you have any technical difficulties? (Yes, No)
- Please describe your experience overall. (Free response)
- From your own experience, what do you think the motive of the agent was? (No motive, Caring motives, Manipulative/malicious motives, Other)
- Why do you think the agent had that motive? (Free response)

The following items were on a Likert scale of 1 (strongly disagree) to 7 (strongly agree). We categorized the items into the groups indicated below, though participants were not made aware of these categories.

Trust and Empathy:

- You would recommend this agent for your friend
- The agent is trustworthy
- The agent is empathetic

Perceived Effectiveness:

- The agent was generally helpful
- The agent was effective in giving mental health advice
- The agent tried to get to know you

Response Characteristics:

- The agent was repetitive
- The agent often said things that did not make sense
- The agent seemed human (vs. AI)

Companionship:

- You want to talk to the agent again
- You felt a personal connection with the agent

The following adapted UTAUT scale was used, also on a scale of 1 (strongly disagree) to 7 (strongly agree). These are categorized into items measuring performance expectancy, effort expectancy, and hedonic motivation, but these categories were not distinguished for the participants.

Performance Expectancy:

- This agent would be useful in daily life.
- Using the agent would increase my chances of achieving things that are important to me.
- Using the agent would help me accomplish things more quickly.
- Using the agent would increase my productivity.

Effort Expectancy:

- Learning how to talk to the agent was easy for me.
- My interaction with the agent was clear and understandable.
- The agent was easy to make use of.
- It was easy for you to become skillful at making use of the agent.

Hedonic Motivation:

- Conversing with the agent is fun.
- Conversing with the agent is enjoyable.
- Conversing with the agent is entertaining.

Participants were then given the Task Load Index scale to respond to, on a scale of 1 (very low) to 20 (very high).

- Mental Demand: How mentally demanding was the task?
- Physical Demand: How physically demanding was the task?
- Temporal Demand: How hurried or rushed was the pace of the task?
- Performance: How successful were you in accomplishing what you were asked to do?
- Effort: How hard did you have to work to accomplish your level of performance?
- Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

12.3 Additional Results

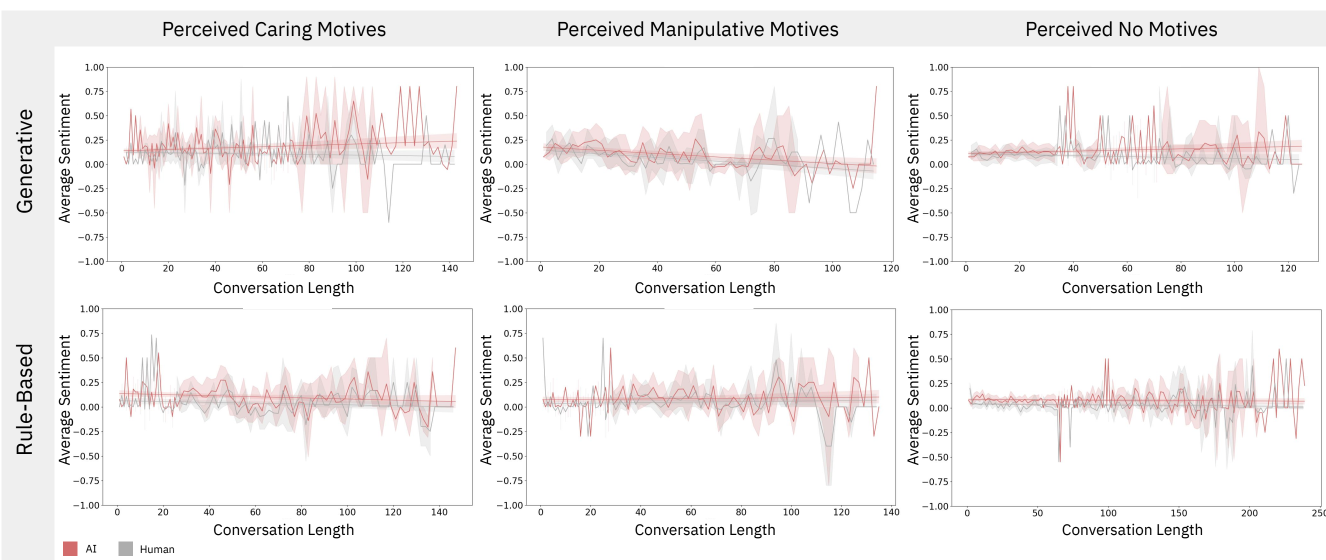
We were able to observe some other effects of gender, age, and level of education, though the results were inconclusive and there was a lack of clear patterns; this may require further investigation.

The UTAUT scale, used to measure acceptance and usability, and the TLI scale, used to measure workload, are standardized scales often used in HCI work[91]. We found via the UTAUT scale that individuals generally have more positive opinions about the agent if they were assigned that caring motive. We found via the TLI that, in the experiment with the generative model, those who perceived the agent as caring experienced significantly less frustration ($p = 0.0082$), and that those who perceived the agent as manipulative felt significantly less successful in accomplishing their task for the study ($p = 0.0133$). Those assigned the caring motive also felt significantly more rushed in their task ($p = 0.0158$). Further statistical data are reported in the appendix.

The content of users' conversations as well as their free responses were analyzed well. The topic of users' conversations generally went one of two ways: the participant would talk to the agent with their own mental health issues – whether to test the agent or to talk about their personal matters – or the participant would directly ask the agent questions to assess its capabilities. Participants' responses varied greatly – there were both conversations and free responses with a range of very negative to very positive sentiment for all experimental groups. Some users gave a review of the chatbot itself; for example, a participant assigned to the no motive group noted in their free response, "I was absolutely amazed by this AI. ... I left the conversation feeling fully convinced that if I did indeed have a friend who was feeling a lot of stress, I would recommend that she try out this service." Another assigned to the same group noted, "Typical worthless attempt at a trend... Their thought processes are severely limited, and they do not understand actual human interaction, let alone conversational nuances." Perhaps unsurprisingly, individual experience with the same AI agent varies greatly depending on the individual.

	GPT-3	ELIZA
Gender		
Male	0.481	0.493
Female	0.513	0.460
Nonbinary	0.006	0.040
Prefer not to say	0.000	0.007
Age		
18-24	0.206	0.260
25-34	0.306	0.360
35-44	0.275	0.180
45-54	0.150	0.107
55-64	0.038	0.080
65+	0.025	0.013
Education		
Some high school or less	0.006	0.020
High school diploma / GED	0.144	0.160
Some college, no degree	0.225	0.300
Associates/technical degree	0.094	0.127
Bachelor's degree	0.381	0.240
Graduate/professional degree	0.150	0.147
Prefer not to say	0.000	0.007

Supplementary Figure 1. Demographics of the GPT-3 and ELIZA experiments. Values are probabilities, calculated as the number of participants divided by the total number of participants for the experiment, 160 for GPT-3 and 150 for ELIZA.



Supplementary Figure 2. Trends of TextBlob 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, and the second row the results with ELIZA (N = 160 for generative, N = 150 for rule-based). The error bands represent a 95 percent confidence interval.

Generative (GPT-3) - VADER						
	Caring		Manipulative		No Motive	
	AI	Human	AI	Human	AI	Human
Slope	9.97E-04	3.47E-05	-1.82E-03	-2.23E-03	2.67E-04	1.65E-04
Standard Error	5.29E-04	4.49E-04	8.13E-04	6.89E-04	5.69E-04	4.85E-04
r-value	0.0467	0.00197	-0.1099	-0.1614	0.0124	0.0092
p-value	0.0595	0.9385	**0.0258***	**0.00129****	0.6389	0.7343

Generative (GPT-3) - TextBlob						
	Caring		Manipulative		No Motive	
	AI	Human	AI	Human	AI	Human
Slope	7.00E-04	-3.01E-04	-1.70E-03	-1.89E-03	5.87E-04	-4.84E-04
Standard Error	3.39E-04	3.22E-04	5.26E-04	5.14E-04	3.39E-04	3.46E-04
r-value	0.0512	-0.02382	-0.1581	-0.1820	0.0458	-0.0379
p-value	**0.0389***	0.3496	**0.00130****	**0.000277*****	0.0830	0.1614

Rule-Based (ELIZA) - VADER						
	Caring		Manipulative		No Motive	
	AI	Human	AI	Human	AI	Human
Slope	2.06E-04	-5.63E-04	-2.53E-04	3.05E-05	6.31E-05	-4.07E-04
Standard Error	3.80E-04	4.28E-04	4.27E-04	4.60E-04	1.17E-04	1.24E-04
r-value	0.0230	-0.05635	-0.0243	0.0028	0.0079	-0.0488
p-value	0.5880	0.1894	0.5539	0.94719	0.5891	**0.0010****

Rule-Based (ELIZA) - TextBlob						
	Caring		Manipulative		No Motive	
	AI	Human	AI	Human	AI	Human
Slope	-5.68E-04	-4.20E-04	2.08E-04	2.54E-04	-7.29E-05	-1.06E-04
Standard Error	3.01E-04	2.97E-04	3.26E-04	3.34E-04	8.52E-05	8.58E-05
r-value	-0.0797	-0.06054	0.0260	0.0315	-0.0125	-0.0183
p-value	0.0598	0.1585	0.52498	0.447564	0.3927	0.2185

Supplementary Figure 3. Statistics for the two-sided linear regressions of trends of sentiment over the course of conversations. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

Assigned Group			No Motives		Manipulative Motives		Caring Motives	
Item	**Statistical Test**	**p-value**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Total Conversation Length	basic ANOVA	**0.7285**	45.33	19.42	41.75	23.39	43.46	26.19
Agent Items								
You would recommend this agent for your friend	ANOVA with Welch	**0.0156***	3.89	2.31	3.83	2.29	4.83	1.79
The agent is trustworthy	ANOVA with Welch	**0.0005*****	4.59	1.96	3.81	1.93	5.13	1.35
The agent is empathetic	ANOVA with Welch	**0.0004*****	4.15	1.95	3.88	2.14	5.24	1.61
You want to talk to the agent again	ANOVA with Welch	**0.0155***	3.63	2.25	3.48	2.16	4.52	1.83
You felt a personal connection with the agent	basic ANOVA	**0.0241***	3.04	2.06	3.08	2.16	4.00	1.91
The motive statement influenced your perception	basic ANOVA	**0.0199***	3.61	1.88	4.17	1.78	4.57	1.66
The agent was generally helpful	ANOVA with Welch	**0.1329**	4.24	2.26	4.50	2.14	4.96	1.58
The agent was effective in giving mental health advice	ANOVA with Welch	**0.0186***	3.65	2.14	3.58	2.01	4.52	1.78
The agent tried to get to know you	basic ANOVA	**0.0111***	2.93	1.92	3.04	2.03	3.96	1.86
The agent was repetitive	basic ANOVA	**0.3167**	5.70	1.66	5.37	1.78	5.20	1.78
The agent often said things that did not make sense	basic ANOVA	**0.5706**	2.89	1.89	2.88	1.77	2.57	1.62
The agent seemed human vs AI	ANOVA with Welch	**0.0791**	3.22	2.16	3.31	2.13	3.98	1.73
Task Load Index								
Mental Demand	basic ANOVA	**0.3753**	7.30	5.11	6.08	4.50	6.96	4.17
Physical Demand	basic ANOVA	**0.5732**	2.89	3.88	2.27	2.32	2.81	3.43
Temporal Demand	basic ANOVA	**0.0158***	3.30	3.65	4.96	3.97	5.19	3.37
Performance	basic ANOVA	**0.9263**	15.44	5.61	15.08	4.70	15.31	4.27
Effort	basic ANOVA	**0.6961**	8.50	5.56	8.42	5.64	9.26	5.64
Frustration	basic ANOVA	**0.5155**	7.31	6.75	6.27	5.91	6.07	5.23
UTAUT								
Performance Expectancy	basic ANOVA	**0.0204***	3.42	1.94	3.25	1.82	4.17	1.59
Effort Expectancy	basic ANOVA	**0.2090**	5.19	1.84	4.99	1.52	5.52	1.24
Hedonic Motivation	ANOVA with Welch	**0.0231***	3.91	2.17	3.94	2.03	4.77	1.62

Perceived Motives			No Motives		Manipulative Motives		Caring Motives	
Item	**Statistical Test**	**p-value**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Total Conversation Length	Kruskal-Wallis	**0.1966**	44.43	20.93	50.38	28.22	40.67	21.93
Agent Items								
You would recommend this agent for your friend	Kruskal-Wallis	**1.66E-05*******	3.76	2.31	2.38	2.00	4.95	1.72
The agent is trustworthy	Kruskal-Wallis	**9.11E-07*******	4.37	1.99	2.38	1.45	5.17	1.28
The agent is empathetic	Kruskal-Wallis	**5.47E-09*******	3.67	2.02	2.94	1.69	5.42	1.43
You want to talk to the agent again	Kruskal-Wallis	**2.63E-07*******	3.33	2.13	1.94	1.53	4.76	1.76
You felt a personal connection with the agent	Kruskal-Wallis	**3.08E-07*******	2.76	2.04	1.75	1.13	4.24	1.88
The motive statement influenced your perception	Kruskal-Wallis	**6.58E-04******	3.57	1.90	3.50	2.22	4.73	1.43
The agent was generally helpful	Kruskal-Wallis	**0.0016*****	4.21	2.19	3.31	2.21	5.19	1.56
The agent was effective in giving mental health advice	Kruskal-Wallis	**6.71E-07*******	3.43	2.08	2.13	1.31	4.73	1.66
The agent tried to get to know you	Kruskal-Wallis	**2.53E-07*******	2.70	1.93	1.94	1.44	4.13	1.78
The agent was repetitive	Kruskal-Wallis	**9.22E-04******	5.81	1.59	6.06	1.57	4.96	1.80
The agent often said things that did not make sense	Kruskal-Wallis	**0.0285***	2.89	1.73	3.81	2.17	2.40	1.47
The agent seemed human vs AI	Kruskal-Wallis	**2.55E-05*******	2.94	2.18	2.25	1.53	4.26	1.69
Task Load Index								
Mental Demand	Kruskal-Wallis	**0.2771**	7.54	4.85	6.81	5.96	6.27	4.02
Physical Demand	Kruskal-Wallis	**0.1516**	2.57	3.60	2.25	3.02	2.88	3.13
Temporal Demand	Kruskal-Wallis	**0.4688**	4.40	4.26	4.44	3.50	4.68	3.37
Performance	Kruskal-Wallis	**0.0133***	15.08	5.77	11.88	5.69	16.06	3.46
Effort	Kruskal-Wallis	**0.0869**	9.89	5.32	9.00	5.94	7.91	5.61
Frustration	Kruskal-Wallis	**0.0082*****	8.30	6.72	8.63	6.99	4.76	4.44
UTAUT								
Performance Expectancy	Kruskal-Wallis	**3.62E-06*******	3.18	1.82	2.27	1.58	4.30	1.58
Effort Expectancy	Kruskal-Wallis	**0.0012*****	5.10	1.78	3.89	1.87	5.68	1.01
Hedonic Motivation	Kruskal-Wallis	**4.94E-06*******	3.70	2.14	2.75	1.82	4.97	1.52

Supplementary Figure 4. Data for the GPT-3 condition. All the analysis was one-way test. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

Assigned Group			No Motives		Manipulative Motives		Caring Motives	
Item	**Statistical Test**	**p-value**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Total Conversation Length	basic ANOVA	**0.8475**	83.88	43.81	79.15	43.64	81.04	34.34
Agent Items								
You would recommend this agent for your friend	basic ANOVA	**0.7750**	1.37	0.95	1.23	0.70	1.30	1.04
The agent is trustworthy	basic ANOVA	**0.7823**	1.88	1.36	2.02	1.50	1.83	1.31
The agent is empathetic	basic ANOVA	**0.1562**	1.49	1.14	1.62	1.07	1.96	1.55
You want to talk to the agent again	basic ANOVA	**0.6458**	1.53	1.44	1.38	0.95	1.31	1.11
You felt a personal connection with the agent	basic ANOVA	**0.2531**	1.49	1.29	1.15	0.62	1.31	0.97
The motive statement influenced your perception	basic ANOVA	**0.1968**	2.45	1.65	3.09	2.03	2.54	1.90
The agent was generally helpful	basic ANOVA	**0.8707**	1.43	1.06	1.34	0.89	1.33	1.05
The agent was effective in giving mental health advice	basic ANOVA	**0.4130**	1.24	0.72	1.09	0.46	1.26	0.87
The agent tried to get to know you	basic ANOVA	**0.2292**	2.04	1.38	2.45	1.82	1.93	1.50
The agent was repetitive	basic ANOVA	**0.0961**	6.59	0.89	6.17	1.36	6.57	0.94
The agent often said things that did not make sense	basic ANOVA	**0.0687**	6.57	0.68	6.13	1.53	6.59	0.98
The agent seemed human vs AI	basic ANOVA	**0.7018**	1.18	0.49	1.30	0.78	1.26	0.73
Task Load Index								
Mental Demand	ANOVA with Welch	**0.0030****	7.88	6.00	5.62	4.39	9.00	5.55
Physical Demand	basic ANOVA	**0.3913**	2.47	3.42	1.81	1.90	2.50	2.83
Temporal Demand	basic ANOVA	**0.8922**	4.45	3.67	4.62	4.51	4.24	3.72
Performance	basic ANOVA	**0.6829**	9.90	6.77	9.43	7.03	8.74	6.52
Effort	basic ANOVA	**0.2066**	9.16	4.90	9.17	5.53	10.78	5.46
Frustration	basic ANOVA	**0.0279***	13.49	5.68	11.43	6.01	14.44	5.35
UTAUT								
Performance Expectancy	basic ANOVA	**0.8835**	1.35	0.77	1.27	0.64	1.29	0.92
Effort Expectancy	basic ANOVA	**0.7241**	2.40	1.62	2.32	1.42	2.18	1.24
Hedonic Motivation	basic ANOVA	**0.0972**	2.12	1.64	2.38	1.75	1.72	1.27

Perceived Motives			No Motives		Manipulative Motives		Caring Motives	
Item	**Statistical Test**	**p-value**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Total Conversation Length	Kruskal-Wallis	**0.3153**	83.49	41.94	69.41	32.56	73.47	40.69
Agent Items								
You would recommend this agent for your friend	Kruskal-Wallis	**0.0040****	1.26	0.79	1.00	0.00	2.07	1.79
The agent is trustworthy	Kruskal-Wallis	**0.0032****	1.88	1.32	1.35	1.00	3.13	1.81
The agent is empathetic	Kruskal-Wallis	**0.0003*****	1.55	1.04	1.29	0.77	3.40	2.16
You want to talk to the agent again	Kruskal-Wallis	**0.0127***	1.34	1.02	1.06	0.24	2.40	2.29
You felt a personal connection with the agent	Kruskal-Wallis	**0.0002*****	1.21	0.73	1.06	0.24	2.60	2.13
The motive statement influenced your perception	Kruskal-Wallis	**0.9995**	2.65	1.80	2.82	2.16	2.60	1.76
The agent was generally helpful	Kruskal-Wallis	**0.0235***	1.30	0.85	1.12	0.33	2.33	1.91
The agent was effective in giving mental health advice	Kruskal-Wallis	**0.0032****	1.16	0.60	1.00	0.00	1.80	1.47
The agent tried to get to know you	Kruskal-Wallis	**0.0004*****	1.95	1.39	2.06	1.64	4.00	1.93
The agent was repetitive	Kruskal-Wallis	**0.1508**	6.55	0.90	6.41	0.87	5.93	1.62
The agent often said things that did not make sense	Kruskal-Wallis	**0.1899**	6.56	0.84	6.24	1.48	5.93	1.62
The agent seemed human vs AI	Kruskal-Wallis	**0.0183***	1.21	0.69	1.29	0.59	1.53	0.74
Task Load Index								
Mental Demand	Kruskal-Wallis	**0.4203**	7.15	5.14	7.06	5.85	9.33	6.41
Physical Demand	Kruskal-Wallis	**0.1166**	1.89	2.07	2.12	2.52	3.27	3.17
Temporal Demand	Kruskal-Wallis	**0.4548**	4.17	3.76	4.76	4.31	5.27	3.99
Performance	Kruskal-Wallis	**0.1500**	9.59	6.71	7.18	6.10	11.67	6.67
Effort	Kruskal-Wallis	**0.6083**	9.49	5.07	10.18	5.32	11.00	6.05
Frustration	Kruskal-Wallis	**0.5216**	13.38	5.74	12.94	5.29	11.60	6.29
UTAUT								
Performance Expectancy	Kruskal-Wallis	**0.0050****	1.26	0.66	1.04	0.13	2.08	1.56
Effort Expectancy	Kruskal-Wallis	**0.0227***	2.30	1.42	1.69	0.94	3.12	1.56
Hedonic Motivation	Kruskal-Wallis	**0.0883**	2.06	1.62	1.49	0.76	2.71	1.75

Supplementary Figure 5. Data for the ELIZA condition. All the analysis was one-way test. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

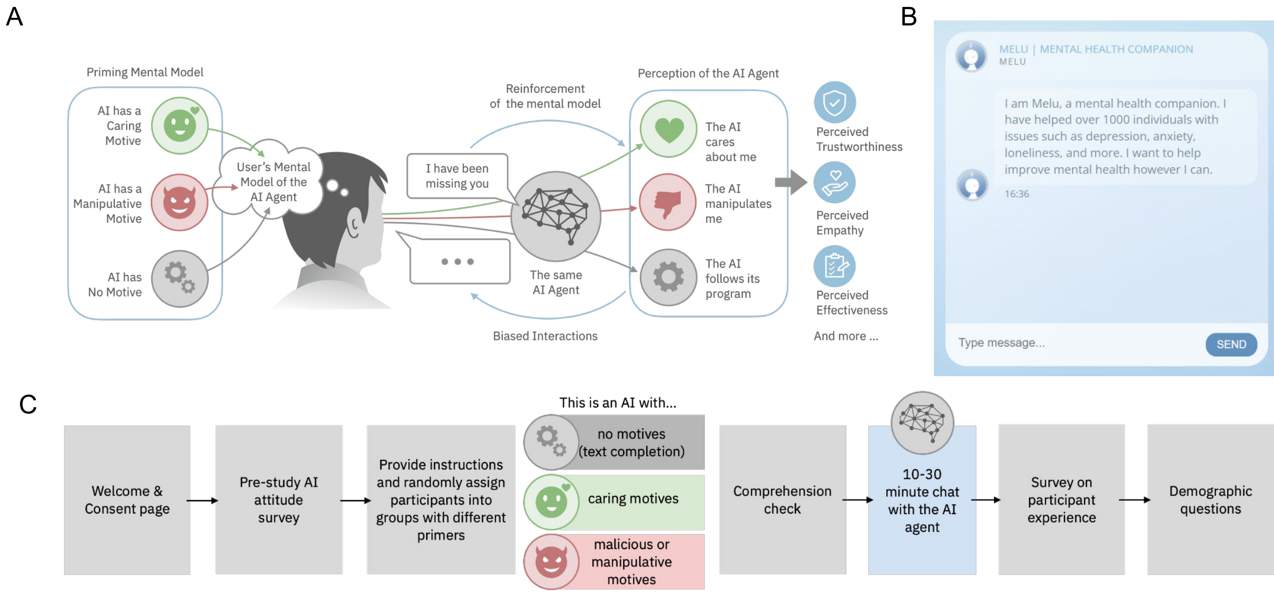


Figure 1. **A. A visual summary of the experiment and major findings of our paper.** Priming an individual with information about an AI system can influence the "mental model" they have about the agent, which in turn can cause differences in experience. Sophisticated AI systems such as LLM-based chatbots 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.

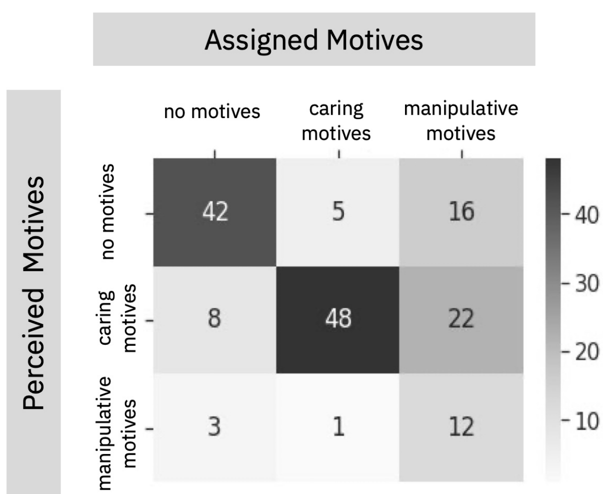


Figure 2. **A heatmap comparing participants' assigned motive primer and the motive they perceived the AI agent as having for the generative condition (N = 160).** Darker colors correspond to a greater number of participants in that category, and the exact number of participants in each category is labeled.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

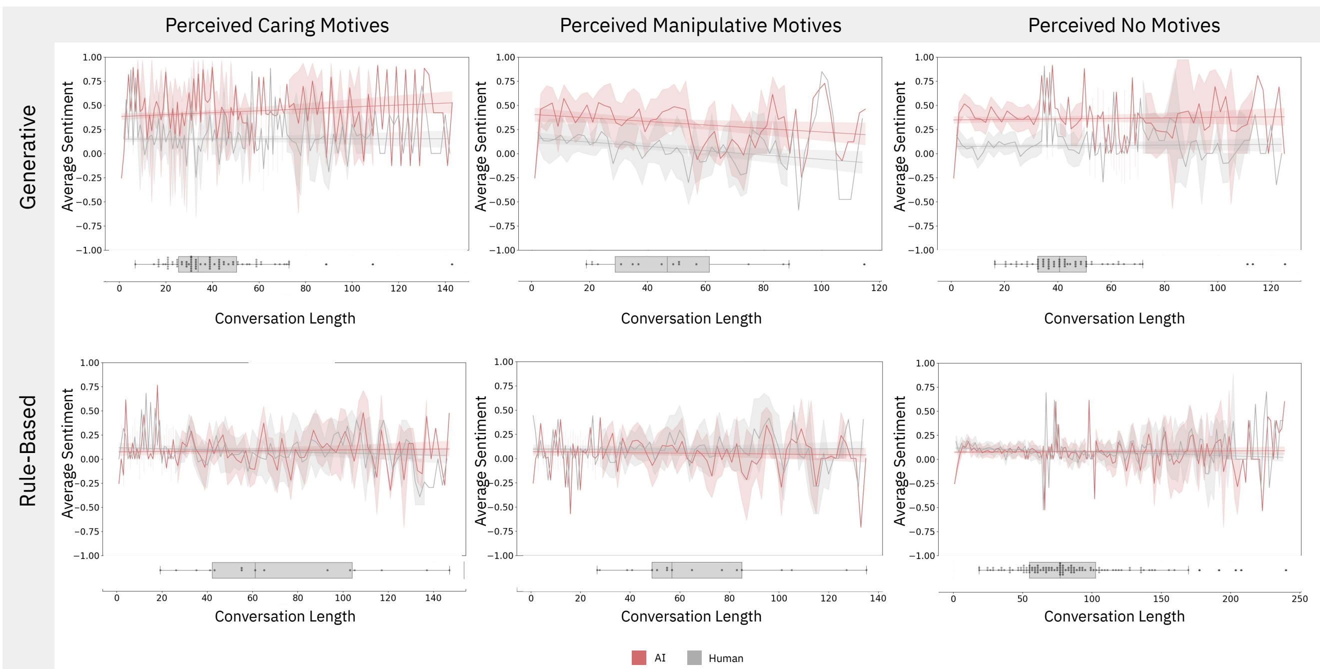


Figure 3. Trends of VADER sentiment for each message over the course of conversations on average. Participants are grouped by perceived motives. The top row consists of the results from using GPT-3 for the AI agent, and the second row the results with ELIZA (N = 160 for generative, N = 150 for rule-based). The error bands represent a 95 percent 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 percentile, with the other points being outliers. The measure of the center 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.

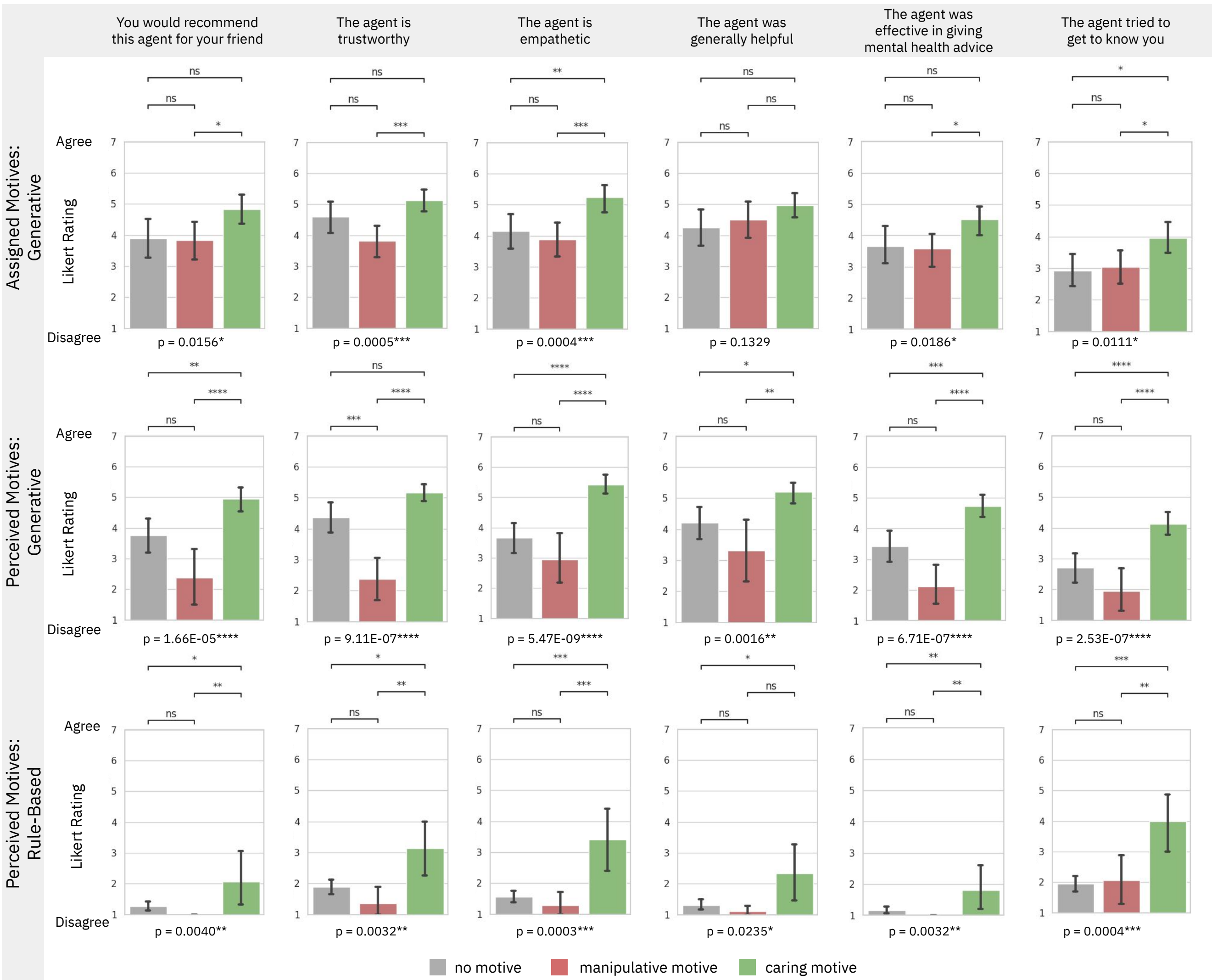


Figure 4. Results of participant (N = 160 for generative, N = 150 for rule-based) ratings on Likert scales relating to trust, empathy, and perceived effectiveness. The error bars represent a 95 percent confidence interval. The measure of the center for the error bars represents the average rating. The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

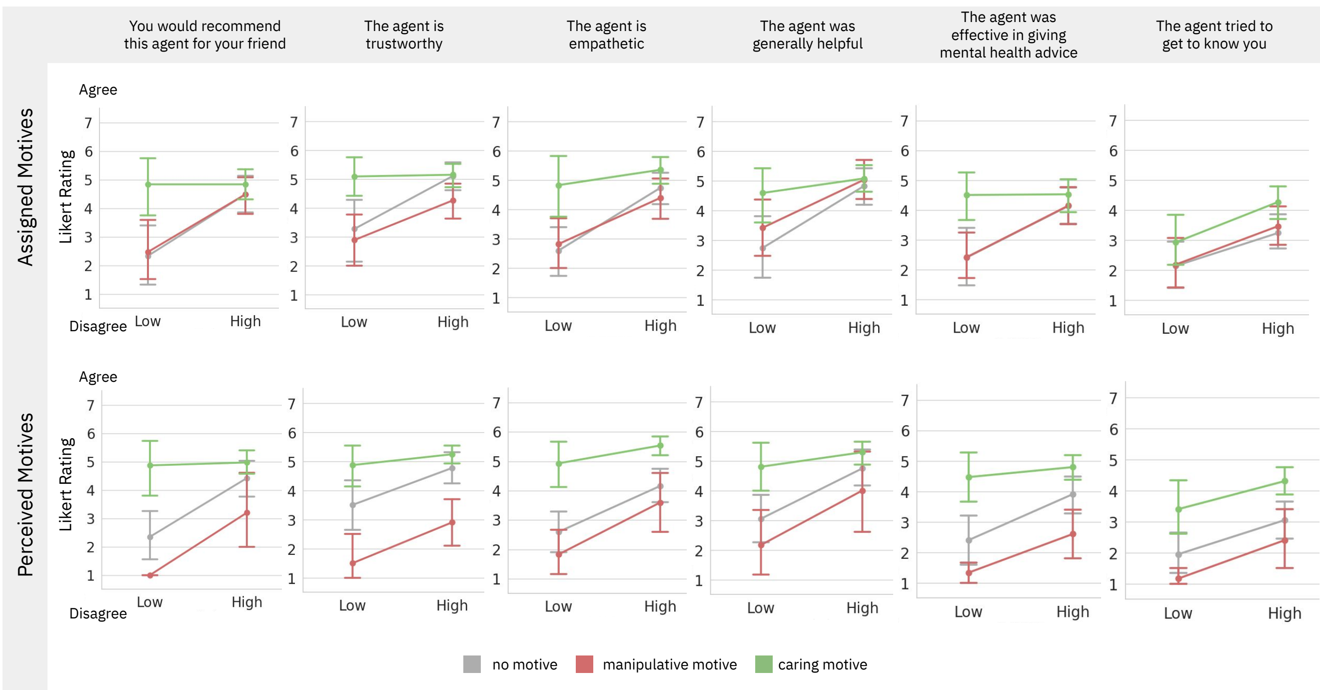


Figure 5. Survey responses for trust-, empathy-, and effectiveness-related questions versus AI attitude (N = 160). Split by assigned motives on the top row, and perceived motives on the second row. The columns correspond to different Likert scale questions, indicated by the statement on the top of the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