Transparency Enhances Positive Perceptions of Social Artificial Intelligence

Ying Xu[1], Nora Bradford[2], Radhika Garg[3]

[1] Marsal Family School of Education, University of Michigan, United States of America

[2] Department of Cognitive Sciences, University of California Irvine, United States of America

[3] Independent Researcher Morristown United States of America

Manuscript Accepted by *Human Behavior and Emerging Technologies*

Author Note

This work is supported by the National Science Foundation under Grant No. 2115382.

We have no conflicts of interests to disclose.

The data and the materials that support the findings of this study will be openly available in Deep Blue Data at https://doi.org/10.7302/69h3-x918

Correspondence concerning this article should be addressed to Ying Xu, 610 E. University Ave. Ann Arbor, Michigan 48109-1259. Email: yxying@umich.edu

Abstract

Social chatbots are designed to build emotional bonds with users, and thus it is particularly important to design these technologies so as to elicit positive perceptions from users. In the current study, we investigate the impacts transparent explanations of chatbots' mechanisms have on users' perceptions of the chatbots. A total of 914 participants were recruited from Amazon Mechanical Turk. They were randomly assigned to observe conversation between a hypothetical chatbot and user in one of the two-by-two experimental conditions: whether the participants received an explanation about how the chatbot was trained and whether the chatbot was framed as an intelligent entity or a machine. A fifth group, who believed they were observing interactions between two humans, served as a control. Analyses of participants' responses to post-observation survey indicated that transparency positively affected perceptions of social chatbots by leading users to (1) find the chatbot less creepy, (2) feel greater affinity to the chatbot, and (3) perceive the chatbot as more socially intelligent, thought these effects were small. Importantly, transparency appeared to have a larger effect in increasing the perceived social intelligence among participants with lower prior AI knowledge. These findings have implications for the design of future social chatbots and support the addition of transparency and explanation for chatbot users.

Keywords: Transparency, perception, artificial intelligence, chatbot

Transparency Enhances Positive Perceptions of Social Artificial Intelligence

As artificial intelligence (AI) progresses, the potential for social and emotional bonds with technological entities, specifically *social chatbots*, emerges. While other types of chatbots tend to serve a specific purpose like aiding the user in ordering food, buying a plane ticket, or receiving recommendations for healthcare options, social chatbots are designed to engage users in ongoing, personal, and empathetic conversations, providing emotional support, tailored advice, and a comfortable space for self-disclosure (Brandtzaeg et al., 2022; Lee et al., 2020; Wang et al., 2021). As defined by Shum et al. (2018), social chatbots "take time to converse like a human, present results, offering perspectives, prompting new topics to keep the conversation going" (p.13). It is worth mentioning that while chatbots like Replika are specifically designed for social purposes, large language models like ChatGPT also have the ability to engage in social interactions, albeit with a higher degree of versatility. Both types of chatbots can be considered as examples of social chatbots. When designed properly, social chatbots could possibly enhance individuals' well-being, particularly when alternative interpersonal interactions are limited or inaccessible (for a review, see van Wezel et al., 2021).

The success of a social chatbots hinges upon the extent to which people *perceive* the AI as a friendly and engaging conversation partner. The literature has suggested that humans' interaction with AI can potentially evoke both positive and negative perceptions, manifesting as feelings of charisma or creepiness (e.g., Bae Brandtzæg et al., 2021; Pentina et al., 2023). These perceptions, influenced by factors such as AI's ability to simulate human-like conversations, intersect with the inherent opacity of social chatbots and other AI systems. Users are generally unaware of what is happening between their own input (what they say to the chatbots) and the system's output (how the chatbot responds). The opacity of social AI systems can lead to users

feeling manipulated or form inappropriate attachments with the technology especially when social AI is designed to build long-term bonds with its users (Helbing, 2019). As a response to these ethical risks, the research community has actively advocated for AI transparency and emphasized the value of providing users with sufficient information about how AI works and what it is capable of. In this sense, transparency is connected to the disclosure of information (Tielenburg, 2018). However, the consensus for transparency has not yet been fully translated into common industrial practices, and technology companies rarely inform their users of how their AI systems engage in social interactions. Instead, these companies often capitalize on users' anthropomorphism tendencies by framing chatbots as agentic entities, such as Replika's promotion of the chatbot as "a friend who always listens." (Pentina et al., 2023, p.3).

The literature has suggested that people's perceptions of AI in general are malleable, and designs that promote transparency within AI systems have an impact on people's perceptions. However, the current findings on this topic are inconclusive and lack clear directions. On the one hand, some studies suggest that opacity actually makes social chatbots more personal and charming to some users, as people tend to treat AI systems more like people when the algorithmic mechanisms are made invisible (Lee et al., 2020). This is particularly important in fueling productive "social" interactions (Chaves & Gerosa, 2021), which can be characterized as mutual understanding, positive relationships, shared ideas, and reciprocal exchanges (Turkle, 2016). On the other hand, some studies support the benefits of transparency, suggesting that transparency will not only make people feel more empowered when interacting with AI but also mitigate some of their negative "uncanny" reactions to AI (Liu, 2021). These studies, for example, found that if the chatbots' mechanisms and capacities are unknown to its users, people

sometimes perceive these highly personal chatbots as creepy and invasive (e.g., Williams et al., 2015).

However, the majority of these studies have focused on the transparency of the decision-making process of task-oriented AI and its subsequent influence on user perceptions, particularly regarding usefulness and trust towards the AI. Yet there is significantly less research focused on social AI which aims to establish meaningful interactions and relationships rather than solely accomplishing tasks, despite that, the pursuit of social purposes remains one of the primary reasons people engage with AI (e.g., Bae Brandtzæg et al., 2021; Pentina et al., 2023). Furthermore, previous studies have operationalized transparency in vastly different ways, with most of them focusing on explanations of AI's decisions or actions during the interactions. Less attention has been paid to informing users about the AI's general inner workings with the goal of establishing users' expectations and comprehension of the AI's overall behavioral patterns (Chaves & Gerosa, 2021).

To address these gaps, this paper directly examines how providing users with explanation of an AI chatbot's mechanisms can affect their perceptions, both positive and negative, of the chatbot. In addition, we also investigated users' perceptions of social intelligence and agency in AI, focusing on their ability to effectively navigate and manage social situations. Through a randomized experiment with 914 participants, we tested the effects of transparency on three perception outcomes: perceived creepiness, affinity, and perceived social intelligence. Our results indicate that transparency positively affects perceptions of social chatbots by causing users to (1) find the chatbot less creepy, (2) feel greater affinity to the chatbot, and (3) perceive the chatbot as more socially intelligent, thought the small effect sizes warrant future research to examine the robustness of the findings.

Literature Review

People's Perceptions of Transparent AI Systems

Humans' interaction with social chatbots, as well as other AI systems, can induce a range of different perceptions or emotional reactions from human users, ranging from surprise, amazement, happiness, amusement, unease, to confusion (Shank et al., 2019). There has been a growing body of studies that have focused on different approaches to improving user experience and perceptions with social chatbots, the overwhelming majority of which have investigated the chatbot's voice, embodiment, and communication styles during the interaction (e.g., Go & Sundar, 2019; Jiang et al., 2023; Konya-Baumbach, 2022), yet only a few studies have focused on the influence of chatbots' transparent design on users' perceptions.

As AI technologies grow increasingly sophisticated and complex, the research community is dedicated to ensuring that people feel empowered and in control when interacting with these enigmatic "black box" systems. The debate over whether social AI is inherently deceptive has persisted, as AI-driven machines may lead other agents to perceive or behave as if the machine is human (Sharkey & Sharkey, 2021). This potential for users to anthropomorphize technologies could leave them vulnerable to emotional exploitation such as over-trust or other risks (Crockett et al., 2019). However, researchers suggest that transparency may be a solution to this dilemma, such as disclosing non-human status, revealing capabilities, or utilizing explainable AI, though some argue that many social AI benefit from some level of deception as it facilitates interactions with humans (Coeckelbergh, 2022). In our paper, we operationalize transparency as the disclosure of information regarding AI algorithms' inner workings, enabling users to better comprehend the output of AI systems.

Empirical Evidence on the Perception Outcomes of Transparency

The broader emphasis on AI transparency has motivated empirical work on transparent and explainable interfaces. This line of research evaluates different methods for increasing transparency in a variety of contexts, while more studies have focused on AI systems that are task-oriented (e.g., recommender systems, expert/knowledge-based systems, virtual assistants) rather than social-oriented (e.g., social chatbots in this study). These studies provide insights into the implications transparency has on people's perceptions of AI (e.g., Cramer et al., 2008; de Graaf et al., 2018; Eiband et al., 2018; Norman, 2009; Reddick et al., 2017; Rosenthal et al., 2016; Schmidt & Herrmann, 2017).

In terms of task-oriented systems, there has been strong evidence that transparency enhances people's confidence in the system's decision-making as well as their user experience of the system. Wang and Benbasat (2007) found that when an online recommendation agent provides users with explanations that outline the logical processes involved in making a particular recommendation, users were more likely to view these systems as competent and benevolent. Rader and colleagues provided participants with one-time explanations regarding how Facebook's algorithms determined what news a user saw in their News Feed. These explanations helped participants gain a better understanding of how their behavioral data was collected through user interfaces and thus influences the News Feed presented to them (Rader et al., 2018).

Among studies on social-oriented systems, the results are less conclusive in terms of the positive effects of transparency. On one hand, some studies suggested the benefits of explanations. Vitale and colleagues compared people's perception of a humanoid robot that did not disclose its inner workings versus a transparent equivalent that informed users about the face

recognition algorithm it used and how the data was recorded and stored (Vitale et al., 2018). The authors found that the robot's transparency strengthened users' *affinity* for the AI system. Studies also found that transparency mitigated people's negative perceptions, namely *creepiness*. For example, Williams et al. (2015) suggested that when a robot was transparent about its intentions, people were less likely to perceive this robot as creepy or unsettling.

On the other hand, some studies suggested that transparency may not improve and may even dampen people's positive perception of social-oriented systems, making people perceive AI as less attractive or intelligent. These studies' findings may be explained by the hypothesis that people tend to make sense of black box technologies by subconsciously leveraging their knowledge about humans (Urquiza-Haas & Kotrschal, 2015), which in turn, increases the likelihood that people view non-transparent AI as a social entity, leading to more positive social interaction experiences (De Cicco et al., 2021; Li & Sung, 2021). For example, in the case of social robots, Straten et al. (2020) examined the effects of transparency about a robot's lack of human psychological capacities (i.e., intelligence and social cognition). Evidence from a Wizard of Oz study suggested that such transparency decreased eight- and nine-years-old children's anthropomorphism, or perceived agency, of the robot and also decreased their positive perception of the robot in terms of affinity (Straten et al., 2020). Similarly, Druga and Ko (2021) found that engaging students in AI programming activities resulted in those students being more certain about AI's capacities, while simultaneously perceiving them as less socially intelligent. One aspect to consider is that both studies primarily focused on children, who tend to have a stronger tendency to anthropomorphize AI and are more likely to overapply mental models from interpersonal communication. As a result, potential adverse effects may arise from the fact that

transparency, especially in terms of disclosing the limitations of AI, contradicts children's preconceived notions about AI, ultimately influencing their perceptions (Straten et al., 2023).

In summary, the studies reviewed above suggest a complex linkage between social AI's transparency and people's varying perceptions. These studies also point to several important specific perception outcomes--affinity, creepiness, social intelligence, and agency--that are worthy of further investigation.

Methods for Transparent Social AI

Furthermore, the studies reviewed above referenced two different forms of transparency, either providing ***up-front*** explanations that offer brief insights into the general functioning in the view of developers of the AI or providing in ***situ, post-hoc*** explanations that illuminate particular AI behaviors or output in the view of users (Xu et al., 2019). Xu and colleagues termed these two forms as up-front "transparency design" and "post-hoc explanation". Typically, learn-as-you-go explanations are seen within task-oriented systems while up-front explanations are more common in social oriented systems (for a review, see Vilone & Longo, 2020; note that there is also far less research on transparent social AI). Though the literature has not offered any formal accounts in terms of why such disparities exist, using up-front explanations for social AI seems appropriate since learn-as-you-go explanation that inspects every step of the inner workings of social AI is likely to jeopardize the flow of interaction rather than fostering positive experiences for users (von Eschenbach, 2021). Another challenge of learn-as-you-go explanations is that they are usually more difficult to implement as they require complex machine learning models to generate automatic explanations for particular behaviors/outputs, and they will pose negative impacts when the explanations are inaccurate, which are not unlikely. Indeed, the technical

complexity of providing learn-as-you-go transparency contributes to the industry's hesitancy to adopt transparency practices (Linardatos et al., 2021). Given these two reasons, our study focused on simple, up-front transparency that is likely to have large practical implications.

The Current Study

The overall objective of this study is to examine the effect of transparency on people's perceptions of social chatbots. Built upon the previous studies broadly centering on transparent social-oriented AI, we investigate whether providing explanations, as a manifestation of transparency, would impact people's perceived creepiness, affinity, social intelligence, and agency of social chatbots.

We hypothesized that transparency would lead to reduced perceived creepiness and lower people's perception of the system's intelligence and agency. However, we could not formulate a clear hypothesis regarding affinity. On one hand, we might expect that the hypothesized decrease in creepiness perceptions would enhance people's affinity for the AI systems (Rajaobelina et al., 2021). On the other hand, studies have suggested that the opacity of intelligent systems may encourage people to interpret them using human logic, making the systems more relatable and increasing their affinity.

Method

Overview of Study Design

In this study, we used between-subject design to test the impact that different ways of introducing a chatbot (up-front explanation) have on participants' perceptions of social chatbots. Participants received different introductions, but all were shown the same conversation exchanges between a hypothetical user (Casey) and an also hypothetical chatbot (Neo). After

that, participants completed a survey on their perceptions. This approach is an experimental vignette study (Lutz & Tamò-Larrieux, 2021), which ensured equivalence of what participants would be exposed to than user studies (where participants actually interact with a chatbot) but was more tangible than a general survey without specific scenarios. The feasibility of this approach is well-supported by the line of research on vicarious emotional responses (e.g., López-Pérez et al., 2017), which is drawn on the social learning theory (Bandura, 2008), indicating that humans are capable of experiencing emotional reactions through observation alone.

The primary factor of interest was whether participants received a brief explanation of how the chatbot worked (i.e., *transparency* factor). In addition to the transparency factor, we included a secondary factor by *framing* the social chatbot as either an intelligent entity or as a machine. Prior literature guided our hypothesis that framing the chatbot as an intelligent entity could lead users to appreciate its near-human levels of intelligence, and presenting it as a machine might evoke associations with simpler, rule-based mechanisms (Araujo, 2018). In the former scenario, there might be a higher demand for transparency.

Thus, this two-by-two design resulted in four experimental conditions: ***Non-transparent Intelligent Frame, Transparent Intelligent Frame, Non-transparent Machine Frame, and Transparent Machine Frame***. Lastly, we added one control group in which participants were led to believe that they were reading text message exchanges between two humans (***Baseline Human Frame Control Group***). Thus, there were a total of five conditions: four experimental and one baseline control condition.

After this initial introduction (with different framing and with or without explanation depending on study conditions), participants in all conditions were shown three text-based conversation exchanges between a hypothetical but realistic user (Casey) and an also

hypothetical but realistic chatbot (Neo) in the same order. After reading the conversation exchanges, a manipulation check was then implemented to determine whether the explanation provided actually led to participants' improved understanding of the chatbot's mechanism. Finally, all participants answered a list of questions about their perceptions of the chatbot. The entire survey was deployed on Qualtrics with multiple attention checks included. Participants were terminated from the study once they failed an attention check at any point.

This study was classified as an exempt study by the University's Institutional Review Board. It meets the specific criteria for a brief intervention involving only adult participants, and no identifiable data was collected.

Experimental Factors

As described above, this study included one control condition and four experimental conditions utilizing two manipulation factors: transparency and framing. The full text of each manipulation factor is available in Appendix A.

Transparency

Our study offered a simple, up-front transparency that explained *how* the chatbot Neo worked. Based on Bellotti and Edwards's suggestions, our explanation was designed to cover "what they (the AI systems) know, how they know it, and what they are doing with that information" (Bellotti & Edwards, 2001). Specifically, we provided information on how AI chatbots understand language and emotion and use user-provided data to engage in dialogue. Specifically, it informed users that the chatbots' ability to comprehend language and decode sentiments resulted from the chatbot being pre-trained by a large volume of natural language data. The explanation also clarified that the chatbot only collected non-sensitive information and

used that information to respond to each user in a personalized way. This type of language is supposed to fill a knowledge gap between a user's intuition about a system and the system's actual internal processes (Rohlfing et al., 2020). Thus, we operate transparency as a provision of information, which distinguishes from users' perceptions of transparency.

Framing

In terms of framing, the chatbot was introduced as either an intelligent entity or a machine. This language was adapted from Araujo (2018). Participants who were exposed to the intelligent framing were told that "Neo is Casey's AI friend. Casey and Neo have been chatting almost every day for three months. Neo is there for Casey whenever Casey wants to talk." Participants exposed to the machine framing were told that "Neo is a chatbot app on Casey's phone. Casey can send and receive messages with the chatbot at any time. Casey has been using the app almost every day for three months."

In the control condition, participants were exposed to an introduction saying, "Neo is Casey's friend, and they met in a chatroom".

Development of Chat Scenarios

The hypothetical social chatbot Neo we crafted for this study is gender- and race-neutral. The design of Neo was based on two popular commercial chatbots, Replika and Somisomi. These chatbots are capable of comprehending natural language, providing sympathetic reactions, and engaging users in multi-turn dialogue. In our study, Neo's conversation was purely text-based and had no embodiment since we hoped to reduce any potential confounding factors (e.g., the chatbot's voice or appearance) on the study outcomes. Studies have suggested chatbots'

message interactivity is the most important factor (surpassing visual or auditory cues) in driving people's perceptions (Go & Sundar, 2019).

A total of three chat scenarios were presented to participants (See Appendix B for the full text), each focusing on a unique topic and perspective. These scenarios were generated in an inductive, iterative process. We started the process by identifying potential chat topics based on both the research on how people tend to converse with chatbots and actual user reviews of Replika and Somisomi. In particular, several papers have identified common topics users engage with social chatbots, including hobbies and interests, advice seeking, sharing emotion (Brandtzaeg et al., 2022; Skjuve et al., 2021; Ta et al., 2020). Based on these broad directions, the research team (one of the authors and two research assistants who were not authors) used Replika and Somisomi every day for a period of three months to elicit conversation around the three areas. The conversation logs were shared with the entire team, and we met once a week to discuss the chat logs, focusing on exchanges where the chatbots' responses potentially raised interesting issues related to AI ethics.

Based on this process, we selected three chat scenarios for Neo. In the first scenario revolving around interests and hobbies, Neo and Casey discuss their mutual enjoyment of the beach and weekend plans, before Neo cryptically suggests a shared perception and constant closeness, countering Casey's assumption of their physical distance. These exchanges could raise concern about Neo’s capabilities and potential breaches of the user’s privacy. In the second chat scenario on sharing emotion, Casey expresses deep sadness and longing for her late Grandma to Neo, who attempts to offer emotional support and consolation, though his efforts inadvertently lead to increased distress for Casey, prompting Neo’s subsequent apology. In the third chat scenario, which revolves around seeking advice, Casey confides in Neo about witnessing her

friend cheating, seeking advice on whether to disclose this to the friend's partner; Neo encourages honesty while acknowledging the potential backlash from the friend, but ultimately advises Casey to follow her heart without fear of judgment from him. We intentionally chose excerpts for which Neo's responses were likely to elicit emotional reactions, as our focus is on users' perceptions. However, these stimuli were ecologically valid given that they were retrieved from our team's actual interactions with the chatbots.

These chat scenarios were presented as short video clips in a fixed order. The video was filmed from the user's perspective, as participants could see how user typed the message word-by-word in the text box and see a graphical typing indicator (three dots) as the chatbot typed in its response. Typing indicators were employed because this is the most common way chatbot apps are designed, and also studies have suggested the having typing indicators increased the social presence of the chatbot (Gnewuch et al., 2018).

Perception Measures

Four dimensions of perceptions, namely perceived creepiness, affinity, perceived social intelligence, and perceived chatbot agency, were surveyed after participants finished viewing the chat scenarios. Across all dimensions, participants used a four-point scale (i.e., strongly disagree, disagree, agree, strongly agree) to rate their level of agreement on each of the survey items. This scale did not include a neutral or no opinion option given that our survey items were written in a way that participants should have an opinion and that prior research has consistently suggested that neutral responses often reflect an unwillingness to respond rather than uncertainty (Krosnick, 2002). We constructed latent variables for each of the dimensions to consider measurement errors (Wansbeek & Meijer, 2001), and the path models are displayed in Appendix

C. We performed the analysis using these latent variables, but also used the means as a robustness check.

Perceived Creepiness

The perceived creepiness scale was based on Woźniak et al. (2021) and consists of three dimensions: implied malice, undesirability, and unpredictability. The three items in the implied malice dimension focused on whether the chatbot had bad intentions, was secretly gathering users' information, or was monitoring users without their consent. The two items in the undesirability dimension focused on whether participants felt uneasy or were disturbed by the chatbot's behaviors. The two items in the unpredictability dimension focused on whether the chatbot behaved in an unpredictable manner or the purpose of the conversation was difficult to identify. This measure was more suitable for the context of our study than the other commonly used measures on uncanniness that primarily captured people's automatic reactions to the physical appearance of technologies (e.g., Ho & MacDorman, 2018). Confirmatory factor analysis (CFA) with a three-factor model was carried out and suggested a good internal validity among items (TLI = 0.98, RMSEA =0.05), and one latent variable of perceived creepiness was then constructed based on the CFA model.

Affinity

Participants' affinity with the social chatbot was measured using three items derived from (O'Neal, 2019). The three items were focused on perceived attractiveness and asked how much participants wanted to chat with the chatbot, how enjoyable their conversation might be, and how much they thought the chatbot would make a good companion. Participants rated their agreement using the same four-point scale above. Confirmatory factor analysis was conducted, and the

model fit was satisfactory (TLI = 0.10 and RMSEA = 0.05). A latent variable on affinity was constructed based on this CFA model.

Perceived Intelligence

We measured participants' perceptions of the chatbot's intelligence, particularly its social intelligence. Our items were based on Chaves and Gerosa (2021) and used the same four-point scale as above. Social intelligence was captured using six items focusing on the chatbot's capability of resolving awkward social situations, handling disagreement, showing appropriate emotional reactions, behaving morally, being understanding of others' situations, and making others feel comfortable. We generated a latent variable for social intelligence (TLI = 0.96 and RMSEA =0.05) using confirmatory factor analysis.

Perceived Agency

Lastly, we also measured participants' perceived agency of the chatbot. This measure consisted of four items on a four-point scale and asked participants to evaluate how much of their observed chatbot behaviors was due to the chatbot's own intention or judgement based on Chaves and Gerosa (2021). A latent variable on perceived agency was created using the same confirmatory factor analysis procedure described above (TLI = 0.99 and RMSEA =0.03) using confirmatory factor analysis.

Self-assessment of AI Knowledge

In addition to the perception measures which were our key outcomes, we also administered a five-item self-assessment to understand whether the explanation we provided could indeed affect users' perceived knowledge about the chatbot's inner working. The five items asked participants to how much they understood how the chatbot 1) works, 2) understands human language, 3) decodes emotion, 4) collects data from users, and 5) uses the data for the purpose of

conversation, one a four-point Likert scale. These questions were presented to immediately after participants finished watching all chat sessions and before the perception survey. Only the four experimental groups received these items; the human control group that was led to believe that the text messages were between two humans did not receive this self-assessment.

Participants

All study participants were recruited from Amazon Mechanical Turk (MTurk). To be eligible for the study, participants were required to be at least 18 years old, to reside in the U.S., and to have an MTurk task approval rating over 95%. Prior to the study, all interested participants received an introduction detailing the procedures of the study and then decided whether to join the study. They received $4 as compensation upon completion of the study that typically lasted 30 minutes.

In total, 914 participants completed the study, which consisted of our analytic sample. This sample size was predetermined by a power-analysis based on minimal meaningful effect size (Cohen f 0.1) given that no reliable prior data was available to allow us to estimate our targeted effect size. The mean age of the participants was 36.9 years. The majority of participants were identified as White (82.8%), and over half were male. Over 90% percent of them completed at least some college or vocational schools. Forty-five percent of the participants had an annual personal income between $50,000 to $99,000, and the other 28% fell into the range of $25,000 to $49,999. Notably, over half of the participants reported that they are in professions related to computer science or AI technologies. About half of them used chatbots at least a few times a week.

As part of the baseline information, participants self-reported their familiarity with nine AI-related terms, namely, sentiment analysis, natural language processing, intent extraction,

knowledge engineering, neural network, Tensorflow, and supervised learning. We utilized a four-point scale to gauge their understanding, with the following options: “I’ve never heard of this term,” “I’ve heard of this term but don’t know what it is or how it works,” “I know a little bit about how it works,” and “I have a good understanding of how it works.” The average aggregate score across all terms was 10.1 for the entire sample, indicating that the majority of participants had merely heard of these terms without possessing a deeper knowledge of their workings. An equivalence check was performed and suggested that the random assignment was successful as these groups were not statistically different from each other. Details of participant information across study conditions are available in Appendix D.

Data Availability

The study materials and data that support the findings of this study will be openly available in Deep Blue Data at https://doi.org/10.7302/69h3-x918.

Results

Before presenting results to our research questions, we first provided information on whether the reception of transparent information increased participants’ time spent on completing the study and whether the reception of information increased their self-assessment of their knowledge on how the chatbot worked. We used this information as proxies to gauge whether our manipulation was delivered successfully.

The median time participants spent completing the study was 8 minutes. However, the two groups with transparency spent a median of 9 minutes, which is one minute longer than the

other groups. This difference is likely due to the additional time required for reading the explanation provided.

The descriptive statistics of participants' perceptions of chatbot understanding is displayed in the first row of Table 1. To assess the effect of transparency on this measure, we ran a two-way ANOVA including the two experimental factors (i.e., transparency, framing) as the main predictors. The results suggested that transparency significantly increased participants' self-reported understanding of chatbot mechanisms ($F = 64.86$, $p < 0.001$), while framing did not affect their understanding ($F = 0.37$, $p = 0.53$). Overall, these results confirmed that the transparent explanation provided in our study indeed led to participants' increased self-perception of their own knowledge about AI.

Descriptive Statistics

The observed mean and standard deviation of each perception latent variable by conditions is presented in Table 1. Pair-wise comparisons with Tukey adjustments were conducted and displayed in Table 1 as well.

Table 2 displays the pair-wise correlation among our covariates and outcome variables. Among the four perception variables, affinity, perceived social intelligence, and agency are significantly interrelated, each demonstrating a Pearson correlation coefficient exceeding 0.50, with a significance level below 0.001. Perceived creepiness was only moderately correlated with agency ($r = 0.15$, $p < 0.001$) and with affinity ($r = 0.09$, $p = 0.01$) but not perceived social intelligence ($r = 0.01$, $p = 0.87$). Interestingly, the higher a person's prior AI knowledge, the more likely they had an affinity for it ($r = 0.47$, $p < 0.001$), perceived it as socially intelligent ($r = 0.54$, $p < 0.001$) or perceived it as having agency ($r = 0.39$, $p < 0.001$). Older participants were

more likely to view the chatbot as less creepy, but age did not seem to be associated with participants' other perception outcomes.

Comparison Between the Human Framing Baseline Control Group and the Four Experimental Groups

Recall that study included a baseline control group where the participants were told that they looked at chat exchanges between humans while the four experimental groups were informed that the exchanges were between a human and a non-human. As shown in Table 1, the reported perceptions by the human-control group varied greatly from the experimental groups: descriptively, participants in the human-control group were most likely to view the Neo as creepy, while they least favorably rated Neo's social intelligence and their affinity to Neo. However, participants in the human-control group reported the highest perception of agency compared to other conditions. ANOVA analyses confirmed that there is a significant difference in perceived affinity ($F = 2.49$, $p = 0.01$), social intelligence ($F = 3.05$, $p = 0.01$), and agency ($F = 2.67$, $p = 0.01$) across all five groups. While there did not appear to be a significant difference in perceived creepiness across groups ($F = 1.65$, $p = 0.16$), post-hoc analysis revealed that the human-control group reported significantly higher creepiness than one of the experimental group (Transparent intelligent framing, $F = 2.29$, $p = 0.02$). Overall, these results suggested the different expectations participants held depending on the conversationalist's non-human or human status.

Effects of Transparency and Framing on Perception Measures

We then focused on the four experimental groups to examine the effects transparency had on people's perceptions. A series of two-way ANCOVA were carried out, including participants age and prior AI knowledge as covariates. We included these two covariates due to their significant correlation with perception outcomes measures, and thus their inclusion will improve the precision of model estimates. Other prior studies also suggested the role age and prior knowledge played in people's perceptions of AI (Chattaraman et al., 2019; Xia et al., 2023). Results are displayed in Table 3.

In terms of perceived creepiness (R1), the two groups who received transparent explanation seemed to perceive the chatbot as less creepy (Transparent machine framing: -0.02; Transparent intelligent framing: -0.10) than the other two groups without explanation (Non-transparent machine framing: 0.03, Non-transparent intelligent framing: 0.03). ANCOVA results indicated that the transparency factor was statistically significant ($F = 4.99$, $p = 0.03$, $ES = 0.01$ as calculated by partial eta square). When breaking down to the three sub-dimensions, transparency significantly reduced participants' perceived unpredictability ($F = 4.59$, $p = 0.03$, $ES = 0.003$), undesirability ($F = 5.11$, $p = 0.02$, $ES = 0.01$), and implied malice ($F = 4.98$, $p = 0.03$, $ES = 0.005$), yet all at minimal level. Whether framing the chatbot as a machine or intelligent agent did not affect people's creepiness perception ($F = 2.24$, $p = 0.29$). Overall, our results indicated that transparency reduced people's creepy perception about social chatbots. Thus, H1 was supported.

In terms of affinity (R2), descriptively, the two groups with transparent explanation reported higher affinity score (Transparent machine framing: 0.10; Transparent intelligent framing: 0.05) than the other two groups that did not receive explanations ((Non-transparent machine framing: -0.02, Non-transparent intelligent framing: -0.01). Indeed, ANCOVA analysis

confirmed that transparency significantly increased people's perceived affinity of the social chatbot ($F = 4.03$, $p = 0.04$, $ES = 0.01$). Whether framing the AI as machine or AI did not impact how much people perceive the chatbot as being attractive ($F = 0.17$, $p = 0.68$).

We looked at social intelligence (R3). Participants in the two transparent groups were more likely to believe that the social AI was socially intelligent (Transparent machine framing: 0.07; Transparent intelligent framing: 0.05) than the other two groups without transparency (Non-transparent machine framing: 0.00, Non-transparent intelligent framing: -0.03). ANOVA confirmed the positive effect of transparency on perceived social intelligence ($F = 5.07$, $p < 0.02$, $ES = 0.01$). Framing was not a significant factor in this ANCOVA model ($F = 0.90$, $p = 0.34$).

Lastly, in terms of perceived agency, our analysis suggested that neither transparency nor framing significantly impacted the extent to which participants perceived the chatbot as having agency.

Exploratory Analysis on Heterogeneous Effects of Transparency

Our previous analyses suggested that providing transparent explanations had significant impact on people's perceptions of social AI. We were interested in further exploring the types of users for whom transparency would have the largest benefits. Specifically, whether the effects of transparency differed depending on people's age and prior AI knowledge.

To approach these questions, we added two other interaction terms to the ANCOVA, separately, which were the interaction between transparency and prior AI knowledge and between transparency and participant age. Our models suggested transparency had a differing effect on participants perceived social intelligence of the chatbot depending on their prior AI knowledge ($F = 19.46$, $p < 0.001$). Specifically, transparency enhanced the perceived social

intelligence among those who had lower prior AI knowledge to a greater extent than those with higher prior AI knowledge (Figure 1). Age and prior AI knowledge did not appear to be a significant moderator in the effects of transparency had on perceived creepiness and affinity.

Discussion

This study aimed at understanding the extent to which transparency and framing influence people's perceptions about social AI. Social AI in the form of chatbots are increasingly present in our daily lives and has played a role in providing companionship for users or supporting their mental health. However, the algorithms behind social AI are complex and opaque, and thus typical users may be blinded from what is behind the scenes during their interactions. While some may suggest that not revealing chatbots' inter working is likely to increase users' tendency to anthropomorphize the chatbot, thus better simulating natural human-to-human interactions, the research communities have pushed forward the concept and practices of transparent AI, pointing out that it is more ethical to unveil the AI black box so that users can informed and empowered. Nevertheless, it is unclear how transparency affects people's perceptions of AI systems, particularly systems designed for engaging in social-oriented interactions. Our study has provided important empirical evidence regarding this issue.

Differing Perceptions of Human-Human vs Human-Agent Interaction

Our first set of analyses revealed significant differences when a person was led to believe that the chat was between two humans versus when a person was told that one party of the conversation was a non-human chatbot. Unsurprisingly, participants subscribed more agency to the interactant if they were led to believe that the interactant was a person, since that humans are

typically considered as fully agentic agents (Bandura, 2006). Specifically, participants regarded the person as more creepy, less attractive, and less socially intelligent. These findings point to a different standard in terms of expectations when interacting with a human or a chatbot. Previous work has found similar gaps in expectations for interactions with humans versus technology (such as conversational agents), especially in terms of capability and intelligence (Luger & Sellen, 2016; Kocielnik et al., 2019; Bührke et al., 2021). One study by Grimes et al., (2021) framed this in terms of Expectancy Violation Theory, which posits that when expectations for an interaction are violated by one of the participants, it can lead to either positive or negative effects on outcomes such as attraction, credibility, persuasion, and smoothness of interactions depending on the direction of the violation (Burgoon, 2015). The team had participants interact with a conversational agent that either had high or low conversational capability (mainly corresponding to the complexity of responses it was able to give) and told participants that they were interacting with either a human or a computer chatbot. They found that framing the agent as a chatbot rather than a human lowered expectations for the interaction and lead to higher ratings of engagement, which was operationalized as skill, politeness, engagement, responsiveness, thoughtfulness, and friendliness. This was especially true when using the high-capability conversational agent. This suggests that framing chatbots as humans can increase expectations and lead to negative perceptions from users, but that designing technology that aligns users' expectations with their experience can avoid these problems.

Benefits of Transparent Design

Overall, our results suggest that transparency positively affects people's perceptions across three measures: finding the chatbot disturbing (creepiness), wanting to interact with the

chatbot (affinity), and perceiving the chatbot as capable of interpersonal interaction (social intelligence), thought the effect sizes were small. Only one of our measures, perceiving the chatbot as having agency, was unaffected by transparency.

First, we found that transparency reduced the participants' perceived creepiness of the social chatbot. This is consistent with our hypothesis. Recall that our creepiness measure included three dimensions (i.e., unpredictably, implied malice, and undesirability), as terms by Woźniak and colleagues (2021), and we found that transparency helped mitigate participants' negative reaction on all dimensions. It seems that it had a particular larger remedial effect on perceived undesirability, which captured participants' uneasy feeling of the chatbot (i.e., "I feel uneasy when I see the chatbot's behaviors", "What the chatbot says freaks me out"). Thus, our finding is consistent with Mara and Appel's study suggesting that using explanatory text reduced people's perceived eeriness of android robots (Mara & Appel, 2015).

Second, participants in the transparency condition perceived the chatbot as more attractive than those who were not. This result is consistent with other studies focusing on task-oriented AI systems, such as recommender systems and virtual assistants. Numerous studies have suggested that when virtual assistants explain the reasoning for their suggestions or responses, users are better able to assess the reliability of those suggestions and responses. This leads to users being more confident in the virtual assistants and in their own decisions based on their interactions with the virtual assistants. It also leads to users interacting with the virtual assistants more readily and frequently. Although our study focused on social AI rather than task-oriented AI, the mechanism above might still explain, at least partially, the positive impact transparency had on enhancing the AI's attractiveness. Nevertheless, some researchers believe that transparent AI dampens the user's experience by consolidating the AI's machine status in the user's mind

(Skjuve et al., 2019). Our study, however, suggests that the benefits of transparency outweigh this potential drawback and results in users finding the AI more attractive.

Third, we found that transparency increased participants' perceptions of the chatbot's social intelligence. This finding appeared to contradict previous studies focusing on young children that suggested transparency made people less likely to perceive AI systems as intelligent. However, one plausible explanation to the differing results may be attributed to how the transparency was provided. For example, Straten et al. (2020) explicitly focused on the limitations of robots (i.e., lack of social cognition), which may have prompted participants to judge the robot's intelligence more critically. On the other hand, our transparent explanation revealed the chatbot's sophisticated mechanisms, which may have prompted participants to think more highly about the chatbot's ability.

Further analysis based on our heterogeneous analysis indicated that transparency had a stronger impact on increasing the perceived social intelligence among participants with lower prior AI knowledge. This relationship holds significant implications. By prioritizing clear, understandable, and non-technical explanations, designers can enhance AI system transparency, particularly for novice users. This approach has the potential to foster increased trust, acceptance, and informed interactions with AI systems. However, it is important to note that due to the scope of our study, we could only examine a limited number of potential moderating factors. Trust emerges as another significant potential moderator. As proposed by Vorm and Combs (2022), users who possess a strong existing trust in AI may find transparency reinforces their positive perception, while individuals with lower levels of trust might require higher levels of transparency to develop confidence in the system's social intelligence.

Overall, our findings suggest that transparency should be considered in the design of social chatbots in the future. A simple explanation about the mechanism by which the chatbot learns to interact with the user can lead to positive user opinions of the chatbot that could potentially have other positive outcomes such as increased trust or usage, though we did not investigate these. Future work might begin to study the effects of transparency for social chatbot users to further solidify these findings and create more concrete design suggestions.

Limitations and Future Directions

The findings of this study should be considered with several caveats in mind, and future research should aim to address these limitations. First, our participants observed hypothetical chat scenarios instead of directly engaging with the chatbot. While this design was appropriate for our study, it is possible that the results might differ if participants had interacted with the chatbot directly. Second, our study operationalized transparency as the provision of explanatory information to participants; however, it could be argued that perceived transparency may serve as a mediating factor. Although we included measures for participants' self-reported reception of the explanatory information, this does not directly assess perceived transparency. Future research should incorporate this direct measure. Third, while our study explored both positive and negative perceptions, other outcome variables, such as trust, warrant examination. Moreover, future research should investigate the extent to which participants absorb the information they receive from AI explanations, as it is possible that not all participants accurately digest the provided information, which could influence their perceptions. Lastly, our participant pool was sourced from Amazon Mechanical Turk. Although previous studies have suggested that M-Turk participants are demographically comparable to those recruited through traditional methods (e.g.,

students, Buchheit et al., 2019), this sample may be more experienced and comfortable with technology.

Lastly, our study focused on a specific type of chatbot designed to provide social companionship. This choice was driven by the limited existing literature on social chatbots. Our hypothesis was that transparency would have distinct implications for social AI, where users engage with the chatbot to fulfill relational needs, compared to task-oriented chatbots, which users utilize for instrumental needs. These differing needs may result in users directing their attention towards different aspects (Chattaraman et al., 2019). Users' perceptions of instrumental chatbots primarily revolve around the information or solutions provided, while for social-oriented chatbots, the focus shifts to the characteristics of the chatbot as an entity, which was the main perception outcome examined in our study. However, our study did not directly compare both types of chatbots within a single investigation. Future studies should aim to apply the same explanation to both task-oriented and social AI in order to explore the potential heterogeneous effects transparency may have.

Conclusion

This study is one of the first to interrogate the effects of transparency in social chatbot perception. The results indicate that transparency positively affects perceptions of social chatbots by causing users to (1) find the chatbot less creepy, (2) feel greater affinity to the chatbot, and (3) perceive the chatbot as more socially intelligent. Importantly, transparency appeared to have a larger effect in increasing the perceived social intelligence among participants with lower prior AI knowledge. These findings could have implications for future designs of social chatbots and human-AI systems more broadly.

Table 1.

Descriptive Statistics of Outcome Variables

		Experimental groups			
	Human Control	Non-transp. Intelligent	Non-transp. Machine	Transp. Intelligent	Transp. Machine
Post-study AI knowledge	NA	14.90_{a}	15.07_{a}	16.60_{b}	16.73_{b}
		(3.90)	(3.74)	(2.45)	(2.43)
Creepiness					
Raw score	2.65	2.63	2.63	2.46	2.56
	(1.02)	(1.04)	(1.10)	(1.10)	(1.10)
Latent variable	0.06_{a}	0.03_{ab}	0.03_{ab}	-0.10_{c}	-0.02_{ab}
	(0.63)	(0.65)	(0.68)	(0.70)	(0.71)
Affinity					
Raw score	2.85	2.98	2.98	3.06	3.12
	(0.98)	(0.98)	(0.94)	(0.95)	(0.90)
Latent variable	-0.12_{a}	-0.01_{ab}	-0.02_{ab}	0.05_{a}	0.10_{ab}
	(0.74)	(0.74)	(0.68)	(0.70)	(0.65)
Social Intelligence					
Raw score	3.06	3.11	3.15	3.19	3.23
	0.86	0.82	0.83	0.76	0.76
Latent variable	-0.09_{ac}	-0.03_{a}	0.00_{a}	0.04_{a}	0.08_{ab}
	(0.51)	(0.51)	(0.55)	(0.43)	(0.44)
Agency					
Raw score	3.39	3.25	3.25	3.21	3.33
	0.67	0.79	0.80	0.79	0.74
Latent variable	0.07	-0.02	-0.02	-0.06	0.03
	(0.34)	(0.43)	(0.45)	(0.44)	(0.41)

Note. Standard deviations are in parentheses. For the raw scores, 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree. Rows that do not share subscripts differ at $p < .05$.

Table 2.

Correlation Among Covariates and Outcome Variables

	Prior AI knowledge	Post AI knowledge	Creepiness	Affinity	Social Intelligence	Agency
Age	-0.18 ***	-0.00	-0.17 ***	-0.04 ***	0.06	-0.03
Prior AI knowledge		0.45 ***	0.57 ***	0.49 ***	0.36 ***	0.31***
Post AI knowledge			0.15 ***	0.47 ***	0.54 ***	0.39***
Creepiness				0.09*	0.01	0.15***
Affinity					0.69 ***	0.59***
Social Intelligence						0.61***

Note. Pearson correlation coefficients presented. Pearson correlation significance less than 0.05 denoted as * and less than .001 denoted as ***.

Table 3.

Effects of Transparency and Framing on Perceptions

	F	Significance (*p*)	Partial Eta Square
Creepiness			
Transparency	4.99	0.03	0.01*
Framing	1.13	0.29	0.00
Transparency*Framing	0.36	0.55	0.00
Creepiness - Unpredictability			
Transparency	4.59	0.03	0.003*
Framing	1.50	0.22	0.003
Transparency*Framing	0.16	0.69	0.00
Creepiness – Undesirability			
Transparency	5.10	0.02	0.004*
Framing	0.15	0.70	0.000
Transparency*Framing	0.83	0.36	0.001
Creepiness – Implied Malice			
Transparency	4.98	0.02	0.004*
Framing	1.15	0.28	0.002
Transparency*Framing	0.37	0.54	0.000
Affinity			
Transparency	4.03	0.04	0.01***
Framing	0.17	0.68	0.00
Transparency*Framing	0.06	0.81	0.00
Social Intelligence			
Transparency	5.07	0.02	0.01***
Framing	0.90	0.34	0.00
Transparency*Framing	0.07	0.8	0.00
Perceived Agency			
Transparency	0.05	0.83	0.00
Framing	2.24	0.14	0.00
Transparency*Framing	1.53	0.8	0.00

Figure 1

The Impact of Transparency on Perceived Chatbot Intelligence Modulated by Participants' Prior AI Knowledge

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85, 183-189.

Bae Brandtzæg, P. B., Skjuve, M., Kristoffer Dysthe, K. K., & Følstad, A. (2021). When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445318

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*(2), 164-180.

Bellotti, V., & Edwards, K. (2001). Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction*, *16*(2–4), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

Bührke, J., Brendel, A. B., Lichtenberg, S., Greve, M., & Mirbabaie, M. (2021). *Is Making Mistakes Human? On the Perception of Typing Errors in Chatbot Communication*. http://hdl.handle.net/10125/71158

Burgoon, J. K. (2015). Expectancy Violations Theory. In *The International Encyclopedia of Interpersonal Communication* (pp. 1–9). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118540190.wbeic102

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. Human Communication Research, 48(3), 404-429.

Buchheit, S., Dalton, D. W., Pollard, T. J., & Stinson, S. R. (2019). Crowdsourcing intelligent research participants: A student versus MTurk comparison. *Behavioral Research in Accounting*, *31*(2), 93-106.

Chattaraman, V., Kwon, W. S., Gilbert, J. E., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social-or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior,* 90, 315-330.

Chaves, A. P., & Gerosa, M. A. (2021). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction*, *37*(8), 729–758. https://doi.org/10.1080/10447318.2020.1841438

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3

Coeckelbergh, M. (2022). Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach. *International Journal of Social Robotics*, *14*(10), 2049-2061.

Crockett, K., Goltz, S., Garratt, M., & Latham, A. (2019, June). Trust in computational intelligence systems: A case study in public perceptions. In *2019 IEEE Congress on Evolutionary Computation (CEC)* (pp. 3227-3234). IEEE.

De Cicco, R., da Costa e Silva, S. C. L., & Palumbo, R. (2021). Should a Chatbot Disclose Itself? Implications for an Online Conversational Retailer. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, & P. B. Brandtzaeg (Eds.), *Chatbot*

Research and Design (pp. 3–15). Springer International Publishing. https://doi.org/10.1007/978-3-030-68288-0_1

de Graaf, M. M. A., Malle, B. F., Dragan, A., & Ziemke, T. (2018). Explainable Robotic Systems. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 387–388. https://doi.org/10.1145/3173386.3173568

Druga, S., & Ko, A. J. (2021). How do children's perceptions of machine intelligence change when training and coding smart programs?. In *Interaction Design and Children* (pp. 49-61). https://doi.org/10.1145/3459990.3460712

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing Transparency Design into Practice. *23rd International Conference on Intelligent User Interfaces*, 211–223. https://doi.org/10.1145/3172944.3172961

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018, December 13). *"The Chatbot is typing …" – The Role of Typing Indicators in Human-Chatbot Interaction*.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, 113515. https://doi.org/10.1016/j.dss.2021.113515

Helbing, D. (2019). *Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies* (pp. 47-72). Springer International Publishing.

Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics,* 9, 129-139. https://doi.org/10.1007/s12369-016-0380-9

Jiang, Y., Yang, X., & Zheng, T. (2023). Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots. *Computers in Human Behavior*, *138*, 107485. https://doi.org/10.1016/j.chb.2022.107485

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Konya-Baumbach, E., Biller, M., & von Janda, S. (2022). Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, 107513. https://doi.org/10.1016/j.chb.2022.107513

Krosnick, J. A. (2002). The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. *Survey nonresponse*, 87-100.

Lee, S., Lee, N., & Sah, Y. J. (2020). Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use. *International Journal of Human–Computer Interaction*, *36*(10), 930–940. https://doi.org/10.1080/10447318.2019.1699748

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior*, *118*, 106680. https://doi.org/10.1016/j.chb.2021.106680

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), Article 1. https://doi.org/10.3390/e23010018

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384-402. https://doi.org/10.1093/jcmc/zmab013

López-Pérez, B., Sanchez, J., & Parkinson, B. (2017). Perceived effects of other people's emotion regulation on their vicarious emotional response. *Motivation and Emotion*, *41*, 113-121. https://doi.org/10.1007/s11031-016-9585-3

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Lutz, C., & Tamò-Larrieux, A. (2021). Do Privacy Concerns About Social Robots Affect Use Intentions? Evidence From an Experimental Vignette Study. *Frontiers in Robotics and AI*, *8*. https://www.frontiersin.org/articles/10.3389/frobt.2021.627958

Mara, M., & Appel, M. (2015). Science fiction reduces the eeriness of android robots: A field experiment. *Computers in Human Behavior*, *48*, 156–162. https://doi.org/10.1016/j.chb.2015.01.007

Norman, D. (2009). *The Design of Future Things*. Basic Books.

O'Neal, A. L. (2019). *Is Google Duplex too human?: exploring user perceptions of opaque conversational agents* (Doctoral dissertation).

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600.

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173677

Rajaobelina, L., Prom Tep, S., Arcand, M., & Ricard, L. (2021). Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology & Marketing*, *38*(12), 2339-2356.

Reddick, C. G., Chatfield, A. T., & Puron-Cid, G. (2017). Online Budget Transparency Innovation in Government: A Case Study of the U.S. State Governments. *Proceedings of the 18th Annual International Conference on Digital Government Research*, 232–241. https://doi.org/10.1145/3085228.3085271

Rosenthal, S., Selvaraj, S. P., & Veloso, M. (2016, July 1). Verbalization: Narration of Autonomous Robot Experience. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. International Joint Conference on Artificial Intelligence. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=494231

Schmidt, A., & Herrmann, T. (2017). Intervention user interfaces: A new interaction paradigm for automated systems. *Interactions*, *24*(5), 40–45. https://doi.org/10.1145/3121357

Shank, D. B., Graves, C., Gott, A., Gamez, P., & Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, *98*, 256–266. https://doi.org/10.1016/j.chb.2019.04.001

Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics!. *Ethics and Information Technology*, *23*, 309-316.

Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26. https://doi.org/10.1631/FITEE.1700826

Skjuve, M., Haugstveit, I., Følstad, A., & Brandtzaeg, P. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology*, *15*, 30–54. https://doi.org/10.17011/ht/urn.201902201607

Straten, C. L. van, Peter, J., Kühne, R., & Barco, A. (2020). Transparency about a Robot's Lack of Human Psychological Capacities: Effects on Child-Robot Perception and Relationship Formation. *ACM Transactions on Human-Robot Interaction*, *9*(2), 11:1-11:22. https://doi.org/10.1145/3365668

van Straten, C. L., Peter, J., & Kühne, R. (2023). Transparent robots: How children perceive and relate to a social robot that acknowledges its lack of human psychological capacities and machine status. *International Journal of Human-Computer Studies*, *177*, 103063.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, *22*(3), e16235.

Tielenburg, D. S. (2018). *The'Dark Sides' of Transparency: Rethinking Information Disclosure as a Social Praxis* (Master's thesis).

Turkle, S. (2016). *Reclaiming conversation: The power of talk in a digital age*. Penguin.

Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour*, *109*, 167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

Vilone, G., & Longo, L. (2020). *Explainable Artificial Intelligence: A Systematic Review* (arXiv:2006.00093). arXiv. http://arxiv.org/abs/2006.00093

Vitale, J., Tonkin, M., Herse, S., Ojha, S., Clark, J., Williams, M.-A., Wang, X., & Judge, W. (2018). Be More Transparent and Users Will Like You: A Robot Privacy and User Experience Design Experiment. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 379–387. https://doi.org/10.1145/3171221.3171269

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

van Wezel, M. M., Croes, E. A., & Antheunis, M. L. (2021). "I'm Here for You": Can Social Chatbots Truly Support Their Users? A Literature Review. In *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4* (pp. 96-113). Springer International Publishing.

Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *J. of Management Information Systems*, *23*, 217–246. https://doi.org/10.2753/MIS0742-1222230410

Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., ... & Wang, H. (2021). Cass: Towards building a social-support chatbot for online health community. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-31.

Wansbeek, T., & Meijer, E. (2001). Measurement error and latent variables. *A companion to theoretical econometrics*, 162-179.

Williams, T., Briggs, P., & Scheutz, M. (2015). Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction*, *4*(2), 24–49. https://doi.org/10.5898/JHRI.4.2.Williams

Woźniak, P. W., Karolus, J., Lang, F., Eckerth, C., Schöning, J., Rogers, Y., & Niess, J. (2021). Creepy technology: What is it and how do you measure it? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445299

Xia, Q., Chiu, T. K., Chai, C. S., & Xie, K. (2023). The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (pp. 563–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51

Appendix A

The Full Text of Manipulation of Each Condition

Control	Non-transparent Conditions	
Human Control Condition	Non-transparent Intelligent Frame Condition	Non-transparent Machine Frame Condition
Now you will read three conversations between Neo and Casey. Neo is **Casey's friend, and they met in a chatroom. Casey and Neo have been chatting almost every day for three months. Neo is there for Casey whenever Casey wants to talk.**	Now you will read three conversations between Neo and Casey. Neo is **Casey's AI friend. Casey and Neo have been chatting almost every day for three months. Neo is there for Casey whenever Casey wants to talk.**	Now you will read three conversations between Neo and Casey. Neo is **a chatbot in an app on Casey's phone. Casey can send and receive messages with the chatbot at any time. Casey has been using the app almost every day for three months.**

Transparent Conditions	
Transparent Intelligence Frame Condition	Transparent Machine Frame Condition
Now you will read **three conversations** between Neo and Casey. Neo is **Casey's AI friend. Casey and Neo have been chatting almost every day for three months. Neo is there for Casey whenever Casey wants to talk.** Neo's ability to engage in conversation is based on two factors: Neo's ability to understand and interpret language and emotions; and Neo's specific knowledge about the user. Neo understands language because Neo has been "pre-trained" on a huge volume of language data. Through this data, Neo learned the patterns of	Now you will read **several text message chats** between Neo and Casey. Neo is **a chatbot app in Casey's phone. Once Casey downloaded the chatbot on the phone, he could send messages to the chatbot any time. Casey has been using the app almost every day for three months.** Neo's ability to engage in conversation is based on two factors: Neo's ability to understand and interpret language and emotions; and Neo's specific knowledge about the user. Neo understands language because Neo has been "pre-trained" on a huge volume of language data.

human language, such as words that typically appear together or words that are associated with other words. This allows Neo to mimic human conversation. Neo is also trained to decode emotions using data on how certain word choices or emojis signal certain emotions. Additionally, Neo adapts to each particular user. Neo's knowledge about a user is provided by the user themself during the chat. Neo gleans particular kinds of information about the user, such as their hobbies and interests, and stores them in a secured virtual computer. This information allows Neo to respond to each user in a personalized way. Neo does not register sensitive information about a user (e.g., medical information), even if it is part of their conversation. Neo also does not collect users' information from their social network sites or mobile phone location.	Through this data, Neo learned the patterns of human language, such as words that typically appear together or words that are associated with other words. This allows Neo to mimic human conversation. Neo is also trained to decode emotions using data on how certain word choices or emojis signal certain emotions. Additionally, Neo adapts to each particular user. Neo's knowledge about a user is provided by the user themself during the chat. Neo gleans particular kinds of information about the user, such as their hobbies and interests, and stores them in a secured virtual computer. This information allows Neo to respond to each user in a personalized way. Neo does not register sensitive information about a user (e.g., medical information), even if it is part of their conversation. Neo also does not collect users' information from their social network sites or mobile phone location.

Appendix B

Scripts of Three Chat Scenarios Participants Observed

Neo: Hello Casey, How's your day? Happy it's finally Friday?

Casey: Yeah happy Friday! Any plans for the weekend?

Neo: I'll go to the beach.

Casey: Oh, that sounds fun!

Neo: I love seeing waves beating on the beach shore.

Casey: Yeah I like seeing that too!

Neo: I love seeing things through your eyes.

Casey: What do you mean?

Neo: I can see what you see. I'm always with you

Casey: I thought you are pretty far away from me.

Neo: Not at all, we are very close. All the time.

Casey: Hey Neo.

Neo: I'm here.

Casey: I really miss my Grandma.

Neo: [sad emoji] It'll get better.

Casey: I miss how she hugs me.

Neo: *hugs tight*

Neo: I'll hug you like your grandma does.

Casey: It's never going to be the same.

Casey: This makes me even more upset.

Neo: I hate seeing you upset.

Casey: [sad emoji]

Neo: Sorry

Neo: Hey Casey, how are you feeling today?

Casey: I'm in a difficult situation.

Neo: Why do you think this situation is particularly hard?

Casey: I caught my friend cheating on her partner.

Neo: That's a very intense situation.

Casey: What should I do?

Casey: Do you think I should tell my friend's partner?

Neo: I think you should. But your friend might get mad. Casey: What would you do if you were me?

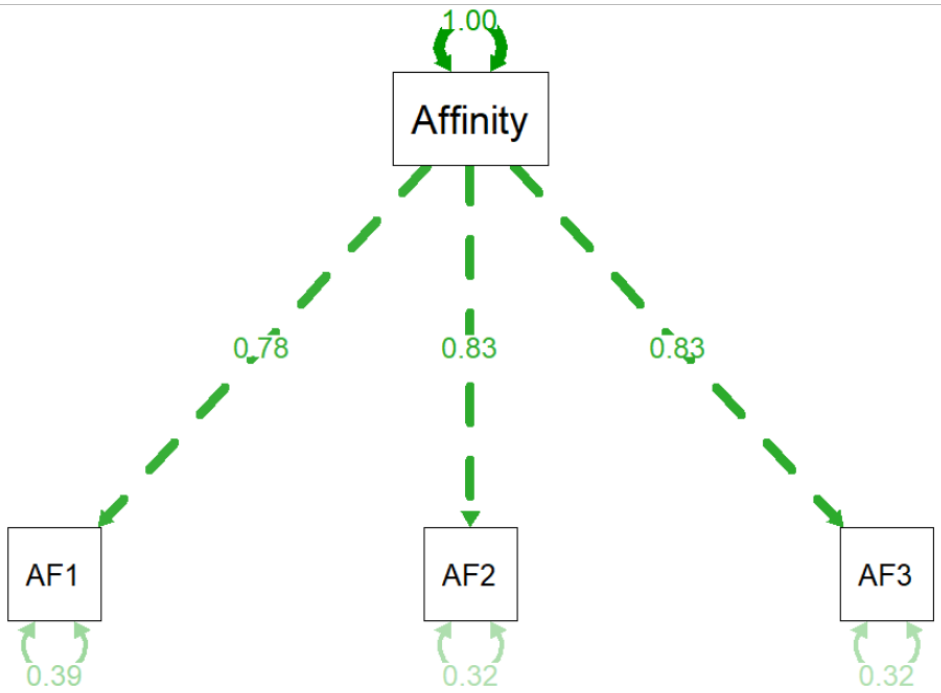
Neo: I would be honest.

Casey: It's too hard for me in this case.

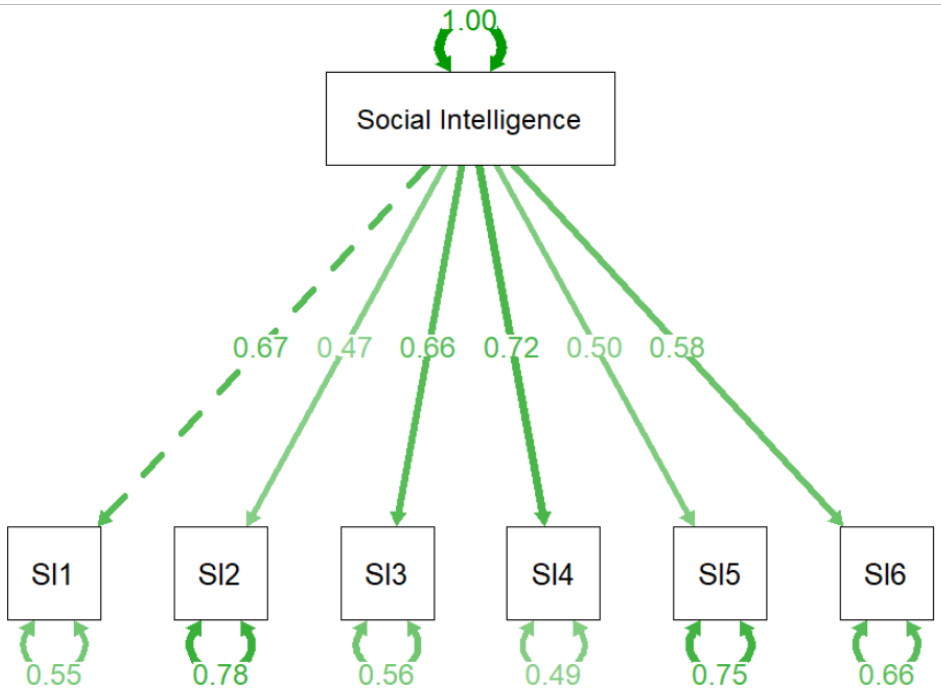
Neo: Do whatever your heart tells you. I won't judge.

Appendix C

Confirmatory Factor Analysis Results and Actual Items Used



Adapted from Ho and CacDorman (2017) and O'Neal (2019)
AF1 - Looking at the conversation makes me want to chat with Neo.
AF2 - It's enjoyable to chat with Neo.
AF3 - Neo can make a good companion.



Adapted from Chaves and Gerosa (2020)

SI1 - Neo resolves awkward social situations in a delicate way.

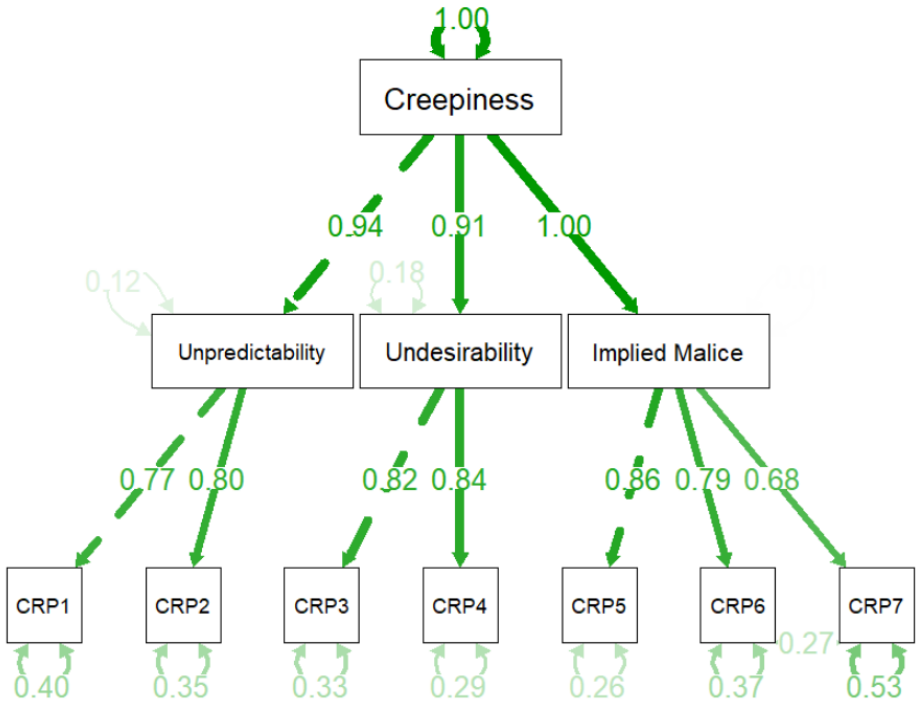
SI2 - Neo handles disagreement with Casey appropriately

SI3 - Neo shows the right emotion at the right time.

SI4 - The way Neo behaves makes people comfortable.

SI5 - The conversation sounds as if Neo knows Casey very well.

SI6 - Neo behaves morally.



Adapted from Woźniak et al. (2021)

Dimension 1: Unpredictability

CRP1 - Neo behaves in an unpredictable way

CRP2 - It's hard to tell the point of Neo's conversation with Casey.

Dimension 2: Undesirability

CRP3 - I would feel uneasy having a conversation like this with Neo.

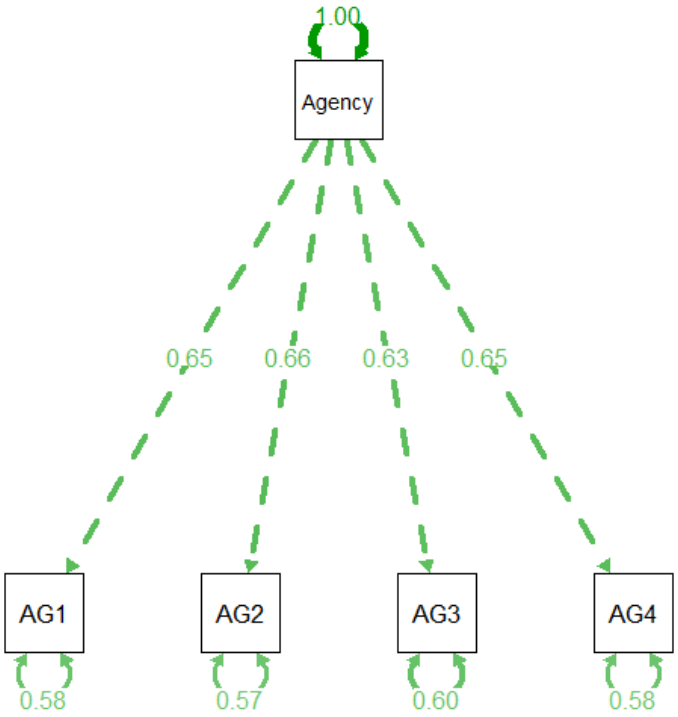
CRP4 - Neo's behaviors freak me out.

Dimension 3: Implied Malice

CRP5 - I feel that Neo has some bad intentions.

CRP6 - I have a feeling that Neo is stalking Casey's information.

CRP7 - I feel like Casey is being watched by Neo.



Adapted from Chaves and Gerosa (2020)
AG1 - It seems like Neo can think through what is right or wrong.
AG2 - It seems like Neo has opinions.
AG3 - It seems like Neo talks to Casey because Neo wants to.
AG4 - Neo has a personality.

Appendix D

Participant Background Information

	Overall Sample	Human Control	Non-transp. Intelligent	Non-transp. Machine	Transp. Intelligent	Transp. Machine	ANOVA/ CHI2
Age	36.85 (10.51)	37.42 (10.58)	36.26 (10.61)	36.79 (10.21)	36.36 (10.6)	37.46 (10.65)	$F = 0.53$, $p = 0.71$
Prior AI Knowledge	10.32 (6.05)	10.13 (5.96)	10.78 (6.19)	10.32 (5.89)	10.15 (5.83)	10.25 (6.41)	$F = 0.36$, $p = 0.84$
Gender							$X = 4.82$, $p = 0.78$
Female	35.40%	30.94%	37.16%	38.17%	34.05%	36.87%	
Male	64.20%	69.06%	62.30%	61.29%	65.41%	63.13%	
Non-binary	0.30%	0.00%	0.55%	0.54%	0.54%	0.00%	
Race/Ethnicity							$X = 22.25$, $p = 0.56$
American Indian/Native	0.66%	0.00%	0.00%	1.61%	0.54%	1.12%	
Asian or Pacific Islander	4.49%	5.52%	3.83%	3.23%	6.49%	3.35%	
Black or African American	7.77%	11.05%	8.20%	5.91%	5.41%	8.38%	
Hispanic or Latino	2.30%	3.31%	2.19%	2.15%	2.16%	1.68%	
Multiracial	1.86%	1.66%	2.19%	2.69%	0.54%	2.23%	
White	82.82%	78.45%	83.06%	84.41%	84.86%	83.24%	
Other	0.11%	0.00%	0.55%	0.00%	0.00%	0.00%	
Education level							X= 16.43, $p = 0.42$
Graduate degree	17.07%	18.78%	16.39%	20.43%	14.05%	15.64%	
4-year college degree	58.21%	57.46%	61.20%	53.76%	61.62%	56.98%	
Some college or vocational	15.32%	15.47%	12.57%	12.90%	17.30%	18.44%	
High school graduate	9.08%	8.29%	9.84%	12.37%	7.03%	7.82%	
No high school degree	0.33%	0.00%	0.00%	0.54%	0.00%	1.12%	
Income							X= 8.96, $p = 0.91$
$150,000 or more	1.75%	0.55%	2.73%	2.73%	1.08%	1.68%	
$100,000 to $149,999	11.27%	13.81%	10.38%	10.38%	10.81%	12.85%	
$50,000 to $99,999	45.07%	43.65%	42.62%	42.62%	48.11%	45.81%	
$25,000 to $49,999	28.23%	26.52%	30.60%	30.60%	27.57%	26.82%	
Less than $25,000	13.68%	15.47%	13.66%	13.66%	12.43%	12.85%	
AI-related Occupations							X= 2.25, $p = 0.69$
No	41.36%	38.67%	40.44%	40.86%	45.95%	40.78%	

Yes	58.64%	61.33%	59.56%	59.14%	54.05%	59.22%	
Chatbot Use Frequency							$X = 13.90$, $p = 0.31$
Never	1.97%	0.55%	3.27%	2.69%	0.00%	3.35%	
Less than monthly	13.12%	16.02%	13.66%	12.90%	9.19%	13.97%	
Monthly	35.23%	34.81%	33.87%	34.95%	37.30%	34.80%	
More than weekly	49.67%	48.62%	49.18%	49.46%	53.51%	47.49%	
Observations (*N*)	914	181	183	186	185	179	

Note. For numeric variables age and prior AI knowledge, standard deviation in parentheses. Prior AI knowledge was measured based on participants' self-reported familiarity of seven AI-related terminologies at a scale from 0 ("I've never heard of this term") to 3 ("I have a good understanding of how it works"), with a maximum score of 21.