BMC Psychology

RESEARCH Open Access

Counselor type (Human/AI) and consultation intention: a moderated mediation model of trust and psychological counseling scenarios

Wei-zhong Zhang[1*] and Rong Lian[2]

Abstract

Objectives In some services involving social and emotional needs (such as psychological counseling), people seem to prefer human services over AI. Exploring the psychological mechanisms behind this preference can help increase people's acceptance of AI-based psychological counseling. This study aims to explore the differences in people's consultation intention for human versus AI across different counseling scenarios, as well as the role of trust (including cognitive and emotional trust) between humans and AI.

Methods A total of 477 participants (297 in Study 1 and 180 in Study 2) were randomly assigned to different groups for counseling imagination tasks and then completed self-report questionnaires.

Results The results of Study 1 demonstrated a significantly higher consultation intention towards human counselors in social emotional scenarios, while no significant preference was observed in cognitive analytical scenarios. Study 2 replicated the findings of Study 1, and further revealed that: (1) cognitive and affective trust played a multiple mediating role between counselor type and consultation intention in social emotional scenarios; (2) there existed a suppressing effect in the relationship model between cognitive trust, counselor type, and counseling intentions in cognitive analytical scenarios; (3) psychological counseling scenarios moderated the relationship between cognitive trust/counselor type and consultation intention.

Conclusion These findings offer practical guidance for the development of AI-driven psychological consultation products and carry theoretical implications for research pertaining to human-AI interaction.

Keywords AI, Consultation intention, Trust, Psychological counseling scenarios, Algorithm aversion

*Correspondence:
Wei-zhong Zhang
1358860312@qq.com
[1]School of Education, Fujian Polytechnic Normal University, Fuzhou, Fujian 350300, China
[2]School of Psychology, Fujian Normal University, Fuzhou, Fujian 350007, China

Introduction

The scarcity of human resources has presented a problem in the field of psychological health services for a significant period, prompting the exploration of integrating Artificial Intelligence (AI) into psychological counseling as a promising area. Various studies had demonstrated the effectiveness of psychological health counseling chatbots, such as Woebot and Gabby, in mitigating symptoms of depression and automating certain elements of clinical therapy [1, 2]. Moreover, these applications exhibited perspective in terms of efficiency and cost-effectiveness within the realm of online self-help interventions [3]. Such researches provided positive evidence for the practical implementation of AI in psychological counseling services.

However, numerous doubts and concerns persisted among individuals regarding the utilization of AI for addressing psychological problems, thereby fostering a sense of distrust in AI and influencing their intention to utilize it [4–6]. Consequently, it was imperative to delve into people's avoidance responses in various psychological counseling scenarios, along with investigating the underlying mechanisms and conditions that influenced their intention to use AI-driven psychological counseling.

Algorithm aversion and consultation intention

People's intention to utilize AI products has not been particularly robust, despite the evident advantages of AI across various domains. Philosopher Bostrom (2014) expressed concerns regarding potential disasters arising from machines making decisions on behalf of humans, while Elon Musk referred to the ascent of automated machines as the "biggest existential threat" to humanity [7, 8]. Some empirical studies have also indicated that the general public prefers human decision-making over AI decisions [9–11].

The tendency to reject AI products may be due to "algorithm aversion." Algorithms refer to a series of computational steps and rules employed to solve problems or accomplish tasks [12]. Algorithms are the core components of AI systems as they can process large amounts of data, extract useful information, and make appropriate decisions or generate corresponding outputs, allowing AI to mimic or even surpass human performance in certain tasks. Numerous studies have identified the phenomenon of "algorithm aversion" across different AI application scenarios [13–15]. The aversion to algorithms manifests in the cognition and affective responses of individuals [16]. Specifically, individuals display a sense of distrust in the capabilities of algorithms and experience negative emotions when utilizing algorithm decisions or encountering their outcomes [14], even resorting to moral blame [17]. For instance, in the healthcare domain, patients harshly criticized doctors who sought advice from algorithms instead of consulting their peers [18]. When evaluating completely identical artworks, individuals exhibit a stronger preference for those created by humans rather than algorithms [15].

Limited contact with algorithms leads to an asymmetry in perception wherein individuals perceive algorithms as lacking in good decision-making abilities. The short history of interaction between humans and algorithms, along with the nontransparency and unexplainability of algorithms, reinforces the sense of distance between humans and algorithms [19]. This perceptual asymmetry in algorithm transparency weakens people's confidence in understanding algorithm decisions, resulting in an objective lack of understanding of algorithms and a subjective underestimation of their own level of understanding. As a result, resistance to algorithms is reinforced [6, 20, 21]. Additionally, individuals often exhibit skepticism towards the professional competence of algorithms, perceiving that algorithm decisions perform worse than those made by humans [22]. For instance, the reason why people are often unwilling to follow medical algorithm recommendations is that they fail to perceive algorithms as possessing the professional medical competence to provide good suggestions [23].

People also exhibit resistance towards algorithms due to a perceived deficiency in "experience." The Mind Perception Theory presents the notion that individuals tend to deconstruct the concept of "mind" into two dimensions: perceived agency and experience [24, 25]. Agency refers to the ability for analysis and reasoning, while experience involves the capacity for feeling and empathy. Researchers propose that although people might acknowledge algorithms surpassing human intelligence, they still regard "experience" as an exclusively human trait [26]. In other words, people generally believe that human problem-solving approaches are more flexible and humane, whereas algorithmic approaches are more mechanistic and devoid of affective qualities. Psychological counseling processes have distinct social characteristics, such as emotions and interactions, which significantly influence individuals' willingness to continue counseling [27]. From the perspective of the Mind Perception Theory, if individuals perceive a deficiency in the experiential dimension of algorithms, they may negate their humanity and consequently resist engaging in equal interaction with algorithms.

Based on these reviews, the present study hypothesized that:

Hypothesis 1 People's consultation intention toward human counselors was significantly higher than with AI.

Cognitive and affective trust as parallel mediations

Trust promotes the client's consultation intention

Trust is essential in traditional psychological counseling. The Social Exchange Theory posits that the exchange of benefits in social interactions is a voluntary action, and these benefits are not generated based on calculations but rather founded on trust [28]. In traditional psychological counseling, trust plays a vital role in establishing a therapeutic alliance and serves as an inherent mechanism for effective therapy [29]. When a trusting relationship exists between clients and counselors, they are more inclined to engage in genuine dialogues, express their inner concerns, and readily accept advice and support from the counselors.

Emotional trust and cognitive trust in psychological counseling

The trust in psychological counseling includes emotional trust and cognitive trust. Traditional research on interpersonal trust suggests that trust is a multidimensional concept, related to the characteristics, intentions, and behaviors of the interacting entities [30, 31]. These concepts of interpersonal trust have been extended to the relationship between humans and technology [4, 32, 33]. For example, Choung et al. focused on anthropomorphic trust (benevolence and integrity) and functional trust in AI [4]. Huang et al. further classified interpersonal trust in the organizational field and differentiated AI trust into cognitive trust and affective trust [34]. It suggests that trust, whether between individuals or between humans and AI, can be primarily summarized into two dimensions: the rational perception of information accuracy, objectivity, and reliability provided by the interacting entities, and the perception of the social and emotional aspects of the interacting entities [34–36]. Thus, this study adopts the classification of cognitive trust and affective trust in human-AI counseling proposed by Huang et al. (2023) [34]. Cognitive trust refers to people's confidence in the competence and reliability of the service provider, while affective trust is more rooted in emotional communication and connection between the individual entities [31, 34].

Trust between humans and trust between humans and AI

Trust is important, but people seem to be more willing to trust humans rather than AI in certain tasks. Trust also aids in facilitating people's intention to seek counseling in AI counseling scenarios [37]. Studies on human-computer interaction have found that trust increases people's reliance and cooperation levels with AI, as well as promotes acceptance of AI as a cooperative partner [38–40]. Many studies suggest that empathy, personalization, and explainability in AI can enhance people's trust in it [5, 19, 41], but building trust becomes challenging when people realize they are interacting with AI. The Machine Heuristic Model posits that when people perceive they are interacting with a machine rather than a human, they automatically activate stereotypes about machines [42]. While people may accept that AI surpasses humans in terms of capabilities, they still view "feeling" as an exclusive human trait [43]. Additionally, human-AI interactions increase the uncertainty of the interaction [44]. When faced with uncertain situations, people tend to raise their expectations for trust [37, 44], which can lead to distrust in AI and subsequently refusal to utilize AI-driven psychological counseling. Therefore, people may exhibit varying degrees of emotional and cognitive trust toward human and AI counselors, which in turn can influence their willingness to engage in counseling.

Based on the previous reviews, this study hypothesized that:

Hypothesis 2 Cognitive trust and affective trust were proposed as parallel mediators in the relationship between counselor type (human/AI) and consultation intention.

Social emotional scenario as a moderator

According to the Task-Technology Fit Theory, a positive impact on outcomes is observed when there is a match between the characteristics of a technology and the features of a task [45]. The problem-solving approaches employed by humans are considered to be more flexible and compassionate, whereas AI is viewed as a tool capable of assisting in data analysis and processing but lacking the ability to handle individualized or exceptional situations due to its mechanistic processing patterns and limited emotional understanding [46]. For instance, in tasks characterized by a mechanical nature such as work assignment and scheduling, AI decisions are believed to be equally fair and reliable as human decisions. However, in more human-centric tasks like employee recruitment and job evaluations, AI decisions are perceived as less fair, less reliable, and more likely to elicit negative emotions compared to human decisions [47]. Additionally, research indicates that regarding subjective matters like dating advice, people tend to seek recommendations from humans, whereas for objective matters like economic advice, AI suggestions are favored [48]. In other words, individuals' intention to utilize AI is influenced by the specific task at hand, with a preference for human advice in subjective and human-centric affairs.

In AI psychological counseling, people's intention to seek advice may vary depending on the emphasis on cognitive analytical or social affective issues. Educating the public about mental health knowledge and providing psychological counseling support are vital components of the social mental health service system [49, 50]. In knowledge education scenarios that emphasize cognitive

analytical aspects, such as concepts, theories, and principles in the field of psychology, AI may be highly applicable due to its rational and objective cognitive capabilities. Conversely, in psychological counseling scenarios that emphasize social emotional analysis, such as addressing personal-centered emotional distress, psychological stress, interpersonal relationships, and coping strategies, people not only expect counselors to possess professional competence but also seek social and emotional support. As a result, they may exhibit higher resistance towards AI.

Based on these perspectives, this study refers to Wirtz et al.'s (2018) research on service robots and categorizes psychological counseling scenarios into cognitive analytical and social emotional ones [51]. The present study proposes the following hypotheses:

Hypothesis 3 Psychological counseling scenarios moderated the relationship between the counselor type (human/AI) and consultation intention.

Hypothesis 4 Psychological counseling scenarios moderated the relationship between the cognitive trust and consultation intention.

Hypothesis 5 Psychological counseling scenarios moderated the relationship between the affective trust and consultation intention.

The current study

There have been a number of binary comparison studies on AI-human decision preferences, uncovering the phenomenon of people exhibiting a certain degree of aversion towards AI [11, 26, 47]. As AI psychological counseling is a potential direction for AI applications, further research is needed to explore this phenomenon, particularly considering its emphasis on the unique nature of cognitive, social, and emotional interaction between humans and AI. Therefore, Study 1 aims to preliminarily explore people's consultation intention with humans/AI and any differences therein. Study 2 aims to validate a moderated mediation model in order to explore the underlying mechanisms and conditions of people's consultation intention with humans/AI. The current study can help us explore the optimal application potential and future development directions of AI-based psychological counseling, considering the current level of AI development and people's cognitive characteristics.

Table 1 Description of sample characteristics

	Study 1	Study 2
Sex		
male	102 (34.3%)	68 (37.8%)
female	195 (65.7%)	112 (62.2%)
Age	22.28 ± 2.54	21.11 ± 2.40
Profession		
Student	28 (70.4%)	113 (62.8%)
Enterprise staff	59 (19.9%)	43 (23.9%)
Freelance work	209 (9.5%)	24 (13.3%)
Total	297	180

Study 1: a preliminary exploration of the consultation intention to human/ai in different psychological counseling scenarios

Methods

Participants

The current study utilized the online survey platform "Wenjuanxing" to distribute the experiment. A total of 319 participants were recruited, and 297 (93.1%) valid responses were obtained after eliminating data that did not pass the attention check items. In the sample, there were 102 males (34.3%) and 195 females (65.7%), see Table 1. The participants were aged from 17 to 35 years ($M = 22.28$ years, $SD = 2.54$). The participants were randomly assigned to the following groups: Human-social emotional group ($n = 76$), Human-cognitive analytical group ($n = 76$), AI-social emotional group ($n = 69$), and AI-cognitive analytical group ($n = 76$). All participants voluntarily participated in the experiment and provided informed consent.

Procedure

Study 1 consisted of four experimental groups, Human-social emotional group, Human-cognitive analytical group, AI-social emotional group, and AI-cognitive analytical group. Participants were randomly assigned to one of these groups by computer. As shown in Fig. 1, firstly, participants were provided with an overview of the study's objectives and informed consent was obtained. Next, in the main experiment, participants were sequentially presented with the brief introduction of the counselor type (Human/AI), problem scenario materials, and response materials. Thirdly, to ensure that participants had read and understood the scenario materials, participants were asked to answer a manipulation check items (e.g., "In the introduction of Mr. Wang, Is Mr. Wang human or AI robot?" 1 = Human, 2 = AI robot). If the question was not answered correctly, the participant's data would be rejected. Finally, participants were given a questionnaire to gather demographic information and assess they consultation intention.

The social emotional scenario in Study 1 revolved was a psychological health problem caused by a break-up (e.g., "You were fell in love with your partner for three years but recently he/she suggested breaking up. You feel very sad… As time passed, you become more and more depressed, and begin to experience anxiety and self-doubt. So, you seek advice from Mr. Wang to help you get

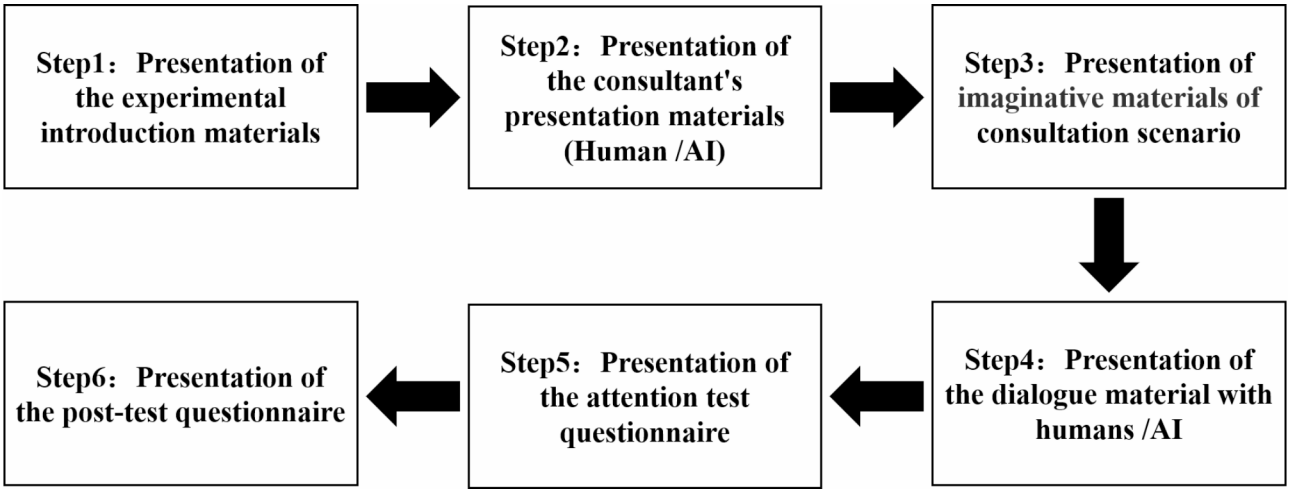

Fig. 1 Flow chart of the experiment

through this terrible time"). The cognitive analytical scenario was related to counseling about "What was exposure therapy". In the current study, response materials were generated by Chat GPT 3.5.

Measures

Consultation intention scale. Consultation intention scale was adapted from previous researches and measured using three items [52, 53]. For example, "I am very willing to accept (AI or human' name)'s answer", and "If there are other relevant questions, I am willing to continue to further communicate with (AI or human' name)". Each item was rated on a 7-point Likert scale ranging from 1 = strongly disagree to 7 = strongly agree. The scale demonstrated good internal consistency reliability with Cronbach's α coefficients of 0.88 in study 1.

Statistical analyses

Study 1 aimed to preliminary explore the differences in individuals' intention for counseling to human vs. AI counselors. Therefore, independent samples t-tests were conducted on the data from the human and AI groups separately, in both social emotional scenario and cognitive analytical scenarios. Additionally, to ensure the robustness of the results, the participants' gender (male = 1, female = 2) and age were included as covariates and analyzed using analysis of variance (ANCOVA).

Results

In the social emotional scenario, the independent samples t-test revealed that consultation intention in the human group ($M = 5.58$, $SD = 1.08$) was significantly higher than AI group ($M = 5.01$, $SD = 1.10$), t (1, 143) = 3.18, $p < 0.01$, *Cohen's d* = 0.53. After controlling for gender and age, the ANCOVA results showed that consultation intention in the human group remained significantly higher than AI group, F (1, 141) = 10.06, $p < 0.05$, $\eta^2 = 0.07$.

In the cognitive analytical scenarios, the independent samples t-test revealed that consultation intention in the human group ($M = 5.91$, $SD = 0.90$) was significantly higher than AI group ($M = 5.68$, $SD = 0.85$), t (1, 150) = 1.59, $p = 0.11$, with *Cohen's d* = 0.26. After controlling for gender and age, the ANCOVA results showed that consultation intention in the human group remained no significantly with AI group, F (1, 148) = 2.25, $p = 0.14$, η^2=0.02.

The preliminary findings of study 1 confirmed hypothesis 1, indicated that even when AI and humans provided the same answers in the social emotional scenario, individuals' intention for counseling with human counselors was significantly higher than AI. However, no such difference was found in the cognitive analytical scenario. To further enhance the robustness of the results obtained in study 1, study 2 expanded on the counseling scenario and delved into the underlying mechanisms by examining the mediating roles of cognitive and affective trust, as well as the moderating effect of psychological counseling scenarios.

Study 2: exploring the mechanisms of consultation intention

Methods

Participants

Study 2 utilized the online survey platform "Wenjuanxing" to distribute the experiment. A total of 185 participants were recruited, and 180 (97.3%) valid responses were obtained after eliminating data that did not pass the attention check items. In the sample, there were 68 males (37.8%) and 112 females (62.2%), see Table 1. The participants were aged from 17 to 31 years ($M = 21.11$ years, $SD = 2.40$). The participants were randomly assigned to

the following groups, Human-social emotional group ($n = 49$), Human-cognitive analytical group ($n = 43$), AI-social emotional group ($n = 45$), and AI-cognitive analytical group ($n = 43$). Participants voluntarily participated in the experiment and provided informed consent.

Procedure

The procedure for study 2 was the same as study 1. In study 2, the social emotional scenarios consisted of psychological health problem caused by interpersonal conflicts (e.g., "You are a student at a university, and you have interpersonal conflicts with your roommate named Xiao Li. You feel that Xiao Li always plays loud music in the dormitory, and… This interpersonal conflict issue has had a negative impact on your psychological well-being. You have started feeling anxious and experiencing insomnia… So, you want to seek advice from AI/Human to help you overcome this difficult time"). The cognitive analytical scenarios involved consulting on knowledge related to childhood autism. After reading the scenario materials and completing an attention check, participants given a questionnaire to gather demographic information and assess they consultation intention, cognitive trust and affective trust.

Measures

Consultation intention scale. Same as study 1. The scale demonstrated good internal consistency reliability with Cronbach's α coefficients of 0.87 in study 2.

Cognitive trust and affective trust Scale. Cognitive trust and affective trust Scale was adapted from Huang et al.'s measurement of trust in robots [34]. Each subscale consisted of three items, such as "I trust (AI or human' name) because it will handle the problem professionally" and "I trust (AI or human' name) because it makes me feel comfortable and at ease." The items were scored on a 7-point Likert scale ranging from "1 = strongly disagree" to "7 = strongly agree." Higher scores indicate a higher level of cognitive or affective trust in the human/AI in the given scenarios. The total scale demonstrated good internal consistency reliability with Cronbach's α coefficients of 0.92, and 0.81, 0.92 for cognitive trust and affective trust, respectively.

Table 2 Scores ($M \pm SD$) of each group on different variables

	Consultation intention	Cognitive trust	Affective trust
Human- counseling question($n = 49$)	5.50 ± 0.92	5.59 ± 0.94	5.52 ± 1.02
Human-knowledge question($n = 43$)	5.30 ± 1.43	5.43 ± 1.16	5.25 ± 1.31
AI- counseling question($n = 45$)	4.65 ± 1.23	4.64 ± 1.16	4.39 ± 1.33
AI- knowledge question($n = 43$)	5.09 ± 0.77	4.77 ± 0.75	4.44 ± 1.04

Statistical analyses

First, to validate the robustness of the findings in Study 1, a two-way analysis of variance (ANOVA) was conducted with consultation intention as the dependent variable and counselor type (AI or human) and psychological counseling scenarios (cognitive analytical, social emotional) as independent variables.

Second, to test the mediating role of cognitive and affective trust, the Model 4 in PROCESS macro by Hayes (2013) for SPSS was utilized [54]. Gender and age were controlled for, cognitive trust and affective trust were proposed as parallel mediators in the relationship between counselor type (human = 0, AI = 1) and consultation intention. A bootstrap sample size of 5000 was used, employing bias-corrected bootstrapping with a 95% confidence interval. This analysis was conducted separately for the two psychological counseling scenarios to examine the mediating effects.

Finally, a moderated mediation analysis was performed using the Model 15 in PROCESS macro for SPSS developed by Hayes [54]. Gender and age were controlled for, and a moderated mediation model was constructed with cognitive trust and affective trust as the mediating variables and psychological counseling scenarios as the moderating variable. Furthermore, in order to understand the essence of the moderation effect, simple slope tests were conducted [55].

Result

Descriptive and differential analysis

The scores of cognitive trust, affective trust, and consultation intention in each group were shown in Table 2. The results of One-way analysis of variance (ANOVA) showed a significant main effect of the counselor type, $F(1, 174) = 10.57$, $p < 0.001$, $\eta^2 = 0.06$. People exhibited significantly higher consultation intention towards human counselor compared to AI. The main effect of psychological counseling scenarios was not significant, $F(1, 174) = 0.61$, $p = 0.44$, $\eta^2 = 0.003$. The interaction effect between counselor type and psychological counseling scenarios was not significant, $F(1, 174) = 2.68$, $p = 0.10$, $\eta^2 = 0.02$. Planned contrast analyses indicated that the AI-social emotional group ($M = 4.65$, $SD = 1.23$) had significantly lower scores compared to the Human-social emotional group ($M = 5.50$, $SD = 0.92$), $p < 0.001$. While the score of AI-cognitive analytical group ($M = 5.09$, $SD = 0.77$) had no significantly difference with human-cognitive analytical group ($M = 5.30$, $SD = 1.43$), $p = 0.37$. The results reconfirmed hypothesis 1, indicated that people's preference for human counselors in the social emotional scenario was robust.

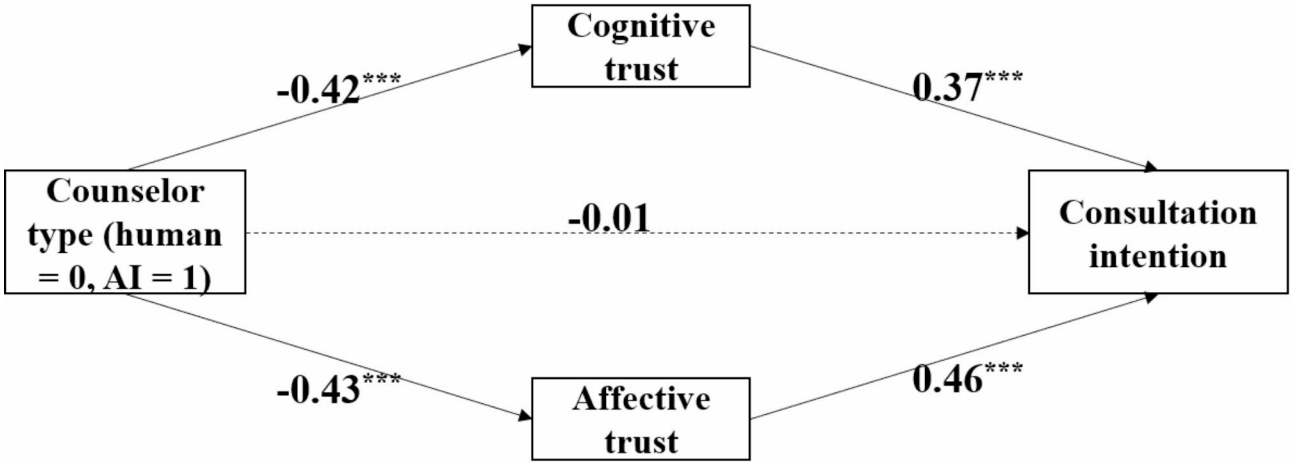

Fig. 2 The fully mediating role of cognitive and affective trust in the social emotional scenario

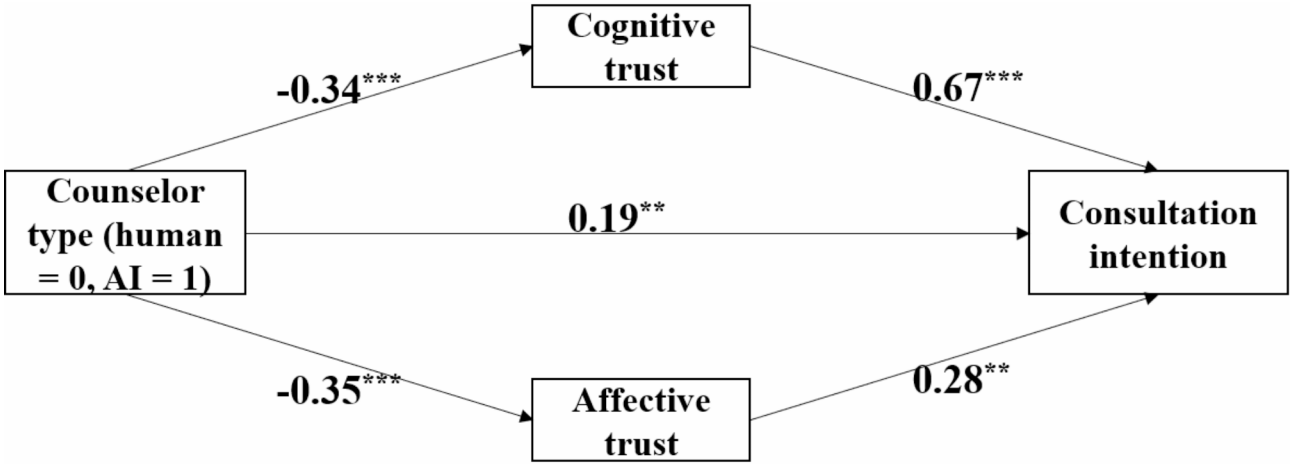

Fig. 3 The suppression effect in the cognitive analytical scenario

The mediating role of cognitive and affective trust

As shown in Fig. 2, in the social emotional scenario, the counselor type (human/AI) significantly and negatively predicted cognitive trust (β = -0.43, $p < 0.001$) and affective trust (β = -0.42, $p < 0.001$). Moreover, cognitive trust ($\beta = 0.37$, $p < 0.001$) and affective trust ($\beta = 0.46$, $p < 0.001$) significantly and positively predicted consultation intention. However, the direct effect between the counselor type and consultation intention was not significant (β = -0.01, $p = 0.92$). These results suggest that cognitive trust and affective trust play a completely mediating role between the counselor type and consultation intention in the social emotional scenario.

Note *ns* = no significant, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The coefficients in the figure are standardized coefficients.

As shown in Fig. 3, in the cognitive analytical scenario, the counselor type (human/AI) significantly and negatively predicted cognitive trust (β = -0.34, $p < 0.001$) and affective trust (β = -0.35, $p < 0.01$). Moreover, cognitive trust ($\beta = 0.67$, $p < 0.001$), affective trust ($\beta = 0.28$, $p < 0.01$) and counselor type ($\beta = 0.19$, $p < 0.01$) significantly and positively predicted consultation intention. These results suggest that cognitive trust and affective trust play a partial mediating role between the counselor type and consultation intention in the cognitive analytical scenario.

These results supported hypotheses 2. Cognitive trust and affective trust were proposed as parallel mediators in the relationship between counselor type (human/AI) and consultation intention.

Examination of the moderated mediation model

The results of testing a moderated mediation model indicated that the interaction between cognitive trust and psychological counseling scenarios significantly predicted consultation intention ($\beta = 0.16$, $t = 2.15$, $p < 0.05$). Similarly, the interaction between the counselor type and psychological counseling scenarios significantly predicted consultation intention ($\beta = 0.10$, $t = 2.19$, $p < 0.05$). However, the interaction between affective trust and

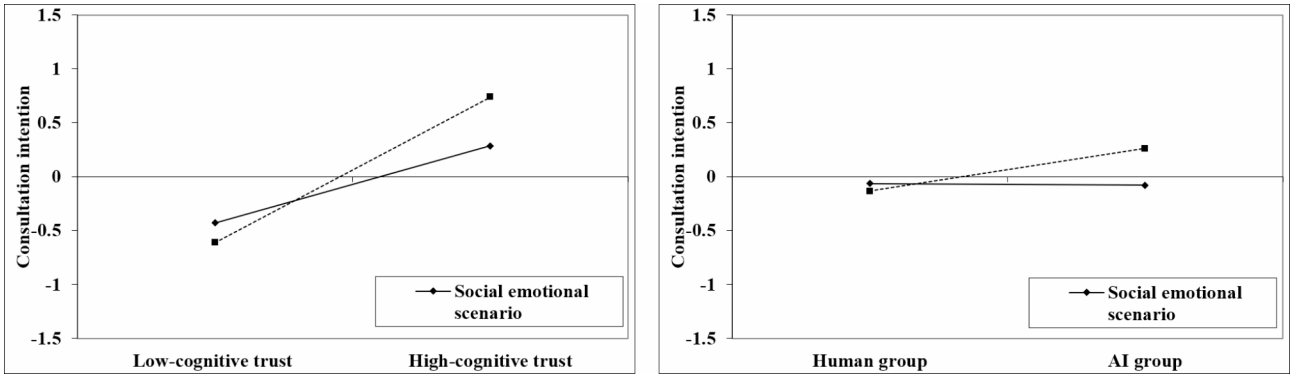

Fig. 4 The moderating effect of psychological counseling scenarios

social emotional scenario did not significantly predict consultation intention (β = -0.09, t = -1.23, $p > 0.05$). These results confirmed hypothesis 3 and 4, but did not provide support for hypothesis 5, which indicated that psychological counseling scenarios moderated the relationship between the counselor type/ cognitive trust and consultation intention.

The results of the simple slope test were shown in Fig. 4. In the social emotional scenario, the predictive effect of the counselor type on consultation intention was not significant (β_{simple} = -0.01, t = -0.06, $p > 0.05$), while the predictive effect of cognitive trust on consultation intention was significant ($\beta_{simple} = 0.36$, $t = 3.46$, $p < 0.001$). In the cognitive analytical scenario, the predictive effect of the counselor type on consultation intention was significant ($\beta_{simple} = 0.20$, $t = 2.99$, $p < 0.01$), as well as the predictive effect of cognitive trust on consultation intention ($\beta_{simple} = 0.68$, $t = 6.86$, $p < 0.001$).

Discussion

The present study aims to investigate the direct impact of counselor type (human/AI) on consultation intention. Furthermore, this study examines the mechanism of cognitive trust and affective trust in the relationship between counselor type and consultation intention, as well as the moderating role of psychological counseling scenarios on the relationship between counselor type/cognitive trust and consultation intention.

Algorithm aversion in social emotional scenario

The results indicate that individuals exhibited algorithm aversion in the social emotional scenario. Even when the same answers were provided in identical scenarios, individuals showed a significantly higher intention to consult with human counselors than with AI. After cognitive and affective trust were added to the model, it was found that both cognitive trust and affective trust fully mediated the relationship between counselor type and consultation intention. It is not surprising that individuals were easier to establish cognitive trust and affective trust in interactions with human counselors [4, 37]. On one hand, these results reflect the specific preferences and psychological needs of individuals for social interaction [34]. When individuals seek counseling for problems of mental health, they often desire emotional connection, an opportunity to express their inner troubles, and to receive emotional support and understanding. On the other hand, the results also shed light on people's negative stereotypical impressions and attitudes towards AI, as lacking experience and being incapable of providing emotional support and understanding similar to human counselors [56]. These perceptions subsequently influence people's intention to utilize AI-driven services.

These findings highlight the significance of establishing social relationships between the client and counselor, whether they are human or AI. Despite researchers suggesting that individuals treat computers in a similar way to social members by incorporating norms, categories, and expectations into human-computer interactions [57], daily encounters with AI often lead to perceptions of AI as lacking complete mental attributes [58, 59] and elicit fewer experiences of social contact during human-AI interactions [9, 60]. According to the Machine Heuristic Model, when individuals perceive they are interacting with a machine rather than a human, it automatically triggers stereotypical impressions about machines, thus influencing their behavior [42]. If individuals perceive AI as lacking experiences, they may deny its humanity and subsequently refuse to engage in equal interaction. Therefore, establishing trust between humans and AI clearly encounters more obstacles compared to establishing trust between humans in social emotional scenarios, resulting in a lower intention to seek consultation.

For AI psychological counselors, it is important to establish emotional trust with clients. However, the limitations of AI counselors in terms of social and emotional capabilities exist. Therefore, it is unreasonable to directly apply theories developed for human-to-human trust relationships. Research has shown that when AI chatbots provide empathetic advice [61] or when AI service agents

are perceived as understanding and displaying human emotions [62], people tend to trust them more and are more willing to adopt the technology. Empathy is a critical skill in psychological counseling, and how to enable people to perceive AI's empathetic capabilities should be an important direction for establishing emotional trust between AI and clients for future research.

AI could serve as a substitute for human in some scenarios

When the task requirements align with the strengths of AI, people are more likely to accept it. This study did not find statistically significant differences in consultation intention between humans and AI in cognitive analytical scenarios. However, when cognitive trust and affective trust were introduced into the model, the counselor type significantly and positively predicted consultation intention. Similarly, the results also show that psychological counseling scenarios moderate the relationship between the counselor type and consultation intention, indicating that in cognitive analytical scenarios, the counselor type (with humans as the baseline) has a stronger predictive ability for consultation intention. This is consistent with previous research findings that people tend to be more inclined to accept AI in contexts that emphasize cognitive analytical [43, 48, 51]. The cognitive advantage in AI's analytical capabilities enables it to meet people's needs for efficient, reliable, and objective services. Therefore, when people realize that AI can better fulfill their requirements in specific tasks, they are more inclined to accept AI as partners or consultants.

Interestingly, there is a suppressing effect in the relationship model between cognitive trust, counselor type, and counseling intentions in cognitive analytical scenarios. In people's perception, AI has a unique advantage in processing large-scale data, conducting complex analysis, and making decisions unaffected by emotions and subjective factors [63]. It is expected that people would find it easier to establish cognitive trust with AI in the cognitive analytical scenario. However, this study reveals contrary findings that the counselor type significantly and negatively predicts cognitive trust and also has a negative impact on consultation intention through cognitive trust in cognitive analytical scenarios. The traditional mediation model introduces a third variable to explain "how X affects Y," and the mediation process provides the "mechanism through which X influences Y." However, in the current study, the sign symbol of the indirect effect is opposite to that of the direct effect, indicating the presence of suppressing effects [64]. In this case, the logical modeling of the mediation model shifts from the traditional mediation model of "how X influences Y" to "how X does not influence Y."

There are some reasons that could explain this result. Firstly, it could be attributed to the fact that during that time, AI systems were not yet highly "intelligent." Research has shown that people have a high sensitivity to errors made by AI algorithms and a low tolerance for them [13, 46]. Therefore, user's past negative experiences might rapidly erode people's trust in AI and make them less inclined to use AI decision-making [22, 65], even in cognitive analytical scenarios that don't emphasize social emotional interaction. Indeed, this further suggests that an important reason for people rejecting AI algorithmic decisions may be a lack of trust in AI algorithms [66], and increasing trust could reduce suspicions about the AI agent and its capabilities. Furthermore, considering that the counselor type negatively affects consultation intention through cognitive trust, it is possible that there are other unaccounted or uncontrolled key factors that positively influence consultation intention. For example, AI and its associated applications demonstrate clear affordability and accessibility, leading to a smaller "intention-action gap" [46]. When faced with relevant issues, seeking advice from a human counselor not only requires more time but sometimes also involves financial costs. In contrast, AI algorithms (programs or applications) with different functionalities are often readily available and free.

Implications

This study has several important implications. Firstly, it represents a pioneering attempt to investigate AI aversion in the context of psychological counseling by examining the impact of the counselor type on consultation intention. The findings reveal that individuals are more inclined to seek further counseling from human counselors in social emotional scenarios, indicating the universality of AI aversion. Secondly, this study extends the concept of interpersonal trust from psychotherapy to the therapeutic relationship between humans and AI. The results validate the mediating role of cognitive and affective trust in the relationship between the counselor type and consultation intention. Furthermore, it suggests that the trust-building process between humans may not directly apply to human-machine trust, although the definition of trust in human-machine interaction is similar to that in human-human interaction [67]. Exploring additional factors that influence the establishment of trust between humans and AI can be helpful for improving people's intention to accept AI psychological counseling. For example, adding transparency design (such as explaining decision logic) and emotional feedback modules to the AI consultation interface can improve user trust. Lastly, this study also examines the differences in people's consultation intention towards humans vs. AI in different psychological counseling scenarios. The results reflect that AI could play a partially substitutive and complementary role in human psychological counseling. It

means that AI pre-processing (e.g. initial screening, questionnaire collection) can reduce the administrative workload of human consultants, allowing them to focus on high-value interventions. Meanwhile, in less developed regions, AI counselling can be used as a transitional solution to alleviate the shortage of professional counsellors and promote universal access to mental health services. Therefore, integrating AI assistants into counseling training programs could help practitioners develop human-AI collaboration competencies, a critical skill for future mental health systems.

Limitations and future research

There are certain limitations of the present study. Firstly, the participants in this study were aged between 17 and 30. However, it is acknowledged that different populations may hold diverse expectations and preferences towards AI, influenced by factors such as educational levels and AI literacy [68]. Thus, future research should aim to include a more representative sample encompassing a broader age range and population. Secondly, given the rapid advancements in AI technology and the increasing prevalence of related products, people's attitudes towards AI may have evolved over time. For example, existing research indicates that human trust in AI robots grows from initially low levels as the level of interaction increases [69]. Incorporating longitudinal tracking methods could provide valuable insights into the development of trust between humans and AI, as well as its impact on consultation intention towards AI. Such an approach would enable a more comprehensive understanding of the dynamics involved. Lastly, it is worth noting that a suppressing effect was observed in the mediation model in the cognitive analytical scenario, indicating that the underlying mechanisms influencing consultation intention were complex and there may be other important factors at play. Exploring the potential mechanisms that shape human/AI psychological consultation intention from multiple perspectives is important for future research.

Conclusion

This study identifies the presence of algorithm aversion in AI psychological counseling, indicating its universality and robustness. Additionally, the concept of trust is extended from human-human psychological counseling to human-AI psychological counseling, and the significance of cognitive trust and affective trust in fostering individuals' intention to seek counseling is empirically verified. Furthermore, AI is found to be a potential substitute in cognitive analytical scenarios, but there remains a need to enhance individuals' cognitive trust in AI. These findings highlight the importance of addressing trust factors in the context of human-AI interactions to improve the effectiveness and acceptance of AI-assisted psychological counseling.

Abbreviations

AI	Artificial Intelligence
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
M	Mean
SD	Standard deviation

Acknowledgements

Not applicable.

Author contributions

Wei-zhong Zhang: Conceptualization, Writing – original draft, Data curation, Writing – original draft. Rong Lian: Supervision, Writing – review & editing.

Funding

This research received no external funding.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

All methods were performed in accordance with the relevant guidelines and regulations. This study was approved by the research ethics committee of School of psychology at Fujian Normal University before data collection. Participants were informed of the study's purpose, procedures, potential risks, and benefits. Written consent was obtained from all participants prior to their inclusion in the study.

Competing interests

The authors declare no competing interests.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 March 2024 / Accepted: 14 April 2025
Published online: 20 April 2025

References

1. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Ment Health. 2017;4(2):e19.
2. Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. JAMA. 2017;318(13):1217–8.
3. Luo B, Lau RYK, Li C, Si YW. A critical review of state-of-the-art chatbot designs and applications. WIREs Data Min Knowl Discov. 2021;12(1).
4. Choung H, David P, Ross A. Trust and ethics in AI. AI Soc. 2022;38(2):733–45.
5. Christoforakos L, Gallucci A, Surmava-Grosse T, Ullrich D, Diefenbach S. Can robots earn our trust the same way humans do?? A systematic exploration of competence, warmth, and anthropomorphism as do?eterminants of trust do?evelopment in HRI. Front Robot AI. 2021;8:640444.
6. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud. 2021;146:102551.
7. Bostrom N, Superintelligence. Paths, strategies, dangers. Oxford, UK: Oxford University Press; 2014.
8. McFarland M. Elon Musk:'With artificial intelligence we are summoning the demon.' The Washington Post Retrieved on Apr 20, 2019 from https://www.washingtonpost.com/news/innovations/wp/2014/10/24. 2014.

9. Acikgoz Y, Davison KH, Compagnone M, Laske M. Justice perceptions of artificial intelligence in selection. Int J Selection Assess. 2020;28(4):399–416.
10. Jones-Jang SM, Park YJ, Yao M. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. J Computer-Mediated Communication. 2023;28(1):1–8.
11. Newman DT, Fast NJ, Harmon DJ. When eliminating bias isn't fair: algorithmic reductionism and procedural justice in human resource decisions. Organ Behav Hum Decis Process. 2020;160:149–67.
12. Kozyreva A, Lorenz-Spreen P, Hertwig R, Lewandowsky S, Herzog SM. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, great Britain, and the United States. Humanit Social Sci Commun. 2021;8(1):1–12.
13. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen. 2015;144(1):114–26.
14. Lee M, Ackermans S, van As N, Chang H, Lucas E, Ijsselsteijn W. Caring for Vincent: a Chatbot for Self-compassion. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; Glasgow, Scotland UK2019. pp. 1–13.
15. Jago AS. Algorithms and authenticity. Acad Manage Discoveries. 2019;5(1):38–56.
16. Zhang Y, Xu L, Yu F, Ding X, Wu J, Zhao L. A three-dimensional motivation model of algorithm aversion. Adv Psychol Sci. 2022;30(5):1093–105.
17. Voiklis J, Kim B, Cusimano C, Malle BF. Moral judgments of human vs. robot agents. In 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); New York, NY, USA IEEE; 2016. pp. 775–80.
18. Shaffer VA, Probst CA, Merkle EC, Arkes HR, Medow MA. Why do patients derogate physicians who use a computer-based diagnostic support system? Med Decis Mak. 2013;33(1):108–18.
19. Zerilli J, Bhatt U, Weller A. How transparency modulates trust in artificial intelligence. Patterns (N Y). 2022;3(4):100455.
20. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav. 2021;5(12):1636–42.
21. Schlicker N, Langer M, Ötting SK, Baum K, König CJ, Wallach D. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. Comput Hum Behav. 2021;122:106837.
22. Prahl A, Van Swol L. Understanding algorithm aversion: when is advice from automation discounted? J Forecast. 2017;36(6):691–702.
23. Promberger M, Baron J. Do patients trust computers? J Behav Decis Mak. 2006;19(5):455–68.
24. Gray K, Young L, Waytz A. Mind perception is the essence of morality. Psychol Inq. 2012;23(2):101–24.
25. Waytz A, Cacioppo J, Epley N. Who sees human?? The stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci. 2010;5(3):219–32.
26. Xie Y, Zhu K, Zhou P, Liang C. How does anthropomorphism improve human-AI interaction satisfaction: a dual-path model. Comput Hum Behav. 2023;148.
27. Xue Y, Xu Z. Impact and application of affective touch on mental health. Adv Psychol Sci. 2022;30(12):2789–98.
28. Nunkoo R, Ramkissoon H. Power, trust, social exchange and community support. Annals Tourism Res. 2012;39(2):997–1023.
29. Fonagy P, Allison E. The role of mentalizing and epistemic trust in the therapeutic relationship. Psychother (Chic). 2014;51(3):372–80.
30. Schoorman FD, Mayer RC, Davis JH. An integrative model of organizational trust: past, present, and future. Acad Manage Rev. 2007;32(2):344–54.
31. Webber SS. Development of cognitive and affective trust in teams. Small Group Res. 2008;39(6):746–69.
32. Calhoun CS, Bobko P, Gallimore JJ, Lyons JB. Linking precursors of interpersonal trust to human-automation trust: an expanded typology and exploratory experiment. J Trust Res. 2019;9(1):28–46.
33. Gillath O, Ai T, Branicky MS, Keshmiri S, Davison RB, Spaulding R. Attachment and trust in artificial intelligence. Comput Hum Behav. 2021;115:106607.
34. Huang D, Chen Q, Huang S, Liu X. Consumer intention to use service robots: a cognitive–affective–conative framework. International Journal of Contemporary Hospitality Management. 2024;36(6):1893-1913.
35. Liu X, Yi X, Wan LC. Friendly or competent? The effects of perception of robot appearance and service context on usage intention. Annals Tourism Res. 2022;92:103324.
36. Wang W, Qiu L, Kim D, Benbasat I. Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. Decis Support Syst. 2016;86:48–60.
37. Park G, Chung J, Lee S. Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention. Current Psychology. 2024;23:4352-4363.
38. Kundinger T, Wintersberger P, Riener A, editors. (Over) Trust in automated driving: The sleeping pill of tomorrow? Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems; 2019.
39. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction ACM, pp 141–1482015. pp. 141-8.
40. Sundar SS, Kim J. Machine heuristic: when we trust computers more than humans with our personal information. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp 538); Glasgow, Scotland, UK: ACM; 2019. pp. 1–9.
41. Watamura E, Ioku T, Mukai T, Yamamoto M. Empathetic robot judge, we trust you. Int J Human–Computer Interact. 2023;40(18):5192–201.
42. Sundar SS. The MAIN model: a heuristic approach to Understanding technology effects on credibility: Cambridge. MA: The MIT Press; 2008. pp. 73–100.
43. Waytz A, Heafner J, Epley N. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol. 2014;52:113–7.
44. Liu B. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. J Computer-Mediated Communication. 2021;26(6):384–402.
45. Zigurs I, Khazanchi D. From profiles to patterns: a new view of task-technology fit. Inform Syst Manage. 2008;25(1):8–13.
46. Du Y. Hate algorithms or appreciate them?— cognitive differences of algorithms and trust construction of algorithms in the era of artificial intelligence. Philosophical Anal. 2022;13(3):151–65.
47. Lee MK. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc. 2018;5(1):205395171875668.
48. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res. 2019;56(5):809–25.
49. Bao Y, Sun Y, Meng S, Shi J, Lu L. 2019-nCoV epidemic: address mental health care to empower society. Lancet. 2020;395(10224):e37–8.
50. Jin Y, Tsai F-S. The promoting effect of mental health education on students' social adaptability: implications for environmental. J Environ Public Health. 2022;2022:1–10.
51. Wirtz J, Patterson PG, Kunz WH, Gruber T, Lu VN, Paluch S, et al. Brave new world: service robots in the frontline. J Service Manage. 2018;29(5):907–31.
52. Park S, Whang M. Empathy in human–robot interaction: designing for social robots. Int J Environ Res Public Health. 2022;19(3).
53. Wu YH, Wrobel J, Cornuet M, Kerherve H, Damnee S, Rigaud AS. Acceptance of an assistive robot in older adults: a mixed-method study of human-robot interaction over a 1-month period in the living lab setting. Clin Interv Aging. 2014;9:801–11.
54. Hayes A. Introduction to mediation, moderation, and conditional process analysis. J Educ Meas. 2013;51(3):335–7.
55. Aiken LS, West SG. Multiple regression: testing and interpreting interactions. Sage Publications, Inc; 1991.
56. Rantanen T, Lehto P, Vuorinen P, Coco K. Attitudes towards care robots among Finnish home care personnel - a comparison of two approaches. Scand J Caring Sci. 2018;32(2):772–82.
57. Nass C, Steuer J, Tauber ER. Computers are social actors. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Boston,Massachusett USA: ACM; 1994. pp. 72–8.
58. Brink KA, Gray K, Wellman HM. Creepiness creeps in: uncanny valley feelings are acquired in childhood. Child Dev. 2019;90(4):1202–14.
59. Swiderska A, Kuster D. Robots as malevolent moral agents: harmful behavior results in dehumanization, not anthropomorphism. Cogn Sci. 2020;44(7):e12872.
60. Noble SM, Foster LL, Craig SB. The procedural and interpersonal justice of automated application and resume screening. Int J Select Assess. 2021;29(2):139–53.
61. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychology Behav Social Netw. 2018;21(10):625–36.
62. Pelau C, Dabija D-C, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Hum Behav. 2021;122:106855.

63. Jiang L, Cao L, Qin X, Tan L, Chen C, Peng X. Fairness perceptions of artificial intelligence decision-making. Adv Psychol Sci. 2022;30(5):1078–92.
64. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. Prev Sci. 2000;1(4):173–81.
65. Filiz I, Judek JR, Lorenz M, Spiwoks M. Reducing algorithm aversion through experience. J Behav Experimental Finance. 2021;31:100524.
66. Lee JD. In: Sons L, editor. Human factors and ergonomics in automation design. John Wiley; 2006.
67. Madhavan P, Wiegmann DA. Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues Ergon Sci. 2007;8(4):277–301.
68. van Berkel N, Goncalves J, Russo D, Hosio S, Skov MB. Effect of information presentation on fairness perceptions of machine learning predictors. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; Yokohama, Japan: ACM; 2021. pp. 1–13.
69. Ullman D, Malle BF. Human-robot trust: just a button press away. Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017.

Publisher's note