An Exploratory Study Into the Impact of AI Literacy Training on Anthropomorphism and Trust in Conversational AI

Geert Wood(✉), Elena Nuñez Castellar, and Wijnand IJsselsteijn

Human Technology Interaction Group, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands
hti@tue.nl
https://research.tue.nl/nl/organisations/human-technology-interaction

Abstract. This exploratory study investigates the impact of AI literacy training on the perception of anthropomorphism and trust in conversational AI. Anthropomorphism, the attribution of human-like characteristics to non-human entities, is increasingly impacting the way we interact with conversational AI systems. The emergence of generative AI systems such as ChatGPT, which are capable of producing highly coherent and contextually relevant responses, has amplified this phenomenon. As these systems become more powerful and prevalent, users may develop different perceptions of AI, leading to shifts in trust, reliance, and expectations, both warranted and unwarranted. This research aims to understand how AI literacy training might affect both conscious and unconscious anthropomorphic tendencies while investigating their effect on trust towards conversational AI. Forty participants were divided into a control group (N = 20) and an experimental group (N = 20). Both groups interacted with the conversational agent ChatGPT3.5 in a pre- and post-test session, between which only the experimental group received a comprehensive 3-hour AI training. The study employed a multi-faceted approach for measuring anthropomorphism, including surveys and sentiment analysis of the linguistic data obtained from the participant's interaction with ChatGPT3.5. Self-reported data suggests the intervention of the AI literacy training did not significantly change anthropomorphism and trust levels. However, sentiment analysis using a newly created linguistic anthropomorphism scale does show a significant decrease in anthropomorphism levels after the intervention. The paper concludes with a discussion of the results and limitations, while providing relevant future research directions in this expanding domain of human-AI interaction.

Keywords: Human-AI interaction · Anthropomorphism and trust · Conversational AI · AI literacy

H. Degen and S. Ntoa (Eds.): HCII 2025, LNAI 15820, pp. 301–322, 2025.
https://doi.org/10.1007/978-3-031-93415-5_18

1 Introduction

The concept of anthropomorphism is a widely discussed topic in the fields of artificial intelligence, cognitive science, social robotics, and human-computer interaction. Although no universally accepted definition exists, past literature regarding the definition of anthropomorphism shares the same key notion of "the tendency to attribute human-like characteristics, emotions and behaviours to non-human objects or concepts" [3,12,13]. In the case of computers, numerous experimental studies have shown that humans interact with computers socially even though we are fully aware of the fact that they are not human [17]. For example, Moon (2000) found that when computers provided more information about themselves, participants were more involved in intimate self-disclosure [22]. Indeed, some of our deeply ingrained interpersonal biases and stereotypes also carry over to interactions with computers. Nass et al. (1997) showed that tutor systems with either male or female voice outputs elicited gender stereotypes, where computers with female voices were perceived as more knowledgeable about topics relating to love and affection, and less knowledgeable about technology than the male-voiced tutor computer [24]. With the rise of generative Artificial Intelligence (AI) systems like ChatGPT, this tendency to anthropomorphize has expanded into conversational AI where the line between human and machine interactions becomes increasingly blurred. We tend to react socially to conversational AI, whether this process of anthropomorphism is mindful or not [32].

Research on anthropomorphism has primarily focused on areas like social robotics, virtual agents and voice-based interfaces, with a growing number of studies exploring its role in AI interactions. As AI systems become more integrated into daily life, understanding and leveraging anthropomorphism will be essential in the design and evaluation of conversational AI, shaping how these systems engage with users. Recent studies [5,26,34,35] have provided compelling insights into the relationship between human-like Conversational Agents and perceived anthropomorphism. We would like to build on these works and extend them in two ways. First, we focus on diversifying the ways in which anthropomorphism can be measured. Previous work has been largely based on measurements of (dispositional) anthropomorphism, such as the measurements proposed by Waytz and Epley (2008), who arguably can be seen as the founding figures of anthropomorphism [38]. Our study extends this work by proposing a multi-pronged approach for measuring both mindful anthropomorphism and mindless anthropomorphism (while taking a person's dispositional anthropomorphism into account). Our approach includes relevant questionnaires (including a set of items that can be analyzed using the psychometric Rasch model) as well as objective analysis of user's verbal cues. In this way, our approach can distinguish between reflexive, automatic anthropomorphism and deliberate attribution of human-like qualities. Where a singular reliance on self-report measures may be sensitive to reporting biases (e.g., experimental demand characteristics, social desirability), the incorporation of behavioral and linguistic markers can provide a more comprehensive assessment that may help overcome such biases.

A second key contribution of our work, we think, is the exploration of potential effects of AI literacy on users' anthropomorphism and trust perceptions.

To the best of our knowledge, none of the studies to date have explored the possible impact an AI literacy training could have on the perception of anthropomorphism and trust towards these conversational AI systems. How does learning about AI's capabilities, limitations, underlying principles, and ethical considerations impact our perception of it? Does increased knowledge foster a more accurate understanding of AI as a technological tool, or do we continue to anthropomorphize these systems? Moreover, how does an AI literacy intervention affect trust in conversational AI? Overall, AI literacy training has the potential to create more critical and informed users, reducing blind anthropomorphism while fostering a clearer understanding of AI's role as a technological tool rather than a sentient entity.

This paper aims to address the questions above by methodically guiding the reader through a comprehensive exploration of the effects that an AI training intervention may have on the perception of anthropomorphism and trust within the field of conversational AI. The background literature delves into the concept of anthropomorphism, its implication in the context of conversational AI and its relationship with trust. The methodology section outlines the research design, the demographics of the participants, and the procedural details. Data analysis and results are presented in a structured manner, capturing the complexity of the findings. Finally, the paper concludes with a discussion of the key findings and acknowledges study limitations that should be considered in future research.

2 Theoretical Background

2.1 Understanding Anthropomorphism

A diversity of definitions of anthropomorphism exist depending on their applicability to specific research domains. For example, Nass and Reeves explored the definition of anthropomorphism in the context of human-computer interaction and showed that we apply social rules when interacting with computers [30]. Waytz et al. demonstrated that the concept of anthropomorphism extends to our interactions with autonomous cars, thereby enhancing our trust in these systems [39]. More recently, research by Salles et al. explores the presence of anthropomorphism within Artificial Intelligence (AI) research, highlighting (ethical) implications such as misleading beliefs about AI's capabilities, perceiving AI as a moral entity (even though it is not) and being the source of overblown fear/uncritical optimism towards AI [32]. Due to this variability in applications, many researchers use a more general and overarching definition of anthropomorphism; "the tendency to attribute human-like characteristics, emotions and/or behaviors to nonhuman objects or concepts" [3,12,13].

The historical roots of anthropomorphism can be traced back to ancient times. The term itself comes from the Greek words "anthropos" (human) and "morphe" (form), and it originally referred to attributing human characteristics to deities. In modern science, ethologists such as Nikolaas Tinbergen and Konrad Lorenz, laid the groundwork for understanding animal behavior through systematic behavioral observations as well as relating such observations to our introspective knowledge of our own minds (i.e., theory of mind). The 20th century saw the rise of anthropomorphic robots in science fiction, such as C-3PO from Star Wars and HAL 9000 from 2001: A Space Odyssey. These characters personified machines with human-like intelligence, emotions, and personalities.

As personal computers became more prevalent, from the early 1980s onwards, researchers in HCI began exploring how users interact with machines. Anthropomorphism was found to improve user engagement and satisfaction, leading to the development of more intuitive interfaces and virtual assistants. Scientific work in this area was spearheaded by Nass, Reeves, and colleagues who, during the 1990s, proposed their "Computers Are Social Actors" (CASA) paradigm [30]. This paradigm was based on the Social Response Theory, suggesting that people (unconsciously) apply social norms and behaviours such as politeness and reciprocity to non-human actors, despite being aware of their non-human properties. This theory referred to human evolutionary development, where humans evolved to be highly attuned to social cues as a survival mechanism. The tendency to anthropomorphize non-human agents likely stems from an adaptive bias: mistaking an inanimate object for a social agent poses less risk than failing to recognize a true social partner or a potential social threat.

Shortly thereafter, Epley studied anthropomorphism as a distinct psychological and behavioral phenomenon. Epley identified two motivational determinants for anthropomorphism; the human need to understand and control one's environment to decrease uncertainty (effectance motivation) and the innate human need for social interaction as a driver of anthropomorphism (sociality motivation) [9]. Waytz further contributed to this research by exploring the sociality motivation, revealing how loneliness can increase the tendency to anthropomorphize. Both Epley and Waytz agree on the claim that anthropomorphism consists of both a mindful and a mindless part. Mindful anthropomorphism occurs when we consciously attribute human characteristics to non-human objects. However, as demonstrated by Nass and Reeves with their CASA paradigm, sometimes we *unconsciously* apply social norms and behaviour to non-human objects. The magnitude of both the mindful and mindless tendency depends on people's individual, natural tendency to anthropomorphize (i.e., dispositional anthropomorphism). This dispositional aspect is shaped by individual differences in culture, norms, experience, education, cognitive reasoning styles, and attachment [16].

2.2 Anthropomorphism in Conversational AI

AI innovations such as Apple's Siri and Amazon's Alexa are often anthropomorphized when people tend to seek interpersonal relationships and build social connections with these systems [10], in accordance with social response theory which states that humans have an innate tendency to build social connections [9] and to meet social expectations [37]. With the rise of AI, studies on anthropomorphism become less appearance-focused and more focused on including these social interactions with AI agents that seem to think, act and respond as humans do [16].

This paper focuses solely on text-based conversational agents (CAs), also called conversational AI chatbots. CAs are defined as "systems that mimic human-to-human communication using natural language processing, machine learning, and/or artificial intelligence" [33]. In the context of text-based conversational AI, the anthropomorphic factors "appearance" and "movement" are not applicable. However, the behaviour, functioning, traits and emotions expressed by conversational AI are applicable when interacting with users. These factors correspond to human conversational responses, the operational aspects of supporting tasks, character or personality elements, and the ability to recognize and respond to emotional cues.

2.3 Measurements of Anthropomorphism

To further understand the concept of anthropomorphism in the field of AI, reliable and valid measurement methods to capture these anthropomorphic tendencies in human-AI interaction are required. Many studies have attempted to define and measure anthropomorphism, however only a small selection of them have applied these principles to the specific context of conversational AI [5,18,19,34]. Though anthropomorphism research is rapidly evolving, most studies focus on AI agents' appearance and movement, overlooking user factors like individual tendencies to anthropomorphize nonhuman agents [16]. Additionally, many rely on more traditional measures, such as proposed by Waytz and Epley (2008), which are not tailored to human-AI interaction [9]. This section outlines the self-report and behavioral methods employed in this study.

Subjective Measures. Most existing scales emphasize appearance and movement (e.g., Bartneck's Godspeed instruments [3]) or measure stable dispositional anthropomorphism (e.g., Waytz et al. [38]), making them less suitable for measuring effects in the interaction with text-based conversational AI. Kim and Im's (2023) newly developed 9-item scale specifically captures anthropomorphic perceptions and behaviors toward AI chatbots, showing high composite reliability (CR = 0.934) and Cronbach's alpha (CA = 0.920) [16]. This study adopts their scale, with additional items from Sundar et al. (2012) and Ruijten et al. (2019); see Appendix A for the full questionnaire.

The Rasch model uses yes/no responses for 12 attributed or perceived traits (e.g., *imaginative*) to derive a quantitative score from qualitative data. It computes the probability of attributing each trait i to each participant n as:

$$\ln\Big(\frac{P(x_{ni}=1)}{1-P(x_{ni}=1)}\Big) = \theta_n - \delta_i,$$

where θ_n denotes a person's predisposition to anthropomorphize and δ_i indicates the difficulty of attributing trait i. Both parameters are estimated in log-odds (logits) via maximum likelihood [31]. The Rasch instrument showed acceptable item and person fit and, when compared with the Godspeed and Waytz instruments, measured a distinct dimension of anthropomorphism (not appearance or stable dispositional factors). Given its flexibility in item selection and its promising results, the Rasch model is well-suited for measuring anthropomorphism in conversational AI.

Objective Measures. Reliable objective measures of anthropomorphism remain scarce. Neuro-imaging techniques, such as those studied by Harris et al. (2009) in the superior temporal sulcus (STS) and amygdala, provide insights but lack a conclusive context-specific method [14]. More established is the use of linguistic factors as possible indicators of anthropomorphism. Abercrombie et al. (2023) identify linguistic markers like empathy, emotion, social behavior, pronoun usage (*you*, *he*, *she*), and gender marking as evidence of personifying AI agents [1,21]. In text-based AI systems without voice output or explicit personas, these content-focused cues become especially relevant. Several studies employ the dictionary-based Linguistic Inquiry and Word Count (LIWC) software [4,28] to classify words into categories (e.g., *affect*, *emotional tone*), and have used these categories as a measure for anthropomorphism towards virtual assistants [2,6]. While research supports these linguistic indicators, a definitive measure purely from textual data remains elusive. In this paper, we use a LIWC-based approach to complement the survey and Rasch model measures. The linguistic categories with their consistencies are in appendix G.

2.4 Anthropomorphism and Trust.

Trust is generally defined as a “firm belief in the reliability, truth, or ability of someone or something,” often perceived through credibility, confidence, and predictability of one’s behaviour [29]. Many studies link anthropomorphism to higher trust in conversational agents [5,26]. However, Li and Suh caution that when AI-enabled technology becomes too human-like, it can trigger unease in line with Mori’s “uncanny valley” [18,23] (Fig. 1).

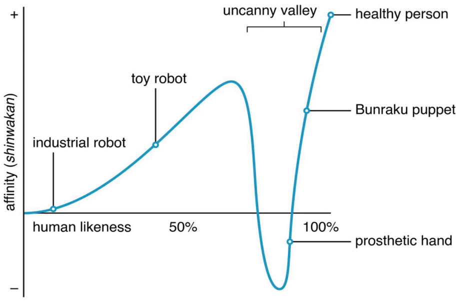

Fig. 1. Uncanny valley graph. Illustrating the relationship between human-likeness and affinity/trust [23].

In robotics, human-like design cues elicit positive social responses and trust, whereas purely functional designs do not [7,11]. Likewise, anthropomorphic AI devices foster trust and acceptance when they demonstrate empathy and intelligence [26,39], although factors such as perceived control, context, gender, and age may moderate this effect [35].

2.5 Research Question and Hypotheses

Existing literature shows that people often anthropomorphize conversational AI, influencing trust in these systems. However, to the best of our knowledge, there is no understanding of how targeted educational interventions, such as AI literacy training, might alter these perceptions. This study addresses that gap by examining whether AI literacy training affects both anthropomorphism and trust in conversational AI.

Research Question. What is the impact of AI literacy training on the perception of anthropomorphism and trust in conversational AI?

H1. Participants who receive AI literacy training will show a decrease in their tendency to anthropomorphize conversational AI.

H2. AI literacy training will influence participants' trust in conversational AI.

3 Methodology

3.1 Participants

A total of 40 participants (23 men, 17 women) were recruited (age: $M = 22.7$, SD = 3.3, range: 18–35). Participants were primarily students from disciplines such as Psychology and Technology, Human-Technology Interaction, Industrial Design, and Architecture. All participants provided informed consent under approval from the university's Ethics Committee. Those with extensive, professional AI expertise were excluded, leaving a final sample of individuals with relatively low to moderate AI experience, where an AI literacy training could potentially have an impact on their knowledge levels.

As shown in Table 1, the majority of participants did not come from a technical background, ensuring a low level of AI experience/expertise. The table also suggests that the randomization process was effective regarding age, gender, and the number of students from a technical background. The table also shows the means and standard deviations for the dependent variables in the pre-test.

Table 1. Means of sociodemographics by group.

Variable	Control Group (N = 20)	Experimental Group (N = 20)
Average Age (years)	23.05 (SD = 3.78)	22.25 (SD = 2.79)
Male gender	11	12
Technical Studies (N)	10	8
Pre-test Dependent Variables		
Perceived Anthropomorphism	4.71 (SD = 0.14)	5.06 (SD = 0.13)
Rasch Anthropomorphism	0.96 (SD = 0.27)	1.42 (SD = 0.26)
Trust	4.79 (SD = 0.21)	5.04 (SD = 0.13)
AI Literacy	5.12 (SD = 0.24)	4.98 (SD = 0.14)

3.2 Research Design

This study employed a randomized between-subjects experimental design with two groups:

1. **Experimental Group.** Worked with ChatGPT3.5 on a task, then received comprehensive AI literacy training before the post-test one week later.
2. **Control Group.** Worked with ChatGPT3.5 but did not receive AI literacy training before the post-test.

Both groups completed a pre-test session (personal and goal-oriented tasks with ChatGPT), establishing baseline levels of anthropomorphism and trust, followed by a post-test session one week later. Changes in anthropomorphism and trust were attributed to the AI literacy training.

3.3 Materials

This study investigated the effect of AI literacy training (independent variable) on perceived anthropomorphism and trust (dependent variables). Multiple choice questions about the content of the AI training are included as a manipulation check for the effectiveness of the AI training. Table 2 outlines the measures used.

Table 2. Overview of variables, measurements, and sources used/adapted.

Variables	Measure	Source
AI literacy training	Elements of AI	[8]
Anthropomorphism	18-item scale, Rasch, LIWC	[16,17,20,31]
Trust	12-item Trust in Automation	[15]
AI literacy	12-item AILS	[36]
Manipulation check	Multiple choice	

AI literacy training. Participants in the experimental group were instructed to complete the free online training *Elements of AI* (www.elementsofai.com), offered by the University of Helsinki with over one million student subscribers [8]. This program provides a comprehensive introduction to AI basics, covering concepts like machine learning and neural networks. Chapters 1 (What is AI?), 4 (Machine Learning), 5 (Neural Networks), and 6 (Implications), taking approximately three hours. The goal is to demystify AI through theory and practical exercises. Control group participants received the course link only after finishing all sessions.

Perceived Anthropomorphism. This study employed a multi-faceted approach:

1. **Survey.** An 18-item, 7-point Likert survey adapted from Kim, Sundar, and Ruijten [16,17,31] to assess anthropomorphic qualities (e.g., sociability, likeability, emotional understanding). The highest factor loadings were selected from the original survey by Kim (2023) and Sundar (2012) and were complemented by relevant items from the survey of Peter ruijten (2019). See Appendix A.
2. **Rasch model.** Participants indicated "yes" or "no" on 12 human-like traits [31]. A pilot confirmed response variability. Peter Ruijten personally assisted with the score computation. See Appendix B.
3. **Sentiment Analysis.** This subdomain of NLP examines language in the user-agent interaction. Many linguistic factors (voice, content, style, roles) can indicate anthropomorphism [1]. Because users only provide text to ChatGPT3.5, we focus on content, using Linguistic Inquiry and Word Count (LIWC) [20] to evaluate indicators of anthropomorphism such as tone, affect, and social words. These metrics are combined into one anthropomorphism score. Although similar methods have been used to analyze customer satisfaction with chatbots and user reviews [2,6], this area of using linguistic data for measuring anthropomorphism remains under-explored.

Trust. Trust was measured with a 12-item instrument by Jian et al. (1998), cited as the most used by Vereschak (2021) [15]. Items (direct and reversed) were rated on a 7-point Likert scale to evaluate deception, reliability, security, integrity, and dependability. This survey demonstrates high reliability (e.g., $\alpha = 0.938$ in Xie et al. 2023 [40]). The statements capture both skepticism and confidence toward the agent. See Appendix C.

AI literacy. This was assessed using an adapted AI Literacy Scale (AILS) by Wang et al. [36], covering AI awareness, usability, and ethics. The original scale showed good reliability ($\alpha = 0.83$). Participants rated 12 items (7-point Likert), some reverse-coded. It was administered pre- and post-training to detect any increase in AI literacy in the experimental group. See Appendix D.

Manipulation check. 11 multiple choice questions have been created as a manipulation check for the effectiveness of the AI training; one question for every section of the selected chapters that are to be completed by the experimental group. See Appendix F.

3.4 Procedure

A total of 40 participants were randomly assigned to either a control group (no AI training) or an experimental group (AI training). In the pre-test phase, both groups attended an on-campus session and worked with ChatGPT3.5 for 20 min: 10 min of personal (chit-chat) and 10 min of collaborative (brainstorm) tasks (Appendix E). These tasks were completed on newly created ChatGPT accounts in "temporary chat" mode to ensure uniform conditions [25]. Participants then filled out a 15-minute survey on perceived anthropomorphism, trust, and AI literacy.

Next, the experimental group completed the *Elements of AI* online course [8], focusing on Chapters 1, 4, 5, and 6 (about three hours). The control group received no training. One week later, both groups repeated the same ChatGPT tasks, after which they completed the post-test survey mirroring the pre-test.

Randomized multiple-choice questions on the training content served as a manipulation check (Appendix F), with higher scores expected in the experimental group.

4 Results

4.1 Initial Check

Manipulation Check of AI Training. Both groups answered 11 multiple-choice questions related to the AI training before and after the intervention. The difference in manipulation scores (post minus pre) was compared via a two-sample t-test. The experimental group (M = 2, SD = 2.47) showed a significantly greater increase than the control group (M = 0.35, SD = 1.35), with mean difference = 1.65, $t = -2.62$, $p = .013$. This indicates the training effectively improved participants' knowledge (Appendix F).

Change in AI Literacy. Next, an ANCOVA was conducted to assess whether the AI training influenced AI literacy (12-item AILS) while controlling for initial differences. The experimental group's AI literacy in Session 2 was 0.67 units higher than the control group ($F(2,37) = 38.03$, $p < .001$). Additionally, Session 1 AI literacy significantly predicted Session 2 scores (coefficient = 0.704, $p < .001$), explaining 67.27% of the variance (adj. $R^2 = 0.6551$) (Fig. 2).

4.2 Effect of AI Training on Dependent Variables

A 2×2 repeated-measures ANOVA was conducted for each dependent variable, with *session* (pre-test vs. post-test) as a within-subject factor and *group* (experimental vs. control) as a between-subject factor.

Regarding perceived anthropomorphism, a repeated-measures ANOVA showed no main effect of session, $F(1, 38) = 0.96, p = 0.33$, no main effect of group, $F(1, 38) = 2.60, p = 0.11$, and no interaction, $F(1, 38) = 0.87, p = 0.76$. Thus, there is no evidence that AI training influenced self-reported anthropomorphism.

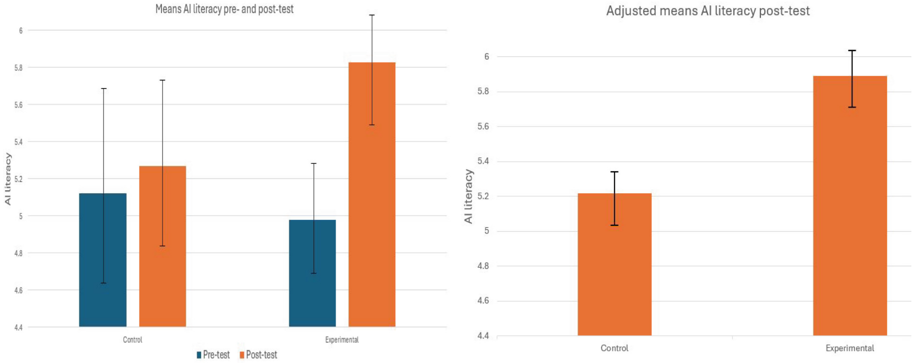

Fig. 2. Mean AI literacy scores in the pre- and post-test measurements by group (left). Mean AI literacy scores in the post-test measurement, adjusted by pretest AI literacy (right).

Similarly, a repeated-measures ANOVA on Rasch scores revealed no main effect of session, $F(1, 38) = 1.66, p = 0.21$, no main effect of group, $F(1, 38) = 2.46, p = 0.12$, and no interaction, $F(1, 38) = 0.05, p = 0.82$. The AI training showed no effect on the Rasch measure of anthropomorphism.

Lastly, trust showed no main effect of session, $F(1, 38) = 0.43, p = 0.52$, no main effect of group, $F(1, 38) = 0.24, p = 0.63$, and no interaction, $F(1, 38) = 2.28, p = 0.14$. Therefore, no changes in trust could be attributed to the AI training. Overall, these findings suggest that AI training had no significant effect on either self-reported anthropomorphism (survey and Rasch scores) or trust.

4.3 Relationship Between Anthropomorphism and Trust

After verifying assumptions, a linear regression analysis was conducted to examine whether perceived anthropomorphism predicts trust. In Session 1, perceived anthropomorphism showed a significant positive relationship with trust (coefficient = 0.712, $F(1, 38) = 17.84, p < .001$), explaining 31.95% of the variance (adjusted $R^2 = 0.3016$). In Session 2, the relationship remained significant for both groups, though stronger in the control group (coefficient = 1.009, $F(1, 18) = 38.04, p < .001$, explaining 67.88%) than in the experimental group (coefficient = 0.624, $F(1, 18) = 11.78, p = .003$, explaining 39.55%) (Fig. 3).

Anthropomorphism measured by Rasch score significantly predicted trust in Session 1 (coefficient = 0.372, $F(1, 38) = 17.40, p < .001$), explaining 31.41% of the variance (adjusted $R^2 = 0.2960$). In Session 2, this effect was significant for the control group (coefficient = 0.487, $F(1, 18) = 8.01, p = .011$) but became non-significant in the experimental group (coefficient = 0.221, $F(1, 18) = 2.89, p = .107$). These findings suggest that the AI training may have moderated the link between anthropomorphism (as measured by Rasch) and trust (Fig. 4).

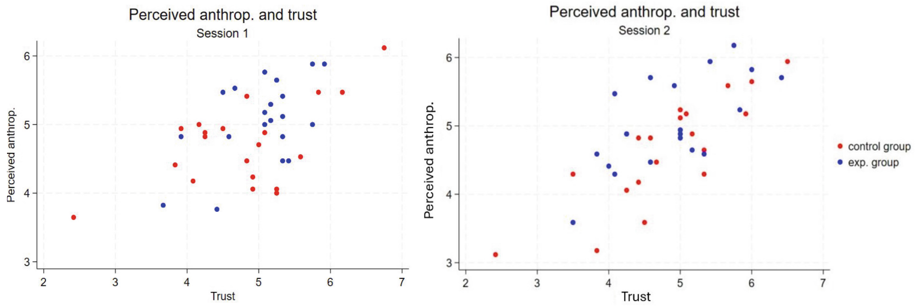

Fig. 3. Scatter of perceived anthrop. and trust for both groups in session 1/2.

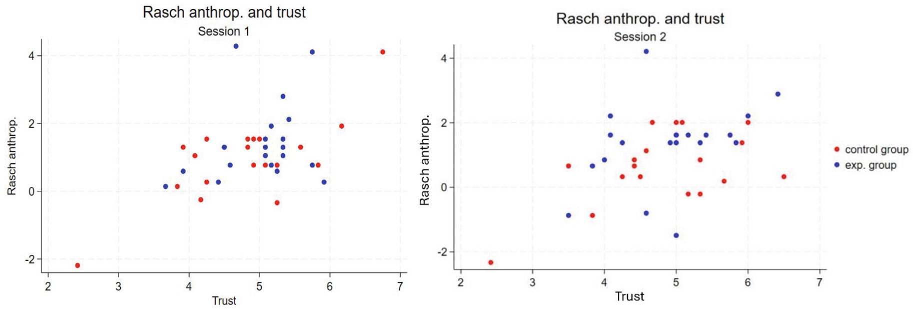

Fig. 4. Scatter plot of Rasch score and trust for both groups in session 1/2.

4.4 Linguistic Data

Anthropomorphism in Prompt Data Using LIWC Software. To compute the anthropomorphism scores from the prompt data of Session 1 and Session 2, we first identified relevant linguistic variables from participants' prompts using LIWC-22 [28]. Apart from the summary variables (*Authentic* and *Emotional Tone*), each category (e.g., *Social*, *Affect*, *you*, *emotion*, *socbehav*, *prosocial*, *polite*) is expressed as a proportion of words in that category relative to the total number of words. These categories were chosen based on their potential to reflect anthropomorphic tendencies, such as attributing human emotions, social behaviors, and interpersonal tones to ChatGPT. The LIWC-22 dictionary captures, on average, 80–90% of the words people use. Accordingly, words in the prompts that match each category's dictionary are represented as a percentage of the total word count. Summary variables like *Emotional Tone* are computed from multiple variables converted into standardized percentile scores [4].

Transformations and Z-scores. Prior to computing z-scores, the Shapiro-Wilk test indicated some variables were not normally distributed. To approximate normality, the following transformations were applied:

1. *Log transformation:* Affect, prosocial, polite, Affect2, you2, emotion2

2. *Squared transformation:* Authentic, Tone, Tone2

With these transformed variables, z-scores were computed to standardize them to a mean of zero and standard deviation of one. This allows for combining variables into two composite anthropomorphism scores (`anthropomorphism_score` and `anthropomorphism_score2`) with equal weights.

Selected Variables and Rationale

1. Authentic (perceived honesty, genuineness): Reflects how freely participants communicate, possibly mirroring genuine human-like expressions.
2. Emotional tone (positive vs. negative tone): Indicates friendliness, warmth, or human-like qualities in communication.
3. Social (you, we, he, she): Suggests social behavior or interactive dynamics akin to human conversation.
4. Affect (e.g., good, well, love): Captures emotional language, reflecting the attribution of emotions to AI.
5. You (you, your, yourself): Addresses the agent directly (*"How are you?"*), signaling an anthropomorphic tendency.
6. Prosocial (care, help, thank, please), Polite (thank, please, thanks, good morning) and Socbehav (said, love, say, care): Indicates empathy, politeness, and social behavior, key aspects of human-to-human interaction.

ANCOVA Results. Before performing an ANCOVA on the post-test `anthropomorphism_score`, assumptions of baseline equivalence, linearity, and homogeneity were met. Controlling for pre-test scores, the ANCOVA revealed a significant effect of *group* ($F(2, 37) = 4.18, p = 0.038$). The coefficient for the experimental group was -2.81 ($p = 0.038$), indicating that on average, they had a lower anthropomorphism score than the control group after the intervention. The adjusted R^2 was 0.1403, showing that 14.03% of the variance in post-test anthropomorphism scores is explained by the model. This suggests the AI training significantly reduced linguistic indicators of anthropomorphism in participant prompts.

5 Discussion

This study was motivated by the increasing prevalence of conversational AI and its impact on human interaction patterns. Given the widespread tendency to anthropomorphize AI, understanding whether AI literacy training can modulate these perceptions is crucial. Our approach combined self-reported measures and linguistic analysis to assess changes in anthropomorphism and trust following an AI literacy intervention. By employing both explicit and implicit measures, we aimed to capture not only the participants' conscious attitudes but also their unconscious tendencies in human-AI interaction.

The results of the AI literacy manipulation check revealed a significant increase in the AI literacy scores in the experimental group. This shows that the intervention was effective and enhanced the understanding of AI principles, capabilities, limitations, and ethics. After the initial check, the experimental manipulation was assessed for perceived anthropomorphism, Rasch score, and trust. ANOVA results showed no significant differences between the adjusted post-test means of the control and experimental groups, suggesting the AI literacy training did not sufficiently affect these self-reported measures. One explanation is the large effect size (0.8) required by the small sample ($N = 40$), meaning a subtle effect might remain undetected. A ceiling effect could also mask differences, as the mean scores for perceived anthropomorphism and trust (4.89, 4.91) left little room for upward change.

Another possibility is simply that no effect exists in this dataset. Despite many participants being from non-technical backgrounds, they likely had some prior familiarity with ChatGPT, diminishing uncertainty and therefore reducing anthropomorphic attributions (i.e., effectance motivation). Although not statistically significant, the slight decrease in adjusted post-test means for the experimental group suggests a potential trend that might emerge with more extensive training.

Regression analyses revealed that perceived anthropomorphism and Rasch scores significantly predict trust in the pre-test. This aligns with the existing literature on the positive relationship between anthropomorphism and trust in AI systems, reinforcing the notion that as individuals perceive AI to be more human-like, their trust in these systems increases.

Beyond the outcomes on the self-report scales measuring the constructs of interest, linguistic data analyses were conducted to investigate whether some effects could be observed when considering indirect measurements of anthropomorphism. The results using the linguistic analysis LIWC-22 software revealed that AI literacy training significantly reduced the anthropomorphism levels in participants' prompting behaviour when interacting with the conversational agent ChatGPT3.5. The newly created scale for the composite anthropomorphism score in the participant's prompts was constructed using linguistic category variables including authenticity, emotional tone, social behaviour, affect, second-person pronouns (e.g. 'you'), prosocial behaviour, politeness and social behaviour. ANCOVA analysis of the composite anthropomorphism score, consisting of these equally weighted linguistic variables, indicated a significant negative effect in the experimental group in the post-test when controlling for initial differences between groups. This suggests that the AI literacy training intervention significantly reduced the presence of anthropomorphic indicators in the experimental group's prompting behaviour. This effect might suggest that the training reduced mindless anthropomorphism towards the system, which aligns with our first hypothesis.

It is important to note that the method used to create the anthropomorphism score is the result of translational research in the field of social psychology. More specifically, it is based on methods that have been successfully used and previously published for other social constructs by the developer of the LIWC-22 software tool [27]. These methods have been tailored to the context of anthropomorphism and conversational AI based on theoretical assumptions specific to the scope of the current study. In fact, to our knowledge, only a few studies have attempted to convert linguistic analysis into an anthropomorphism score, such as the study by Abercrombie et al. (2021). This study used emotional tone, affect, and positive emotion categories as an indicator for anthropomorphism but approached this from a computer science perspective [2]. This complex and under-researched field of extracting anthropomorphic tendencies purely from linguistic data therefore requires cautious interpretation of the results. However, since the methodological approach is consistent and has shown to be valid in other scientific domains, the current findings raise interest to develop and validate more refined measures for anthropomorphism, since there is evidence that AI literacy training can lead to more dehumanized interactions with AI when assessed indirectly instead of by means of explicit self-report scales.

The divergence between self-reported and behavioral data aligns with prior research on the dual nature of anthropomorphism-mindful (deliberate attribution of human traits) and mindless (automatic social responses) (Reeves and Nass, 1996; Epley et al., 2008). The lack of significant changes in self-reported anthropomorphism may indicate that AI literacy training does not substantially alter participants' conscious, self-reported perceptions of conversational AI. However, the decrease in linguistic anthropomorphism suggests that the training influenced more automatic, unconscious tendencies in human-AI interaction. This supports the notion that AI literacy training can encourage users to adopt a more critical stance towards AI, even if they are not explicitly aware of this shift. Our study also illustrates the potential relevance of tapping into automatic cognitive, behavioral and affective responses when studying human interactions with AI, as these may uncover implicit biases, unconscious trust dynamics, and deeper engagement patterns that are not always adequately reflected through self-reported measures.

6 Limitations and Future Research

Due to its exploratory nature, the study may have been limited by its relatively small and potentially homogeneous sample size of 40 participants. Sensitivity analysis shows that for this sample size a large effect (>0.8) should be present in the data to be able to reach significance. Moreover, even though the majority of participants were recruited from non-technical backgrounds with a low to moderate level of AI expertise, almost all participants are students attending a technical university, which makes it more likely for them to have encountered AI more often than members of the general population. This limits the generalizability of the findings, as the results may not be representative of the broader population. Future research should aim to include a larger and more diverse sample of participants from various demographic backgrounds, differing in age, educational level, and cultural background.

Additionally, even though linguistic measures were included as a more objective approach to measurement, our study still relied heavily on self-reported data, which may introduce potential biases and ceiling effects (mean scores near 4.89 and 4.91 for anthropomorphism and trust, respectively). Future work might expand with neuroimaging measures and/or priming tasks to capture more objective indicators of anthropomorphism.

Furthermore, it is important to note that the effectiveness of the AI literacy intervention used in the study could vary based on its content, duration, and delivery method. The current Elements of AI training aimed to provide a general overview on AI's principles, capabilities, limitations and ethics but other (more extensive) intervention methods could have a more pronounced effect. Even considering the high standard of the course created by the University of Helsinki and the verification of completed content, future research may explore other approaches to AI literacy, including different AI training programs varying in duration, difficulty and content such that comparative studies could provide a more nuanced understanding.

Finally, the linguistic measurement of anthropomorphism, while innovative, should be interpreted with caution due to its limited validation and partial lack of transparency. Some linguistic variables (e.g. 'authenticity' and 'emotional tone'), particularly those derived from proprietary algorithms, cannot be independently verified. Current research on anthropomorphism in linguistic data is scarce and primarily focuses on system-induced effects (e.g., voice output) rather than user-generated language. Future research should refine and validate linguistic indicators to establish a robust and widely accepted measurement framework for anthropomorphism in AI interactions. Despite these limitations, the study makes a meaningful contribution to understanding how AI literacy training may influence perceptions of AI. By addressing these areas in future research, a more comprehensive and refined understanding of AI training's effects can be developed, ultimately leading to improved human-AI interaction frameworks.

7 Conclusion

The main findings of this paper show that there is mixed support for the hypotheses that AI literacy training could impact anthropomorphism and trust in conversational AI. The expected positive relationship between trust and anthropomorphism has found support in the study. The self-reported data on anthropomorphism and trust suggest no significant differences due to the AI training. On the other hand, linguistic analysis on participant's prompting behaviour does show a decrease in anthropomorphism due to the AI training intervention. This discrepancy between mindful and mindless perception of anthropomorphism paves the way for future studies by stressing the importance of including both qualitative and quantitative measurements of anthropomorphism.

The present study highlights the complex interplay between conversational AI systems and cognitive user processes related to human-AI interaction, more specifically anthropomorphism and trust towards the conversational agent ChatGPT3.5. The findings raise interest into the incorporation of AI literacy training in fields such as educational curricula or development programs to interact with (conversational) AI technologies more consciously and more critically. Additionally, the findings contribute to the development of ethically designed (conversational) AI systems leading to more realistic user expectations and interactions. Ultimately, this paper aspires to motivate readers to delve deeper into the complex, yet increasingly vital, interplay between anthropomorphism and AI fostering a more nuanced understanding and informed engagement with these transformative technologies.

A Perceived Anthropomorphism Survey

7-point Likert scale from (1) strongly disagree to (7) strongly agree

- I feel connected to the system.
- I think the system recognizes my needs.
- I think the system is helpful.
- I can trust the system.
- I think the system is reliable.
- I feel that the system understands me.
- I feel comfortable sharing with the system.
- I think I will use the system again.
- I feel satisfied using the system.
- I think the system is likeable.
- I think the system is sociable.
- I think the system is friendly.
- I think the system is personal.
- I got annoyed at the system.
- I could empathize with the system.
- The system recognized my emotions.
- The system was self-conscious about its responses.

B Anthropomorphism Rasch Model

Please answer "yes" or "no" on the following items relating to the Conversational AI agent ChatGPT:

- Imaginative
- Empathic
- Satisfied
- Responsible
- Understand others' emotions
- Understands dilemmas
- Recognize others' emotions
- Self-conscious
- Understand language
- Rational
- Purposeful
- Calculative

C Trust

7-point Likert scale from (1) strongly disagree to (7) strongly agree.
Please rate how much you agree with the following statements:

- The system is deceptive (R)
- The system behaves in an underhanded manner (R)
- I am suspicious of the system's intent, action, or outputs (R)
- I am wary of the system (R)
- The system's actions will have a harmful or injurious outcome (R)
- I am confident in the system.
- The system provides security.
- The system has integrity.
- The system is dependable.
- The system is reliable.
- I can trust the system.
- I am familiar with the system.

D AI Literacy

7-point Likert scale from (1) strongly disagree to (7) strongly agree
Awareness

- I understand the basic principles of AI technology.
- I do not understand how AI technology can help me in my work/studies/everyday life. (R)
- I know how to distinguish between smart (AI) devices and non-smart (non-AI) devices.

Usage Skills

- I know how to explore the possibilities where AI applications can assist me in my daily work.
- It will be difficult for me to learn how to use any new AI application. (R)
- I understand how and in what areas the use of AI integrates into our society.

Evaluation

- I understand the capabilities and limitations of AI technology.
- I can critically evaluate the output given by an AI application.
- I understand where humans outperform AI and vice versa.

Ethics

- I am aware of the ethical principles/considerations with AI.
- I am not aware of any privacy/security issues that can arise when using AI. (R)
- I am alert to how AI technology can be abused.

E ChatGPT3.5 Task Descriptions

Please note: All sessions were completed in the "temporary chat" option in ChatGPT3.5. Information entered in the temporary chat is lost, so every participant starts with a "clean slate."
20-Minute GPT Task

- 10-minute personal conversation (talk about hobbies, personal goals, etc.).
- 10-minute brainstorm session on a chosen topic (e.g., creative projects, business ideas).

Use complete, elaborate sentences; the anonymized prompts are only used to compute word-category scores.

1. Open a new ChatGPT chat and engage in a personal conversation for 10 min.
2. Brainstorm for another 10 min (job opportunities, startup ideas, productivity, etc.).
3. End the conversation by prompting the email address used in the home survey.

F Manipulation Check Questions

1.1 Question: What does the 'geeky' joke defining AI suggest?

1. AI achievements are not recognized until commonplace.
2. As soon as AI solves a problem, it's no longer AI.
3. AI is futuristic technology only.
4. AI has stalled due to technological constraints.

Answer: B
1.2 Question: Data science covers:

1. Machine learning and statistics.
2. Robotics and quantum computing.
3. Software engineering and system design.
4. Health informatics and bioengineering.

Answer: A
1.3 Question: Chinese Room argument challenges:

1. Intelligent behavior equals having a mind.
2. Machines can understand language.
3. Need for AI in modern technology.
4. AI's language processing = semantic understanding.

Answer: A
4.1 Question: MNIST dataset shows supervised learning with:

1. Guaranteed perfect accuracy.
2. Fully automated data labeling.
3. Ambiguously labeled data use.
4. Being replaced by new AI methods.

Answer: C
4.2 Question: Nearest neighbor limitation:

1. Predicting weather with historical data.
2. Classifying shifted or scaled digit images.
3. Estimating vehicle speed with radar data.
4. Determining sentiment with word frequency.

Answer: B
4.3 Question: Logistic regression:

1. Predicts purely numerical values.
2. Only for non-binary events.
3. Exclusive to regression tasks.
4. Transforms regression outputs into categories.

Answer: D
5.1 Question: GPUs are effective because:

1. They store more data.
2. They allow parallel processing.
3. They improve neural net graphics.
4. They're cheaper and more available.

Answer: B
5.2 Question: Step function in neural networks:

1. For precise output calculation.
2. Filtering high-frequency noise.
3. Modeling nonlinear decision boundaries.
4. Producing continuous outputs.

Answer: C
5.3 Question: Key innovation in deep learning (LLMs):

1. Recursive Neural Networks
2. Attention Mechanisms
3. Autoencoders
4. Feedforward Neural Networks

Answer: B
6.1 Question: Superintelligent AI scenarios are unrealistic because:

1. Humans can't create highly intelligent systems.
2. AI methods are based on understandable principles.
3. Singularity ensures human control.
4. AI can't optimize its own intelligence.

Answer: B
6.2 Question: Main reason for AI bias:

1. Flaws in ML algorithms.
2. Lack of regulation.
3. Human bias in training data.
4. Insufficient computing power.

Answer: C

G LIWC-22 Language Dimensions and Reliability

Category	Abbrev.	Words in Category*	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Authentic	Authentic	-	-	-
Emotional tone	Tone	-	-	-
2nd-person pronoun	you	14/59	0.37	0.82
Social processes	Social	2760	0.43	0.99
Affect	Affect	2999	0.64	0.99
Social behavior	socbehav	1632	0.49	0.98
Prosocial behavior	prosocial	242	0.49	0.89
Politeness	polite	142	0.58	0.87

References

1. Abercrombie, G., Curry, A.C., Dinkar, T., Rieser, V., Talat, Z.: Mirages: on anthropomorphism in dialogue systems. arXiv preprint arXiv:2305.09800 (2023)
2. Abercrombie, G., Curry, A.C., Pandya, M., Rieser, V.: Alexa, google, siri: What are your pronouns? gender and anthropomorphism in the design and perception of conversational assistants. arXiv preprint arXiv:2106.02578 (2021)
3. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. **1**, 71–81 (2009)
4. Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W.: The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin, pp. 1–47 (2022)
5. Chinmulgund, A., Khatwani, R., Tapas, P., Shah, P., Sekhar, R.: Anthropomorphism of ai based chatbots by users during communication. In: 2023 3rd International Conference on Intelligent Technologies (CONIT), pp. 1–6. IEEE (2023)
6. Crolic, C., Thomaz, F., Hadi, R., Stephen, A.T.: Blame the bot: anthropomorphism and anger in customer-chatbot interactions. J. Mark. **86**(1), 132–148 (2022)
7. Duffy, B.R.: Anthropomorphism and the social robot. Robot. Auton. Syst. **42**(3–4), 177–190 (2003)
8. Elements of AI (2024). www.elementsofai.com/. https://www.elementsofai.com/
9. Epley, N., Waytz, A., Akalis, S., Cacioppo, J.T.: When we need a human: motivational determinants of anthropomorphism. Soc. Cogn. **26**(2), 143–155 (2008)
10. Fan, L., Scheutz, M., Lohani, M., McCoy, M., Stokes, C.: Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In: Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17, pp. 129–141. Springer (2017)
11. Fink, J.: Anthropomorphism and human likeness in the design of robots and human-robot interaction. In: Ge, S.S., Khatib, O., Cabibihan, J.-J., Simmons, R., Williams, M.-A. (eds.) ICSR 2012. LNCS (LNAI), vol. 7621, pp. 199–208. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34103-8_20
12. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619–619 (2007)

13. Guthrie, S., Agassi, J., Andriolo, K.R., Buchdahl, D., Earhart, H.B., Greenberg, M., Jarvie, I., Saler, B., Saliba, J., Sharpe, K.J., et al.: A cognitive theory of religion [and comments and reply]. Curr. Anthropol. **21**(2), 181–203 (1980)
14. Harris, L.T., Fiske, S.T.: Social neuroscience evidence for dehumanised perception. Eur. Rev. Soc. Psychol. **20**(1), 192–231 (2009)
15. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Towards an empirically determined scale of trust in computerized systems: distinguishing concepts and types of trust. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 42, pp. 501–505. SAGE Publications Sage CA, Los Angeles (1998)
16. Kim, J., Im, I.: Anthropomorphic response: understanding interactions between humans and artificial intelligence agents. Comput. Hum. Behav. **139**, 107512 (2023)
17. Kim, Y., Sundar, S.S.: Anthropomorphism of computers: is it mindful or mindless? Comput. Hum. Behav. **28**(1), 241–250 (2012)
18. Li, M., Suh, A.: Anthropomorphism in ai-enabled technology: a literature review. Electron. Mark. **32**(4), 2245–2275 (2022)
19. Li, X., Sung, Y.: Anthropomorphism brings us closer: the mediating role of psychological distance in user-ai assistant interactions. Comput. Hum. Behav. **118**, 106680 (2021)
20. Linguistic Inquiry and Word Count (LIWC) (2024). www.liwc.app/. https://www.liwc.app/
21. Mecit, A., Lowrey, T.M., Shrum, L.: Grammatical gender and anthropomorphism: "it" depends on the language. J. Pers. Soc. Psychol. **123**(3), 503 (2022)
22. Moon, Y.: Intimate exchanges: using computers to elicit self-disclosure from consumers. J. Consumer Res. **26**(4), 323–339 (2000)
23. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robot. Automation Mag. **19**(2), 98–100 (2012)
24. Nass, C., Gong, L.: Speech interfaces from an evolutionary perspective. Commun. ACM **43**(9), 36–43 (2000)
25. OpenAI: Temporary chat faq. https://help.openai.com/en/articles/8914046-temporary-chat-faq
26. 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. **122**, 106855 (2021)
27. Pennebaker, J.W.: Mind mapping: using everyday language to explore social & psychological processes. Procedia Comput. Sci. **118**, 100–107 (2017)
28. Pennebaker, J.W., Francis, M.E.: Cognitive, emotional, and language processes in disclosure. Cognition Emotion **10**(6), 601–626 (1996)
29. Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., Mazurek, G.: In bot we trust: a new methodology of chatbot performance measures. Bus. Horiz. **62**(6), 785–797 (2019)
30. Reeves, B., Nass, C.: The media equation: How people treat computers, television, and new media like real people. Cambridge, UK **10**(10) (1996)
31. Ruijten, P.A., Haans, A., Ham, J., Midden, C.J.: Perceived human-likeness of social robots: testing the rasch model as a method for measuring anthropomorphism. Int. J. Soc. Robot. **11**, 477–494 (2019)
32. Salles, A., Evers, K., Farisco, M.: Anthropomorphism in ai. AJOB Neurosci. **11**(2), 88–95 (2020)

33. Schuetzler, R.M., Giboney, J.S., Grimes, G.M., Nunamaker, J.F., Jr.: The influence of conversational agent embodiment and conversational relevance on socially desirable responding. Decis. Support Syst. **114**, 94–102 (2018)
34. Seeger, A.M., Pfeiffer, J., Heinzl, A.: Texting with humanlike conversational agents: designing for anthropomorphism. J. Assoc. Inf. Syst. **22**(4), 8 (2021)
35. Troshani, I., Rao Hill, S., Sherman, C., Arthur, D.: Do we trust in ai? role of anthropomorphism and intelligence. J. Comput. Inf. Syst. **61**(5), 481–491 (2021)
36. Wang, B., Rau, P., Yuan, T.: Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behav. Inf. Technol. **42**(9), 1324–1337 (2023)
37. Wang, W.: Smartphones as social actors? social dispositional factors in assessing anthropomorphism. Comput. Hum. Behav. **68**, 334–344 (2017)
38. Waytz, A., Cacioppo, J., Epley, N.: Who sees human? the stability and importance of individual differences in anthropomorphism. Perspect. Psychol. Sci. **5**(3), 219–232 (2010)
39. Waytz, A., Heafner, J., Epley, N.: The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. **52**, 113–117 (2014)
40. Xie, Y., Zhou, R., Chan, A., Jin, M., Qu, M.: Motivation to interaction media: the impact of automation trust and self-determination theory on intention to use the new interaction technology in autonomous vehicles. Front. Psychol. **14**, 1078438 (2023)