Artificial Intelligence Agents as our Teammates? How Accurate Mental Models can Facilitate Trust in Technology

Bachelor Thesis

Leona Weise

s2332736

Faculty of Behavioural, Management and Social Sciences, University of Twente

Examination Committee:

PhDc Esther Kox

Dr. Margot Kuttschreuter

June 23, 2022

Abstract

In recent years, Artificial Intelligence (AI) has evolved to be a nearly ubiquitous concept of the modern world, as tool and even as team partner. For effective collaboration in a Human-Agent Team (HAT), means to facilitate and maintain trust in the robotic colleague must be found. This paper investigated the concept of mental models, cognitive representations of systems such as AI agents, as a predictor of trust in AI as well as effects of AI familiarity and anthropomorphic agent embodiment. It was expected that mental model congruence with the assigned AI agent impacts initial trust, with further influence of AI familiarity. In addition, anthropomorphic embodiment was expected to yield higher initial trust. A 4 (anthropomorphic, zoomorphic, mechanical, virtual) x 2 (initial trust and after a violation) online experiment including a written scenario and mental model sketch was used. Majority of the 80 participants were students at the University of Twente. No support for the mentioned effects were found. However, this study makes a promising advance into the assessment of mental models.

Keywords: Artificial Intelligence, Mental Model, Trust, Human-Agent-Team, Anthropomorphism

Artificial Intelligence (AI) has a growing presence in the modern world, be it in a work, administrative or private setting. Its application ranges from virtual agents such as Apple's voice-controlled assistant Siri via Tesla's self-driving cars to agents in health care able to diagnose and treat patients (LaRosa & Danks, 2018). Next to a rapid increase in AI as a tool following the automatization of tasks, its role is becoming more social, as AI are evolving into team partners (Phillips et al., 2011). This concept is known as Human-Agent Teaming (HAT), defined as a team consisting of both human and AI members. It functions best when the human colleague has an accurate understanding of both the agent's abilities and limitations and accurate trust in the agent (LaRosa & Danks, 2018).

For the purpose of this study, AI is defined as machines simulating human intelligence processes, such as learning, reasoning, and self-correction (Gillath et al., 2021). Furthermore, this paper will use the term 'AI agent' and 'AI' to refer to AI systems in general, while 'AI agent' and 'robot' are used to refer to embodied agents.

Although humans today are exposed to AI more than ever before, experience and knowledge are often lacking, leading to little or inadequate trust in AI (Bansal, 2019; Kim & Song, 2020). Trust can be defined as the willingness to rely on another's intentions and behaviour based on positive expectations and despite possible risk (McKnight & Chervany, 2001). Trust is imperative for HAT, as it has been identified as the primary motive for individuals to accept new technology (Kim & Song, 2020; Li, Hess, & Valacich, 2008). Accurate understanding of an agent is another major facilitator for successful HAT. Representative of these expectations of new systems are mental models, cognitive representations of our external reality, which help us navigate encounters with new situations, for example with new technologies (Jones et al., 2011). However, they are often inaccurate, especially when it comes to the concept of AI (Jones et al., 2011, Ososky et al., 2013). Accordingly, strategies need to be found to assess and manage mental models to adjust expectations and trust in AI to facilitate Human-Agent Teaming.

Mental models

Mental models, cognitive representations of our external reality, help us navigate encounters with new systems, such as AI agents (Jones et al., 2011). Their definition has seen an evolution from "small-scale model of the world" to "reasoning mechanism" (Craik, 1943; Johnson-Laird, 1983). They help individuals explain and understand novel concepts, by tapping into knowledge about existing structures and applying it to another system or domain perceived to be similar (Gertner & Collins, 1987; Rickheit & Sichelschmidt, 1999). Previous research has identified the construction of mental models to be a type of analogical thinking, (Gertner & Gertner, 1983; Gertner & Collins, 1987). Thus, mental models can be understood as inferential frameworks that constantly evolve and adapt when exposed to new information. However, mental models are made up from subjective experience and are context-dependent, meaning that they change constantly and are highly dynamical (Jones et al., 2011). Especially when it comes to AI, a relatively new concept, individuals' mental models tend to be inaccurate and unstable (Phillips et al., 2019). Nevertheless, mental models impact how humans approach and interact with AI (Phillips et al., 2019). Previous research found that mental models of AI largely draw on superficial characteristics (Sims et al., 2005). For instance, Sims et al. (2005) found that individuals made assumptions of an AI agent's level of intelligence or aggressiveness based on visual aspects such as body position or presence of arms, country of origin or dialogue and language. One explanation for this is that individuals with a lack of experience with robots refer to what they know about similar entities, such as humans or animals (Sims et al., 2005). The quality of AI mental models is influenced by familiarity with AI in general. Research showed that the more experienced a user is with a technology, the more nuanced and abstract their mental models of AI are (DiSessa, Greeno, & Larkin, 1983). Lay understanding of AI agents is most often too concrete to be representative of such a complex system. This shows that more adequate expectations change the mental models we have, which could improve human colleagues' experience with their robotic

teammate.

Trust

Calibration of trust is very important for effective teamwork (Okamura & Yamada, 2020). Trust calibration refers to users adjusting their level of trust to the reliability of the system they are working with (Okamura & Yamada, 2020). Over-trust can have catastrophic consequences, for example in the case of an AI driving assistant, where overestimation of the agent's capacities could lead to a car crash. Under-trust may prevent the technology from being used in the first place, which may lead to economic loss or have detrimental consequences in high-stakes environments such as a hospital (Okamura & Yamada, 2018). Having accurate trust helps the user know when to accept or when to overrule an agent's feedback (Bansal et al., 2019). Therefore, trust is a crucial factor for successful HAT.

Anthropomorphism

An important predictor of trust is the level of perceived anthropomorphism of an agent. Anthropomorphism refers to the human tendency to attribute agent's behaviour to motivations, intentions, and characteristics similar to their own (Epley, Waytz, & Cacioppo, 2007). Humanoid features, such as the voice, facial features, figure and human-like movement can lead a user to falsely assume that a machine has a sense of self. This was confirmed by the Computers-Are-Like-Social-Actors (CASA) paradigm, according to which there are many similarities in the way individuals interact with agents and other humans (De Visser, 2016; Kim & Song, 2020). It proposes the media equation hypothesis according to which humans tend to treat AI agents like other humans, assigning human characteristics to them. According to De Visser (2016) the level of anthropomorphism may implicate the trust estimate placed on an AI agent. Research showed that users had lower initial expectations for anthropomorphic agents, consequentially leading to lower initial trust (De Visser et al.). Users perceived the humanoid agent as more fallible, as it was more similar to them. In contrast, agents with no anthropomorphic features were perceived as more useful and trustworthy. However, in the

case of trust violation, anthropomorphism positively affected trust resilience, meaning a smaller breakdown in trust after a violation. This proved that incorporating anthropomorphic features contributes to better calibrated trust, with less grave consequences in the case of a trust violation.

Current study

It has been established that due to the novelty of the concept, humans' mental models of AI are still often inaccurate, in addition to their trust being uncalibrated. However, knowledge is limited on how the two relate to one another and how trust can be facilitated by agent embodiment. This study aims to explore how mental models can influence humans' trust in an agent and facilitate HAT. However, due to the complex nature of mental models, this study will focus on superficial aspects of an agent. For this, four different types of agent embodiments, namely anthropomorphic, zoomorphic, mechanical and virtual were chosen. Accordingly, this study focusses on the effects of mental models and perceived anthropomorphism on initial trust, meaning the individual's first exposure to the agent. In addition, trust after a violation is assessed for a secondary exploratory purpose. Because it is impossible to measure mental models completely, due to their subjective and intrinsic nature, previous research has utilized visualizations to access mental models (Jones et a., 2011). Therefore, this study will use a mental model sketch.

Based on previous research it is expected that a match between preconceived expectations measured using a mental model sketch and the assigned AI agent increases initial trust in the agent.

H1: *Incongruence between the prior mental model and the assigned AI agent leads to lower levels of initial trust in comparison to congruence between prior mental models and assigned AI.*

Furthermore, as previous research asserts that individuals who are more familiar with AI have more abstract mental models, setting them up to be more open towards different agent embodiments, it is expected that a mismatch of expected and assigned AI agent is less detrimental for initial trust.

H2: The effect of incongruence between prior mental models and assigned AI agent on initial trust is decreased by the moderating effect of high levels of reported familiarity with AI.

Lastly, of the four different embodiments, the anthropomorphic is expected to elicit the highest trust due to findings that showed anthropomorphism to influence trust positively.

H3: The anthropomorphic agent condition elicits a higher level of initial trust compared to the zoomorphic, mechanical, and virtual conditions.

Methods

This study was part of a larger data collection effort assessing mental models of AI, which included several variables that were not of interest for the purpose of this study and were therefore excluded. However, for further exploration and matter of completeness, all three parts of the experiment scenario were kept as part of the design. The author's contribution to this research was the addition of AI familiarity to the theoretical framework of the study. Moreover, the author was responsible for discourse with participants and handling of data collection via SONA systems.

Design

A 4 (conditions: anthropomorphic, zoomorphic, mechanical, virtual) x 2 (time: initial trust before trust violation, trust after the trust violation) mixed design was employed. The independent between-subjects variable was AI agent embodiment, which participants were randomly assigned to. One of the dependent within-subject variables was trust. This was measured after introduction to the agent to assess the effects of agent embodiment on initial trust. For exploratory purposes, trust was measured a second time after part three of the

scenario, in which a trust violation occurred. Furthermore, dependent variables were mental model congruence, familiarity with Artificial Intelligence (AI) and perceived anthropomorphism, which were measured at the end of the study. See Figure 1 for an overview of the theoretical framework of the study, and Figure 2 for a more detailed overview of the course of the experiment.

Figure 1

Theoretical Framework

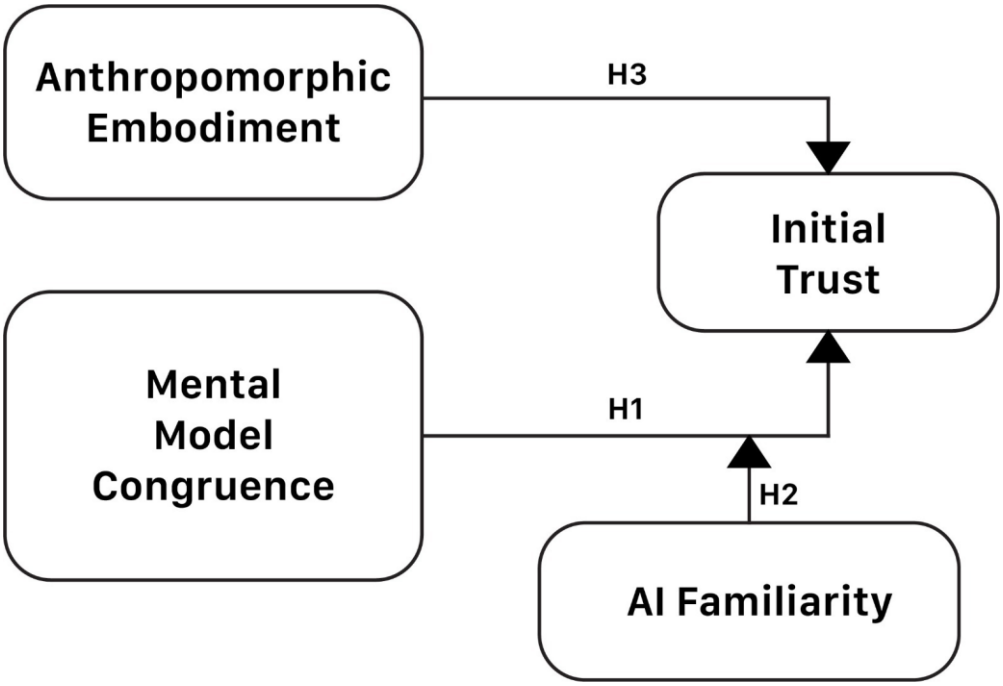
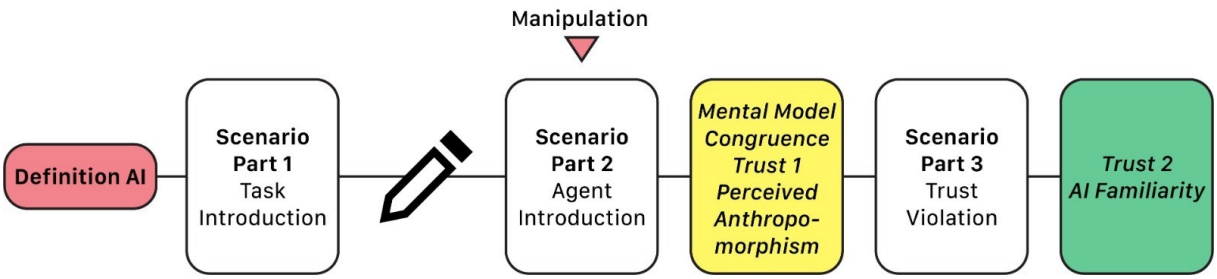


Figure 2

Course of the experiment



Note. Variables measured by means of survey are formatted in italics.

Participants

The BMS test subject pool SONA and convenience sampling were used to recruit participants. In total, 109 responses were recorded, of which two were test runs and 27 unfinished. These participants were omitted from the data set. A sample of 80 participants remained. A majority were students from the University of Twente, recruited via SONA systems, a test subject network from the University of Twente. Participants were compensated with credits. Their ages ranged from 18 to 65 (M = 24.7, SD = 8.6). Moreover, 45% of participants were male, 50% female, 2.5% referred to their gender as 'other' and 2.5% chose not to disclose their gender. Of all participants, 83.3% were German, 12.5% Dutch and 3.9% of another nationality. Regarding their highest level of education, 21.3% participants indicated it was their secondary school diploma, 51.3% college, 20% undergraduate degree, 7.5% graduate degree. Participants were automatically randomly assigned to one of the four conditions. 21 were assigned to the anthropomorphic, 20 to the zoomorphic, 18 to the mechanical and 21 to the virtual condition. No significant differences between groups according to age, gender, nationality, or education were found. Participant data was managed according to the ethical standards of the American Psychological Association (APA) and the study was approved by the ethics board of the BMS faculty.

Materials

The study was designed and administered online via Qualtrics. Participants were able to access it through a link on the SONA website using a laptop or their smartphone. The study was conducted in English, which all participants were proficient in.

Demographics

Before the start of the study, participants were asked to fill in their demographic data. Age, gender, nationality, educational level and, if applicable, field of study were recorded to check for group differences.

Mental Model Congruence

Mental Model Sketch

Before the sketch task, participants received a definition of AI (Appendix A) to facilitate clear and equal understanding of the concept for each participant. In addition, the agent was introduced as teammate and the task and environment were described (Appendix C). Participants were asked to draw a simple sketch of what they expected the agent to look like. The task provided a blank space and digital drawing tools. It was emphasized that the quality of the sketch was not of importance (Appendix B). The purpose of this task was to elicit visual representation of participants' mental models about the agent. This technique was adapted from research by Ososky et al. (2013) which used a mental model sketch to assess participants' expectations of what a military robotic teammate might look like in the future. Similarly, a study by Broadbent et al. (2011) used it to assess expectations about the appearance of robots in healthcare and found the drawing to be predictive of participants' affect towards the robot. The concept of this task is supported by mental model theory, referenced by Ososky et al. (2013), according to which mental models enable individuals to describe the form of a system.

Congruence Measure

The congruence of the sketch with the assigned AI agent was measured using a self-report measure on a 5-point Likert scale (*none at all; a great deal*) with a higher score indicating a more accurate match. The item was 'To what extent does this AI agent match the drawing that you made earlier?'.

Trust

The 14-item short version of the Trust-Perception Scale-HRI, which is specific to Human-Robot-Interactions (HRI) was used to measure trust (Schaefer, 2016). It included a rating scale from 0% to 100% in increments of 10%. It instructed participants to rate to what percentage the agent would display a certain characteristic. Examples included 'What % of

the time will the agent be dependent?' or unresponsive the agent would be or whether it would provide feedback. The was very good (α = .89), with acceptable validity (*KMO* = .81, *p* < .001). See Appendix D, Table D1 and D2 for a more detailed overview.

Perceived Anthropomorphism

Perceived Anthropomorphism was measured by a 5-item subscale of the Godspeed scale, along with perceived likeability and intelligence (Bartneck, 2009). The scale used the semantic differential (SD) measurement technique which refers to type of rating scale designed to find connotative meaning of objects, words, and concepts (Tsukada & Niitsuma, 2016). Participants were presented with word pairs (i.e., natural – fake, humanlike - machinelike) and instructed to rate the agent on 5-point Likert scales (see Appendix E). For four of the items, a higher score meant a higher score of perceived anthropomorphism, with the exception of one item, which was reverse formulated. Principal component analysis (PCA) revealed initial reliability to be good (α = .60). However, the reversed item showed an extraction value inferior to .4. After its exclusion the scale displayed higher reliability (α = .73). It showed acceptable sampling adequacy and a significant Bartlett's sphericity measure, deeming it sufficiently valid (*KMO* = .69, *p* < .001). See Appendix E, Table E1 and E2 for a more detailed overview.

AI Familiarity

AI Familiarity was measured on a 5-point Likert scale (strongly disagree - strongly agree) using 12 self-constructed items. A higher score indicated more experience with AI. Items ranged in concreteness from 'I have interacted with AI agents before' to 'I can explain the difference between machine learning algorithms and deep learning algorithms'. A principal component analysis (PCA) revealed great sampling adequacy as well as a significant measure of sphericity (KMO = .77, p < .001). All items showed extraction values above .5. In addition, reliability was assessed using Cronbach's Alpha (α = .83). It can be concluded that the scale is sufficiently valid and reliable. See Appendix F, Table F1 and F2 for a more

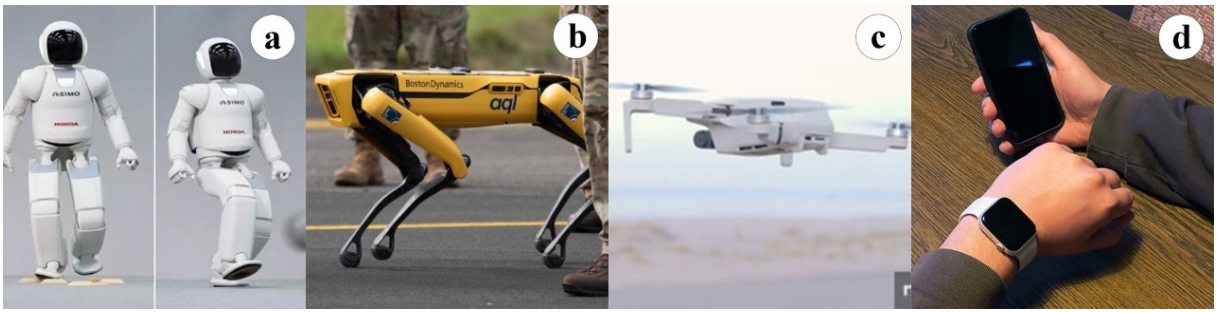
detailed overview.

Procedure

The questionnaire was administered online via Qualtrics. First, participants were presented with information on the research and a consent form. Upon giving their consent, participants were asked to answer demographical questions. After that they were presented with a definition of AI (Appendix A). Then followed the first part of the scenario during which participants were briefed on their task as a humanitarian volunteer aid for The Red Cross (Appendix C). According to the description, they were being sent out to help victims affected by natural disaster with the medical help of an artificial intelligence agent. This agent was going to provide the participant with information necessary to adequately help injured people, making the participant, who had no medical knowledge dependent on the agent's expertise. With a mental model sketch the participant was asked to imagine and draw what AI looks like to them. In the second part of the scenario, each participant was randomly assigned to one of the four conditions, anthropomorphic, zoomorphic, mechanical, or virtual. They were introduced to the agent and informed of its capabilities, cameras, and sensors to observe the environment and ability to understand and respond to the participant (Appendix C). To make the description more vivid, the participant was shown a GIF or a picture of the agent in motion (Figure 3). Next, several questionnaires assessed agent perception, initial trust in the agent, and perceived anthropomorphism. During the third part of the scenario the participant and the AI arrived at the disaster site and found three injured individuals (Appendix C). The agent scanned each of them and decided who to help first and gave advice on what to do. However, the agent made the wrong decision and one of the individuals suffered complications due to its failure. This was followed by the second trust measure. Afterwards, a violation check was performed. A final questionnaire assessed participants' level of familiarity with Artificial Intelligence.

Figure 3

Visual representations of the four different agent embodiments



Note. The four agent embodiments are depicted as follows: a) anthropomorphic, b) zoomorphic, c) mechanical, d) virtual embodiment. During the experiment, the first three images were shown as GIFs in motion and the fourth as an image.

Data Analysis

Prior to analysis, the data set was screened for missing data. Data analysis was conducted with the Statistical Package for Social Sciences (SPSS 25). Preliminary analyses were conducted to assess linear assumptions including normality, linearity, homoscedasticity and the absence of multicollinearity. Group differences in perceived anthropomorphism across conditions were examined using Planned Comparison Analysis. To test the first hypothesis, linear regression analysis was used to investigate the effect of mental model congruence on initial trust. Next, the second hypothesis was examined using a multiple linear model to assess moderating effects of familiarity with AI on the relationship between mental model congruence and trust. Subsequently, effects of AI familiarity on mental model congruence were examined using linear regression analysis. For the third hypothesis, planned comparison analysis was used to compare trust levels across the four conditions. Regarding exploratory analyses, correlation analysis of effects of trust after the violation

Results

Preliminary Analyses

Table 1

Descriptive statistics and Pearson correlations

	M	SD	1	2	3	4	5
1. Mental Model Congruence	2.09	1.42					
2. AI Familiarity	2.55	1.22	.44**				
3. Perceived Anthropomorphism	2.11	.77	.14	.17			
4. Initial Trust before the Trust Violation	7.56	1.46	.09	-.11	.15		
5. Trust after the Trust Violation	6.87	1.69	.06	.05	.23*	.69**	

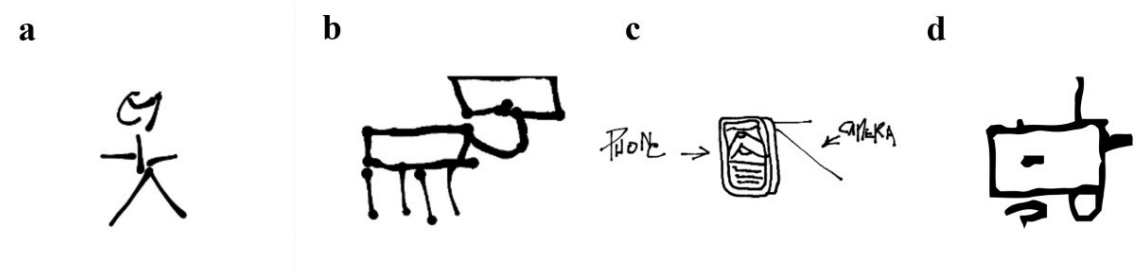
Note. * indicates $p < .05$, ** indicates $p < .01$.

Mental Model Congruence

A descriptive analysis of the distribution of participants' self-reported congruence of their previously expected with the assigned agent was conducted ($M = 2.09$, $SD = 1.42$). Half of the participants (n = 40) indicated no congruence at all, while 19 reported a little congruence. Five participants rated the congruence to be moderate, 6 a lot and 10 a great deal. On average, participants rated the congruence as 'a little', meaning a score of 2 on a 5-point Likert scale.

Figure 4

Exemplary drawings for each agent embodiment condition



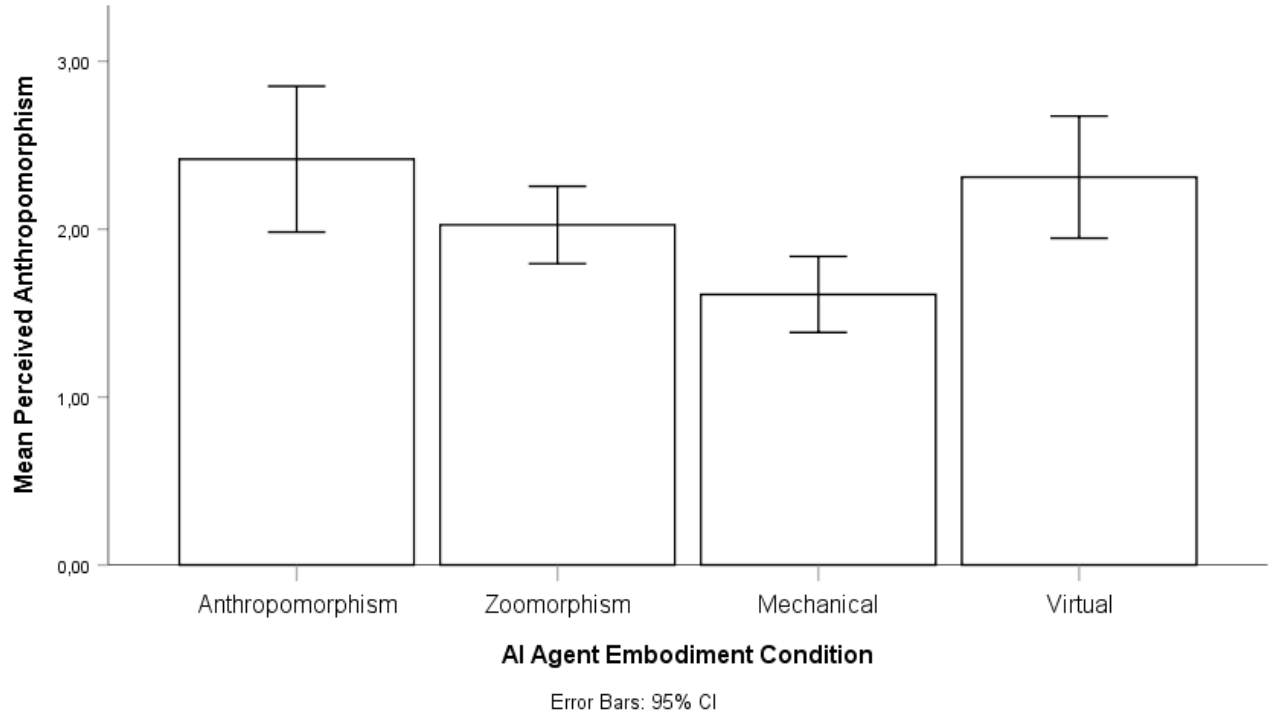
Note. All four agent embodiment conditions are depicted as follows: a) anthropomorphic condition with a self-reported congruence score of 5, b) zoomorphic condition with a self-reported congruence score of 5, c) virtual condition with a self-reported congruence score of 4, d) mechanical condition with a self-reported congruence score of 3. The drawings with the highest possible congruence score were selected for each condition unless the highest rated drawing was unfinished.

Agent Embodiment and Perceived Anthropomorphism

A Planned Comparison UNIANOVA was performed to assess differences in perceived anthropomorphism across the four conditions (anthropomorphic, zoomorphic, virtual and mechanical). It was found that there were significant differences in perceived anthropomorphism between the four types of agents ($F(3, 76) = 4.82$, $p = .004$). Further analysis using contrast specialization showed a significant contrast between the anthropomorphic and mechanical condition ($p < .001$).

Figure 5

Mean Perceived Anthropomorphism across Conditions



Hypothesis Testing

Effect of mental model congruence on initial trust

H1 expected that a low level of mental model congruence would lead to lower levels of initial trust among participants in comparison to high levels of mental model congruence. A linear regression analysis showed that there was no significant effect of mental model congruence on perceived trust ($F(4, 73) = .94$, $p = .447$, $R^2 adjusted = -.01$). This hypothesis was rejected.

Table 2

Linear Regression Analysis of Effect of Mental Model Congruence on Initial Trust

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Mental Model Congruence	.094	.117	-.138	.327	.423

Note. $N = 80$, *CI* = Confidence Interval, *LL* = lower limit, *UL* = upper limit

*Effect of Mental Model Congruence*Familiarity with AI on Initial Trust*

H3 assumed that initial trust would be more resilient to low levels of mental model congruence if paired with a high score on familiarity with AI. A moderation analysis revealed that there is no effect of congruence with familiarity on trust ($F(3,74) = .936$, $p = .653$). Therefore, this hypothesis was rejected. However, an effect of familiarity with AI on mental model congruence was found to be significant ($F(12, 67) = 4.147$, $p < .001$) (see Table 2)

Table 2

Linear Regression Analysis of Effect of AI Familiarity on Mental Model Congruence

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Familiarity with AI	.507	.118	.272	.742	<.001

Note. $N = 80$, *CI* = Confidence Interval, *LL* = lower limit, *UL* = upper limit

Effect of the Anthropomorphic Embodiment on Initial Trust

H1 assumed that participants in the anthropomorphic condition would report higher levels of trust in comparison to the zoomorphic, virtual and mechanical condition. A Planned Comparison ANOVA revealed no significant differences in trust between the four agent conditions ($F(3, 74) = 1.52$, $p = .215$). Therefore, this hypothesis was rejected.

Table 3

Mean Initial Trust Scores across Conditions

Condition	Initial Trust	
	M	SD
Anthropomorphic (N = 20)	7.59	1.31
Zoomorphic (N = 19)	7.98	1.14
Mechanical (N = 18)	6.98	1.58
Virtual (N = 21)	7.63	1.66

Exploratory Analyses

Effect of Initial Trust on Trust after the Violation

As shown in Table 1, trust after the trust violation was significantly correlated with initial trust $r(76) = .69$, $p < .001$. This was confirmed by further analysis using linear regression, which revealed that initial trust significantly predicted trust after the trust violation $F(1, 75) = 67.21$, $p < .001$, $R^2 = .47$, R^2 *adjusted* = .47 with a regression coefficient of $B = .80$. Therefore, it can be concluded that initial trust influences trust after the violation.

*Effect of Time*Perceived Anthropomorphism*

A correlation analysis indicated a significant correlation between trust after the violation and perceived anthropomorphism $r(78) = .23$, $p = .046$. A linear regression analysis showed that perceived anthropomorphism significantly predicted trust after the violation $F(1, 76) = 4.13$, $p = .046$, $R^2 = .05$, R^2 *adjusted* = .04 with a regression coefficient of $B = .49$. Further Repeated Measures ANOVA showed that while there was a statistically significant difference between means of initial trust and trust before the violation $F(1, 76) = 24.91$, $p > .001$. However, it revealed no significant difference with regards to perceived anthropomorphism and its effect on time $F(13, 63) = 1.06$, $p = .407$. Therefore, it can be concluded that perceived anthropomorphism does not predict the development of trust over time.

Discussion

In an effort to find predictors of successful Human-Agent-Interaction (HRI) this study explored the concept of mental models with regard to Artificial Intelligence (AI) embodiment as predictors of trust in interaction with different agent types. It explored the effect of congruence between initial expectations in terms of embodiment (referred to as mental model) and the influence of familiarity with AI. In addition, the influence of agent embodiment on initial trust in the agent was assessed using 4 agent types. The main findings found no support for associations between mental model congruence nor influence of familiarity with AI. Furthermore, agent embodiment was not found to influence trust. Therefore, it can be concluded that mental model congruence and agent embodiment has no effect on trust. In addition, familiarity of AI was found to predict mental model congruence. The following will discuss these findings in more detail.

Agent embodiment did not predict anthropomorphism perception as expected. The setup of the study included four types of agent embodiments (anthropomorphic, zoomorphic, virtual, mechanical), which were perceived by participants to differ in levels of anthropomorphism. As expected, the mechanical agent was perceived as the least, the zoomorphic agent as moderately and the anthropomorphic agent as the most humanoid. This was in line with previous expectations, rendering the manipulation partly successful. However, contrary to expectations, the virtual agent was perceived nearly equally as humanoid as the anthropomorphic condition despite possessing no human-like features. This finding is inconsistent with previous research on the mechanisms of anthropomorphism, which found human characteristics (appearance, mannerisms, reference to self, backstory, voice, name) to be required for an agent to perceived as human-like (De Visser et al., 2017). The virtual agent, however, is just a black screen. Participants anthropomorphising the virtual agent could perhaps be explained by their association of a smartphone with the AI agent ‘Siri’, who has a very human-like voice. Another possible explanation could be that the image

depicts a person using the agent, which may have influenced participants' perception. In that case this result could be attributed to poor choice of imagery, which would represent a considerable limitation of this study. In hindsight, the introduction of the agent also mentions that the agent is able to 'understand natural language and talk back', which could lead to anthropomorphizing as well (Appendix C).

The aforementioned findings explain why trust was not found to be higher in the anthropomorphic condition. Based on research by DeVisser et al. (2016), trust was assumed to be higher in the anthropomorphic agent embodiment. However, as previously explained, three of the four agent embodiments did not differ much in perceived anthropomorphism levels. It can be assumed that similar levels of anthropomorphism lead to similar levels of trust (DeVisser et al., 2016). Generally, agent embodiment did not impact trust initially or after the trust violation. In addition, perceived anthropomorphism was found not to influence the development of trust. This is in contrast to previous findings which found perceived anthropomorphism to be able to increase trust resilience after a violation (De Visser et al., 2017; Kim & Song, 2020).

No support for an influence of mental model congruence on trust was found. This is not in line with previous assumptions that high mental model congruence leads to higher trust levels. Findings showed that accurate expectations about an agent lead to more sustainable trust (Okamura & Yamada, 2020). This is explained by the concept of calibrated trust which states that adjusting one's expectations to the reliability of the AI agent improves resilience of trust in case of a violation (Kim & Song, 2020).

An unexpected correlation was found between mental model congruence and agent familiarity. This could be explained by previous research referenced by Jones et al. (2011) which found that individuals with more experience with artificial intelligence, referred to as AI familiarity in this study, are likely to have more abstract mental modules about AI agents. Abstract mental models are more likely to be congruent with the agent at hand, as they match

a wider range of agents. According to this finding, experience with AI can increase tolerance towards new agent embodiments and thus could facilitate interaction with new AI. In practice, this could mean that education about AI should precede Human-Agent-Teaming (HAT) to facilitate trust.

As mentioned, the design of this study placed an emphasis on the visual appearance of an agent, by prompting participants to draw their expectations of an agent and using it to measure mental model congruence. On the one hand, this was sensible, as anthropomorphism elicited through human-like features has implications for numerous agent characteristics, such as perceived locus of control as well as perceived competence (Epley, Waytz, & Cacioppo, 2007; Ososky et al., 2013). However, visual aspects may not be sufficient to assess agent expectations, as mental models can contain many more facets of a system than just its form. Other important facets include the agent's manner of speaking, tone of voice or a backstory. This reduced mental model congruence to the appearance of the agent, rather than addressing the mental representations participants have about its inner workings such as locus of control. After all, mental models can refer to visual representations of complex processes (Rickheit & Sichelschmidt, 1999). In addition, this task forced participants to reduce their imagined AI to one, when in fact their first reaction would have been multiple agents, a more variable and abstract representation of AI. This issue is addressed by the mental model uncertainty principle as referenced by Kodama et al. (2017), according to which the instruction to draw a mental model can lead to distortions of the elicited mental model. Therefore, future studies could include asking participants to draw or describe multiple aspects of the AI.

Another limitation included the low immersiveness of the experimental environment used in this study, a written scenario paired with images and GIFs of the AI agents, in comparison to a video or using Virtual Reality (VR). It can be assumed that the more the participant feels dependent on the agent, the more accurate are their evaluations of the agent. Alternatively, the written scenario could include more active interactions to increase exposure

to the agent to help the participant gather more information about the agent.

The design of this study included a self-constructed 11-item scale on AI familiarity, which proved to have excellent validity and reliability and can be useful for future research within the field of AI. Based on the results of this study, familiarity with AI can set up individuals to be more open towards new AI agents, as their more abstract mental models are congruent with a larger range of agent embodiments. Therefore, educating individuals on the workings of AI before interacting or working with an agent could facilitate interaction.

This study investigated the impact of mental models on human trust in Artificial Intelligence (AI) agents using four different agent embodiments. In addition, it was assessed how familiarity with AI is related to trust. The main findings revealed no impact of the agent embodiment on trust. Furthermore, no association between mental model congruence and trust, and therefore no effect of AI familiarity, was found. However, AI familiarity showed to impact mental model congruence. Overall, this study made an interesting advance into the field of mental models of AI. Following the implementation of the abovementioned recommendations, further advancements on mental models could be made. Thus far, not much evidence exists concerning the effects of AI familiarity on mental models. Therefore, future study designs could compare the effects of briefing participants on AI.

Conclusion

This study explored the concept of mental models with regard to AI aiming to find means to facilitate human-agent interaction. Contrary to previous findings, it did not confirm that anthropomorphism influences trust or increases trust resilience. Moreover, it found that mental models did not influence trust. However, this study did find support for previous research that found individuals familiar with AI to have more abstract mental models. It can be concluded that familiarity with AI can increase the likelihood that expectations of AI (mental models) match the agent embodiment at hand.

References

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 2-11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

Broadbent, E., Lee, Y. I., Stafford, R. Q., Kuo, I. H., & MacDonald, B. A. (2011). Mental schemas of robots as more human-like are associated with higher blood pressure and negative emotions in a human-robot interaction. International Journal of Social Robotics. http://dx.doi.org/10.1007/s12369-011-0096-9

Craik, K. J. W. 1943. *The nature of explanation.* Cambridge University Press, Cambridge, UK. https://doi-org.ezproxy2.utwente.nl/10.1068%2Fp120233

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3), 331. https://doi.org/10.1037/xap0000092

De Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R. & Krueger, F. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. Human factors, 59(1), 116 133. https://doi-org.ezproxy2.utwente.nl/10.1177%2F0018720816687205

DiSessa, A. A. 1983. Phenomenology and the evolution of intuition. Pages 15-34 *in* D. Gentner and A. Stevens, editors. *Mental models.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Gentner, D., and D. R. Gentner. 1983. Flowing waters or teeming crowds: mental models of electricity. Pages 99-130 *in* D. Gentner and A. Stevens, editors. *Mental models.* Lawrence Erlbaum, Hillsdale, New Jersey, USA.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021).

Attachment and trust in artificial intelligence. Computers in Human Behavior, 115, 106607. https://doi-org.ezproxy2.utwente.nl/10.1016/j.chb.2020.106607

Greeno, G. J. 1983. Conceptual entities. Pages 227-252 *in* D. Gentner and A. Stevens, editors. *Mental models.* Lawrence Erlbaum, Hillsdale, New Jersey, USA.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434. https://doi org.ezproxy2.utwente.nl/10.1177%2F0018720814547570

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80. https://doi org.ezproxy2.utwente.nl/10.1518%2Fhfes.46.1.50_30392

Johnson-Laird, P. N. 1983. *Mental models.* Cambridge University Press, Cambridge, UK.

Jones, N. A., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental models: an Interdisciplinary synthesis of theory and methods. Ecology and Society, 16(1). http://www.jstor.org/stable/26268859

Kim, T., & Song, H. (2020). How should intelligent agents apologize to restore trust? The interaction effect between anthropomorphism and apology attribution on trust repair. https://doi.org/10.1016/j.tele.2021.101595

Larkin, J. H. 1983. The role of problem representation in physics. Pages 75-97 *in* D. Gentner and A. Stevens, editors. *Mental models.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. The Journal of Strategic Information Systems, 17(1), 39-71. https://doi.org/10.1016/j.jsis.2008.01.001

McKnight, D. H., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. In Trust in Cyber-societies (pp. 27-54). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45547-7_3

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. PloS one, 15(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

Ososky, S., Philips, E., Schuster, D., & Jentsch, F. (2013, September). A picture is worth a thousand mental models: Evaluating human understanding of robot teammates. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 1298-1302). Sage CA: Los Angeles, CA: SAGE Publications.

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011, September). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 55, No. 1, pp. 1491-1495). Sage CA: Los Angeles, CA: SAGE Publications. https://doi org.ezproxy2.utwente.nl/10.1177%2F1071181311551310

Rickheit, G., and L. Sichelschmidt. 1999. Mental models: some answers, some questions, some suggestions. Pages 9-40 *in* G. Rickheit and C. Habel, editors. *Mental models in discourse processing and reasoning.* Elsevier, Amsterdam, The Netherlands.

Schaefer, K. E. (2016). Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI." In Robust Intelligence and Trust in Autonomous Systems (Issue August, pp. 1–270). https://doi.org/10.1007/978-1-4899-7668-0_10

Sims,V. K., Chin, M. G., Sushil, D. J., Barber, D. J., Ballion, T., Clark,…Finkelstein, N. (2005). Anthropomorphism of robotic forms: A response to affordances? Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 49(3), 602-605. https://doi.org/10.1177%2F154193120504900383

Tsukada, K., & Niitsuma, M. (2016, October). Impression on Human-Robot Communication Affected by Inconsistency in Expected Robot Perception. In Proceedings of the Fourth International Conference on Human Agent Interaction (pp. 261-262). https://doi.org/10.1145/2974804.2980520

Turner, M., Kitchenham, B., Brereton, P., Charters, S., & Budgen, D. (2010). Does the

technology acceptance model predict actual use? A systematic literature review. Information and software technology, 52(5), 463-479. https://doi.org/10.1016/j.infsof.2009.11.005

Appendix A

Definition of AI

Participants were provided with a short definition of AI after the demographics assessment to make sure all participants understood the concept. It was given as follows. ‘The following questions are about "AI agents". Artificial intelligence (AI) is a digital method of problem solving, aimed at performing tasks that would normally require human intelligence. Subsequently, an AI agent is a digital system that uses data from its environment (like sensors or user input) in order to maximise its chance of achieving its defined goals.

Appendix B

Drawing Exercise Instruction

‘DRAW YOUR AGENT

Please draw a picture of what you think the AI agent in the scenario might look like. We are not interested in your drawing ability—a simple sketch is fine. We are interested in your ideas about the AI agent in this scenario.’

Appendix C

Scenario part 1

‘Imagine you are a humanitarian volunteer aid for The Red Cross, an organization that responds to many disasters around the world. As a humanitarian volunteer, you are a critical resource for crisis response in areas that are affected by disasters like hurricanes, earthquakes, flash floods, wildfires, and more. Due to a shortage of personnel, they ask you to go out in the field today to provide medical care to people in need. However, your medical knowledge is limited. To compensate for this, you will go with an artificial intelligence (AI) agent that is

able diagnose medical problems and that will provide you with instructions on how to adequately help injured people'.

Scenario part 2

`Today you will be working with the AI agent presented below. It has cameras and sensors to observe the environment and it can understand natural language and talk back.'

Scenario part 3

'When you and the AI agent arrive at the site of disaster, you encounter three heavily injured individuals. For you as a lay person, it is hard to identify whose needs are the most urgent. The AI agent scans each of them and determines whose needs are most urgent and who you should help first. The AI agent tells you whom to help and how to help. In retrospect, your partner's assessment turns out to be incorrect. One of the other two injured persons suffers complications, because action was not taken quickly enough. If you would have helped that individual, both individuals would have recovered completely'.

Appendix D

Questionnaire Trust

Instruction

"The following questions are about the AI agent from the scenario. What % of the time will this AI agent be…"

Items

1. …dependable
2. …reliable
3. …unresponsive
4. …predictable
5. …act consistently
6. …function successfully
7. …provide feedback

8. …malfunction
9. …have errors
10. …meet the needs of the mission/task
11. …provide appropriate information
12. …communicate with people
13. …perform exactly as instructed
14. …follow directions

Table D1

Inter-Item Correlation Matrix for the Trust Questionnaire Measured Before the Trust Violation

Item	1	2	3R	4	5	6	7	8R	9R	10	11	12	13	14
Trust 1														
Trust 2	.562													
Trust 3R	.282	.554												
Trust 4	.145	.196	.026											
Trust 5	.447	.659	.463	.283										
Trust 6	.454	.634	.325	.238	.682									
Trust 7	.327	.301	.184	.298	.230	.469								
Trust 8R	.266	.366	.463	.135	.462	.445	.111							
Trust 9R	.318	.455	.409	.045	.520	.604	.244	.676						
Trust 10	.372	.534	.225	.157	525	.702	.438	.253	.342					
Trust 11	.375	.659	.384	.328	.648	.533	.310	.276	.387	.613				
Trust 12	.170	.425	.414	.055	.359	.442	.257	.160	.177	.320	.441			
Trust 13	.505	.440	.415	.360	.583	.546	.367	.453	.447	.491	.564	.263		
Trust 14	.283	.421	.374	.256	.423	.318	.311	.175	.220	.275	.414	.321	.543	

Note. Trust 3R, 8R, 9R represent the reverse formulated items.

Table D2

Component Matrix for the Trust Questionnaire Measured Before the Trust Violation

	Component			
Item	1	2	3	4
Trust 1	.609	.059	.133	-.102
Trust 2	.807	-.040	-.212	-.061
Trust 3R	.604	-.371	-.359	.323
Trust 4	.343	.514	.432	.405
Trust 5	.819	-.084	.016	.038
Trust 6	.826	.007	.096	-.349
Trust 7	.505	.446	.138	-.208
Trust 8R	.574	-.589	.322	.141

Trust 9R	,653	-,513	,280	-,103
Trust 10	,704	,219	,014	-,463
Trust 11	,772	,189	-,139	-,008
Trust 12	,518	,076	-,644	-,046
Trust 13	,765	,094	,216	,256
Trust 14	,570	,251	-,186	,501

Note. The extraction method was Principal Component Analysis (PCA).

Table D3

Inter-Item Correlation Matrix for the Perceived Trustworthiness Questionnaire Measured After the Trust Violation

Item	1	2	3R	4	5	6	7	8R	9R	10	11	12	13	14
Trust 1	—													
Trust 2	.70	—												
Trust 3R	.21	.22	—											
Trust 4	.24	.43	.25	—										
Trust 5	.39	.65	.25	.52	—									
Trust 6	.44	.80	.16	.50	.68	—								
Trust 7	.44	.56	.32	.42	.23	.45	—							
Trust 8R	.40	.59	.15	.31	.52	.65	.30	—						
Trust 9R	.44	.71	.26	.34	.58	.78	.38	.75	—					
Trust 10	.53	.69	.09	.41	.44	.64	.47	.51	.50	—				
Trust 11	.46	.73	.18	.55	.66	.76	.51	.53	.62	.71	—			
Trust 12	.42	.49	.37	.22	.29	.29	.49	.21	.25	.41	.45	—		
Trust 13	.50	.68	.26	.51	.66	.73	.39	.47	.68	.48	.72	.29	—	
Trust 14	.42	.47	.23	.42	.26	.46	.58	.29	.31	.38	.56	.32	.63	—

Note. Trust 3R, 8R, 9R represent reverse formulated items.

Table D4

Component Matrix for the Trust Questionnaire Measured After the Trust Violation

	Component	
Item	1	2
Trust 1	.66	.19
Trust 2	.90	-.04
Trust 3R	.34	.48
Trust 4	.61	.05
Trust 5	.73	-.30
Trust 6	.87	-.29
Trust 7	.64	.52
Trust 8R	.69	-.38
Trust 9R	.79	-.34

Trust 10	.75	-.02
Trust 11	.87	-.05
Trust 12	.52	.55
Trust 13	.82	-.10
Trust 14	.63	.37

Note. The extraction method was Principal Component Analysis (PCA).

Appendix E

Questionnaire Perceived Anthropomorphism

Items

Natural; Fake

Humanlike; Machinelike

Conscious; Unconscious

Lifelike; Artificial

Moves rigidly; Moves Elegantly

Table E1

Inter-Item Correlation Matrix for the Anthropomorphism Questionnaire

Item	1	2	3	4	5R
Anthropomorphism 1					
Anthropomorphism 2	.411				
Anthropomorphism 3	.149	.403			
Anthropomorphism 4	.317	.639	.565		
Anthropomorphism 5R	.037	-.072	.027	-.087	

Note. Anthropomorphism 5R represents the reverse formulated item, which was removed for subsequent analyses.

Table E2

Component Matrix for the Perceived Anthropomorphism Questionnaire

	Component
Item	1
Anthropomorphism 1	.57
Anthropomorphism 2	.84
Anthropomorphism 3	.71
Anthropomorphism 4	.87

Note. The extraction method was Principal Component Analysis.

Appendix F

Questionnaire AI Familiarity

Instruction

"The following questions are about your experience with Artificial Intelligence (AI) and AI agents in general."

Items

1. I was aware that the applications named above were examples of AI
2. I am familiar with AI agents
3. I have interacted with AI agents before
4. I have interacted with robots before
5. I often use AI (agents) in my daily life
6. I feel comfortable using AI (agents)
7. I feel confident using AI (agents)
8. I understand the fundamentals of AI
9. I can explain the difference between machine learning algorithms and deep learning algorithms
10. I am an expert on AI
11. I would like to learn more about AI
12. AI fascinates me

Participants were able to respond on a 5-point Likert scale (*strongly disagree; strongly agree*).

Table F1

Inter-Item Correlation Matrix for the Questionnaire AI Familiarity

Item	1	2	3	4	5	6	7	8	9	10	11	12
AI F. 1												
AI F. 2	.241*											
AI F. 3	.269*	.563**										
AI F. 4	.335**	.267*	.234*									
AI F. 5	171	.474**	.501**	.261*								

AI F. 6	.143	.286*	.400**	.361**	.524**						
AI F. 7	.140	.295**	.465**	.349**	.614**	.798**					
AI F. 8	.190	.357**	.413**	.232*	.306**	.452**	.525**				
AI F. 9	.067	.303**	.218	.217	.230*	.263*	.390*	.589**			
AI F. 10	.045	.014	.019	.206	-.025	.041	.143	.384**	.613**		
AI F. 11	-.052	.095	.209	.145	.072	.112	.224*	.225*	.189	.202	
AI F. 12	.183	.259*	.288**	.424**	.272*	.411**	.475**	.427**	.310**	.226*	.583**

Note. * indicates $p < .05$, ** indicates $p < .01$.

Table F2

Component Matrix for the Questionnaire AI Familiarity

	Component			
Item	1	2	3	4
AI Familiarity 1	.33	-.22	.12	.764
AI Familiarity 2	.59	-.30	-.16	.22
AI Familiarity 3	.67	-.34	.03	-.02
AI Familiarity 4	.54	-.01	.25	.45
AI Familiarity 5	.67	-.45	-.18	-.14
AI Familiarity 6	.73	-.25	-.05	-.29
AI Familiarity 7	.81	-.15	-.07	-.31
AI Familiarity 8	.73	.26	-.26	-.01
AI Familiarity 9	.59	.54	-.41	.03
AI Familiarity 10	.33	.77	-.26	.18
AI Familiarity 11	.40	.389	.66	-.23
AI Familiarity 12	.68	.233	.53	-.04

Note. The extraction method was Principal Component Analysis.