Trust in Artificial Intelligence? The Role of Mental Models, Openness, and Anthropomorphism in Human-Agent Teams

Hannah S. Gräsel
Faculty of Behavioural, Management, and Social Sciences, Department of Psychology
University of Twente, The Netherlands

Graduation Committee members:
Drs Esther S. Kox, Dr Marcel E. Pieterse

July 6th, 2022, Enschede, The Netherlands

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

Abstract

Background. The application of artificial intelligence (AI) as teammates has become increasingly relevant, making humans more dependent on effective use and collaboration. Various studies identified trust as a fundamental dimension to effective collaboration, which is suggested to be enhanced by peoples' accurate mental models of AI, openness towards novelty, and perceived anthropomorphism.

Objective. This study aims to explore how people respond to different AI agents with a focus on mental models, openness, anthropomorphism, and the calibration of trust.

Task and Procedure. An experiment that included several tasks, such as self-report questionnaires and fictional scenarios was designed to assess participants' prior mental models of AI, openness, perceived anthropomorphism, and trust in AI agents. Particularly, participants were randomly assigned to four AI agent conditions which they evaluated in follow-up questionnaires.

Findings. Linear and multiple regression analyses revealed that mental model congruence and openness did not predict the level of initial trust towards an AI agent. Furthermore, panned comparison analysis showed that the level of trust did not depend on the AI embodiment condition a participant was in.

Conclusion. The findings suggest that the assessment of mental models and the appropriate calibration of trust remains a challenge and needs to be further investigated. This study presents critical implications of experimental design choices and measurement of mental models, anthropomorphism, and trust in human-agent teams.

1. An Introduction to Artificial Intelligence in Human-Agent Teams

Today's society is coined by the ever-increasing application of artificial intelligence (AI) technology. Not only does the automatization of tasks through AI applications and tools becomes more relevant, but more prominently, AI is being used in an interactive manner; i.e. utilising technology as teammates to fulfil a certain task. For this study, the following working definition is applied to AI:

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction. Particular applications of AI include expert systems, speech recognition, and machine vision. (Gillath et al., 2021, p.1)

Despite the rise of AI technology, most people do not possess sufficient knowledge (Bansal et al., 2019) and trust in the technology (Kok & Soh, 2020). This lack of knowledge and trust is problematic for effective interaction between humans and AI. For instance, when a user and AI of self-driving cars are not sufficiently aligned with each other, disastrous outcomes can occur. One aspect that has been shown to influence such expectation alignment is the most noticeable characteristic of AI, i.e. the AI embodiment. Therefore, this study aims to investigate different embodiments of AI in social contexts in order to gain insights into predictors for effective human-AI interaction.

Accordingly, AI technology used in human-AI interaction is seen as a social technology as it should be established around a human-centred goal. Such as goal can be achieved by means of a human-agent team (HAT). A HAT refers to "a team consisting of at least one human and one intelligent agent, robot, and/ or other AI or autonomous system" (Kox et al., 2021, p.2). To clarify for this report, the terms AI system and AI agent are interchangeably used to generally refer to different embodiments of AI, ranging from computer systems to embodied agents, such as robots. The terms robot or robotic system is specifically used for embodied robots. A HAT builds the foundation for the transition from tools to teammates as increasingly autonomous and intelligent robots or artificial intelligence interact with humans more naturalistically and mimic a human-human team (Ososky et al., 2013b).

Overall, HATs can be found in various applications. For instance, in the military, robots are deployed to perform tasks in environments hostile to humans, such as conducting search and rescue operations and detecting explosive devices (Ososky et al., 2013a; Ososky et al., 2013b; Kox et al., 2021). Moreover, robotic systems find use on smartphones, including Apple's voice assistant 'Siri' or chatbots that imitate human behaviour (De Visser et al., 2016).

Furthermore, robots operate in the health care sector, being homecare providers and personal assistants (Ososky et al., 2013a; De Visser et al., 2016; Lee & See, 2004; Kox et al., 2021). Thus, AI technology is explored and integrated into several domains of today's society making humans increasingly dependent on effective use and collaboration.

Considering the transition from tools to teammates, certain underlying assumptions and expectations can be detected. First, it is anticipated that with the increasing development of technology, AI agents are able to process the complexity of our world and become active participants in human-agent collaboration (Ososky et al., 2013b). Second, the human teammate of a HAT must have sufficient knowledge about the robot's contributions and limitations in order to establish effective cooperation. However, academics point out peoples' lack of understanding of robotic systems (Jermutus, 2022), potentially leading to misuse and disuse of such systems (Lee & See, 2004). Lee and See (2004) define misuse as "the failures that occur when people inadvertently violate critical assumptions and rely on automation inappropriately" and disuse as "failures that occur when people reject the capabilities of automation" (p.50), both potentially leading to less safety and profitability. Thus, a realistic comprehension and understanding of a robot's functioning and limitations in the context of the interaction are required for effective human-agent collaboration.

1.1 Mental Models

Humans require sufficient knowledge about AI systems in order to make sense of their essence and usability. It has been investigated that in order for a successful human-agent interaction to take place, people need to have an accurate understanding of such technology (Ososky et al., 2013b). People's internal understanding of a system or object can be referred to as mental models, as outlined by Johnson-Laird in 1983 (Al-Diban, 2012). Jones et al. (2011) define a mental model as "a cognitive structure that forms the basis of reasoning, [and] decision making (…), constructed by individuals based on their personal life experiences, perceptions, and understandings of the world. They provide the mechanism through which new information is filtered and stored" (p.1). Considering this definition, mental models are seen as being internal guiding mechanisms of an individual on which a person bases his or her perception of the world. Therefore, people constantly make use of their mental models to interact with the world and engage in certain behaviours. This constant use of mental models is seen as an unconscious rather than conscious process as continuous decisions are made more or less automatically due to people's enduring exposure to their environment.

An important notion is the selective nature of mental models, meaning that due to an individual's construction of mental models over his or her lifetime, mental models do not

always represent the external world accurately. Instead, mental models mirror the person's experience with the external world and expectations of it. Thus, mental models offer an incomplete representation of reality due to the subjective nature of a person's mental models. Furthermore, Jones et al. (2011) mention that "aspects that are represented [within a mental model] are influenced by a person's goals and motives for constructing the mental model as well as their background knowledge or existing knowledge structures" (p.6). Thus, the selective nature of mental models reflects a filter principle through which new information is stored.

Considering the lack of people's knowledge about AI systems, this further implies that people lack proper mental models of such technology. Phillips et al. (2011) point out that "human mental models of intelligent robots are primitive, easily influenced by superficial characteristics, and often incomplete or inaccurate" (p.1491), and that "inaccurate, and/ or incomplete mental models of robotic teammates in high-workload and/or critical situations could ultimately endanger the humans who work with and depend on them" (p.1491). People use existing mental models to make predictions about a new system or object (Ososky et al., 2013b), such as the system's purpose and form, explanations of system functioning, observed system states, and predictions about future states (Ososky et al., 2013a). This is, however, problematic as people's lack of understanding of AI systems leads to incomplete, unstable, and inaccurate mental models. To illustrate, individuals might be intrigued by watching science fiction movies in which they observed robots in the past and consequently hold expectations about robots which might be unrealistic and lead to ineffective human-agent interaction.

1.2 Openness

Despite the selective nature of mental models, previous researchers pointed out that an individual's level of openness can influence his or her acceptance of novel information. Jones et al. (2011) put forward that the "acceptance of new information is also related to personal orientations toward learning" (p.7). This notion makes an individual's willingness to learn something new and/ or their openness towards novelty or even contradictory information to their prior beliefs and its effect on acceptance of novelty interesting to investigate. Therefore, openness might have an influence on peoples' mental models of AI, and consequently, human-agent interaction.

Furthermore, Matthews et al. (2021) mapped the role of personality traits with respect to challenges of interacting with machines and showed that openness is associated with a higher interest in novel systems. It is further suggested that this interest in novelty might be important for engagement with complex and unpredictable machines. In addition, Rossi et al. (2018) found openness to be a significant predictor of facilitated robot interaction. Thus, openness

might play an important role in human-agent interaction especially when people make novel contact with AI agents and lack prior knowledge.

1.3 Anthropomorphism

It has been further investigated that especially people who have no prior experience with AI systems base their perception of a robot on its superficial characteristics. Such characteristics might include the perceived personality traits of a robot based on its features which are assessed in the light of existing mental models (Ososky et al., 2013a). One dimension of superficial characteristics on which people base their assessment of a system is anthropomorphism. Anthropomorphism refers to the "process of attributing human characteristics to animals or non-living entities", which helps people to rationalize the actions or behaviours of non-human objects or things (Phillips et al., 2011, p.1492). Lemaignan et al. (2014) describe anthropomorphism as a dynamic social phenomenon which emerges from the interaction between an AI system and a human user. Overall, research has shown that a system's design features and how they appear to the human user, such as shape, behaviour, and the degree of communication, determine the user's perceived level of anthropomorphism (Lemaignan et al., 2014; Powers & Kiesler, 2006; Phillips et al., 2011).

Next to that, Lemaignan et al. (2014) suggested that the level of anthropomorphism ascribed to one AI system varies across people due to individual and situational differences. Considering individual differences, a person's demographics, individual traits, and other psychological factors, such as emotional states and/ or motivations impact the level of anthropomorphism ascribed to a system (Lemaignan et al., 2014). Additionally, situational factors that influence the level of anthropomorphism include the real or imagined purpose of the system, such as the context in which it is used, the task, and the role in which the system is experienced (Lemaignan et al., 2014). Thus, anthropomorphism can be seen as a multi-faceted concept which is likely to depend on multiple factors that determine the user's perception of the AI system.

Anthropomorphism has been investigated in many studies in the past, indicating that humans perceive AI systems differently depending on their application of anthropomorphic features. Phillips et al. (2011) elaborated on studies which prove that people generally favour robots that are more anthropomorphic and that a human user may develop a "mental model of the robot that is similar to their mental model of humans, causing them to interact with robots and people in the same way" (p.1493). Moreover, studies have shown that people tend to hold richer mental models of anthropomorphic robots compared to mechanic ones (Lemaignan et al., 2014). Additionally, an experiment conducted by Ososky et al. (2013a) demonstrated that

"participants who drew a robot with anthropomorphic or zoomorphic qualities reported more perceived knowledge of their robotic teammate, as well as of their human-robot team [compared to drawings of robots that were mechanical in nature]" (p.1301). Moreover, Powers and Kiesler (2006) reported that the robot's voice and physiognomy changed people's perception of the robot's level of anthropomorphism, knowledge, and sociability. Similarly, Lemaignan et al. (2014) found proof that anthropomorphic robots can elicit social responses from humans which has a positive effect on acceptance.

Despite the fact that anthropomorphism can elicit richer mental models and has a positive effect on the acceptance of AI technology, it should be critically assessed whether anthropomorphism should be triggered within humans. Researchers highlighted that "human-like qualities in computers can create inaccurate models for how computers actually work, often deceiving people into believing these devices poses more capability than they actually do" (Phillips et al., 2011, p.1492). This is especially important when people lack accurate knowledge about a system and consequently form false expectations of its capabilities. Minor anthropomorphic features can lead to the attribution of personality characteristics, such as trustworthiness to an AI system (Sims et al., 2010), making people prone to first impression judgements. Therefore, the application of anthropomorphic features to an AI system should fit its purpose and context.

1.4 Trust

The blurred lines between humans and social technology raise the issue of trust in interactions. It is suggested that the AI system's transition from a tool to a teammate depends on trust (Ososky et al., 2013b), which should be calibrated accurately for effective cooperation. Trust can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p.51). In particular, perceived trust depends on a system's competency to achieve a certain goal, including its ability, reliability, predictability, an agent's characteristics, and its overall purpose (Lee & See, 2015). Related to the concept of trust, the term trustworthiness is used in this paper to refer to the more stable state of being worthy to be trusted (Özer & Zheng, 2017). Hence, the attitude to trust and the assigned level of trustworthiness towards an AI agent are argued to be crucial for effective HATs.

Considering people's lack of appropriate mental models of AI systems, assigned trust can be inaccurate. In the initial trust formation stage, people tend to maintain an automation bias, meaning that they tend to perceive automation as having more authority and being more objective and rational than humans (De Visser et al., 2016). Furthermore, it was found that a

match between peoples' prior mental model and an AI agent enhanced trust and teaming with the agent (Lin et al., 2022). Also, previous research showed that a higher degree of an agent's anthropomorphism led to higher levels of trust and greater resistance to breakdowns in trust regarding the agent (Kox et al., 2021; De Visser et al., 2016; Natarajan & Gombolay, 2020; Kaplan et al., 2021; Jensen et al., 2021). Accordingly, "knowledge estimation, attitudes, and expectancies all play a significant role in the formation of mental models and the level of trust placed in automated systems" (Phillips et al., 2011, p.1492).

Contrarily to the formation of trust, trust can be violated and decrease during an interaction due to the over-estimation of an agent's capabilities and a lack of understanding of its limitations. Humans lose trust in robots more rapidly than trust in humans which might be due to the expectation of a system's perfection (De Visser et al., 2016). Therefore, appropriate trust depends on a sufficiently developed mental model of an AI system which facilitates effective human-agent interaction.

1.5 Current Study

Considering the fact that people lack appropriate mental models of AI agents, accompanied by an inaccurate calibration of trust which people primarily base on a system's superficial characteristics, knowledge is needed about how such systems can be designed to facilitate effective human-agent interaction. Academics have called for continuing research concerning "the impact of specific features on human's expectancies and knowledge-estimation of robots and agents" (Phillips et al., 2011, p.1494). Important dimensions detected in literature are mental models, openness, and anthropomorphism, which are argued to influence peoples' perceived trust in AI systems.

1.6 Research Question and Hypotheses

Hence, the main research question is: *To what extent do peoples' mental models, openness, and perceived anthropomorphism influence the level of trust towards AI agents?* In order to answer this question, this study aims to further explore peoples' mental models of AI systems and their effect on trust. Next to that, the effect of anthropomorphism using different AI agent embodiments is intended to be explored concerning trust. Here, the main focus lies on assessing the trust that people initially assign to AI agents when first exposed to them. Hence, resulting from the literature, the following hypotheses emerged:

H1: *A mismatch between the prior mental model and assigned AI agent (mental model incongruence) leads to lower levels of initially perceived trustworthiness amongst participants, compared to a match between the prior mental model and assigned AI agent (mental model congruence).*

Furthermore, due to the notion in the literature about the effect of openness on the acceptance of novelty and enhanced human-agent interaction, it was hypothesized that:

H2: *The negative effect of mental model incongruence on initially perceived trustworthiness is weakened by openness.*

Considering the impact of anthropomorphic cues on trust as found in previous studies, it was further hypothesized that:

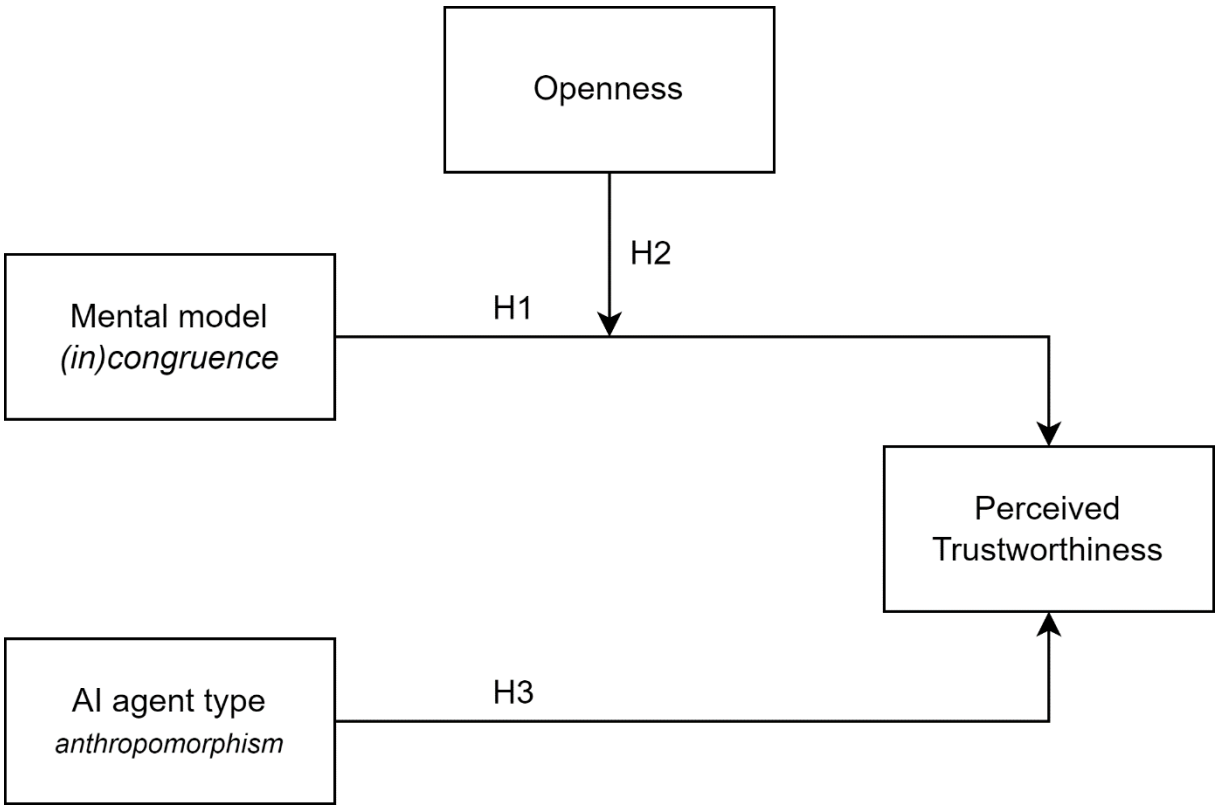
H3: *The anthropomorphic agent condition yields the highest level of initially perceived trustworthiness amongst participants, compared to the zoomorphic, mechanical, and virtual agent conditions.*

1.7 Conceptual Framework

Resulting from the literature review, the conceptual framework in Figure 1 was established to give an overview of the relationships between variables for the hypotheses.

Figure 1

Conceptual Framework



1.8 Additional Exploratory Research

In addition to the main investigations, trust is further examined after a trust violation for exploratory purposes. Such additional investigation was not included as main hypotheses in this study due to limited availability of time and an extensive focus on the investigation of initially perceived trustworthiness. Nonetheless, the design of this study enabled the exploration of a

trust violation in the context of human-AI interaction which was examined subsidiarily due to curiosity.

2. Methods

This study was part of a larger student project, where several more variables were measured apart from the variables of interest included in this report. Therefore, not all variables that had been originally assessed by means of questionnaires were reported here. Nevertheless, all scenario parts of the online experiment were described in this report for the purpose of completeness and additional exploration of variables.

2.1 Design

The first hypothesis was examined by measuring mental model congruence as an independent variable with regards to initially perceived trustworthiness as the dependent variable. For the second hypothesis, openness was treated as an additional independent variable next to mental model congruence which is assumed to moderate the effect on the dependent variable initially perceived trustworthiness. Lastly, the third hypothesis was tested utilizing an experimental between-subjects design, including the independent variable 'Condition' with four different types of AI agents (anthropomorphic, zoomorphic, mechanical, and virtual). Each participant was randomly assigned to one of the four AI agent types. The dependent variable was initially perceived trustworthiness. For exploratory purposes, perceived trustworthiness was measured a second time after the trust violation as a within-subjects variable. Here, a 4 (Condition: anthropomorphic, zoomorphic, mechanical, virtual) × 2 (Time: perceived trustworthiness before the trust violation, perceived trustworthiness after the trust violation) mixed-subjects design was used.

2.2 Participants

Non-probability sampling, including convenience sampling and follow-up snowball sampling were used to recruit participants. Participants who participated through the university's research participation system were offered credits in return for their participation. Treatment of participants was in accordance with the Ethics Committee of Behavioural, Management and Social Sciences (BMS) of the University of Twente. In total, 109 responses were obtained of which 27 remained unfinished and two were test runs, resulting in a final data set of N = 80. The participants' ages ranged from 18 to 65 years (M = 24.7, SD = 8.6). Furthermore, 45% of the participants were male, 50% were female, 2.5% described their gender with the category 'other', and 2.5% preferred not to declare their gender. Considering participants' nationalities, 83.8% were German, 12.5% were Dutch, and 3.9% had another nationality. Moreover, the highest level of education that participants completed resulted in

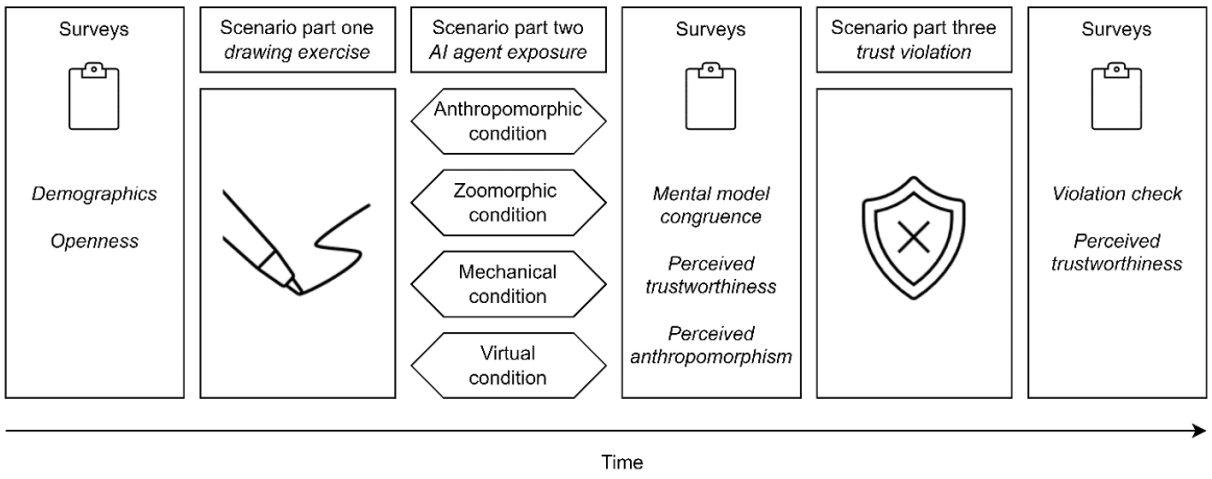
21.3% of participants who finished secondary school, 51.3% who completed college, 20% who held an undergraduate degree, and 7.5% who held a graduate degree. Regarding the four conditions of the experiment, 21 participants were in the anthropomorphic condition, 20 in the zoomorphic condition, 18 in the mechanical condition, and 21 in the virtual condition.

2.3 Task and Procedure

The current study entailed several materials necessary to investigate the hypotheses. Overall, an online environment was created within the software Qualtrics, containing an online experiment and several self-report questionnaires. Therefore, using a computer, tablet, or smartphone was necessary to conduct the study. Furthermore, informed consent had to be signed in order to participate in this study (see Appendix A). In order to give an overview of the study's overall procedure, Figure 2 is displayed below.

Figure 2

The Procedure of the Study: An Overview



2.3.1 Demographics

First, demographic questionnaires were included about the participant's age, nationality, gender, educational level, and field of study (if applicable) in order to assess individual differences.

2.3.2 Openness

In the following, participants filled out an openness questionnaire, which was adopted from Satow (2020). Openness was measured on a 5-point Likert scale, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*), and contained seven items ($M = 3.68$, $SD = .52$) (see Appendix B). An exemplary item was: 'I always enjoy learning new things'. The original reliability of the questionnaire reported in the document of Satow (2020) was good ($\alpha = .76$). However, the reliability found in this study was questionable ($\alpha = .58$). The measure included one reverse formulated item, namely, 'I would prefer everything to stay as it is', which was found to be

problematic. According to the inter-item correlation matrix (see Table B1 in the Appendix), this item demonstrated low and negative correlations with the other items. Removing this item resulted in a Cronbach's Alpha value of .66, representing higher reliability. Therefore, the reverse formulated item was removed from the scale for further analyses. The validity of the openness scale was assessed by conducting Principal Component Analysis, which revealed a Kaiser-Meyer-Olkin value of .67 and a significant Bartlett sphericity test ($p < .001$). A detailed description of the component matrix for the openness questionnaire can be found in Table B2 in the Appendix.

2.3.3 Definition of AI

Participants were provided with a short and general definition of AI, particularly AI agents, to channel their understanding of AI for this study. The definition given was stated 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'.

2.3.4 Scenario Part One

After that, participants were introduced to a fictional scenario consisting of three parts. The overall scenario was that of imagining being a humanitarian volunteer aid for the Red Cross responding to different disasters around the world in cooperation with an AI agent. This context was chosen due to its degree of uncertainty within a critical environment and required dependency on an agent to investigate the variable of interest 'perceived trustworthiness' in the AI agent. The first part of the scenario was stated as follows:

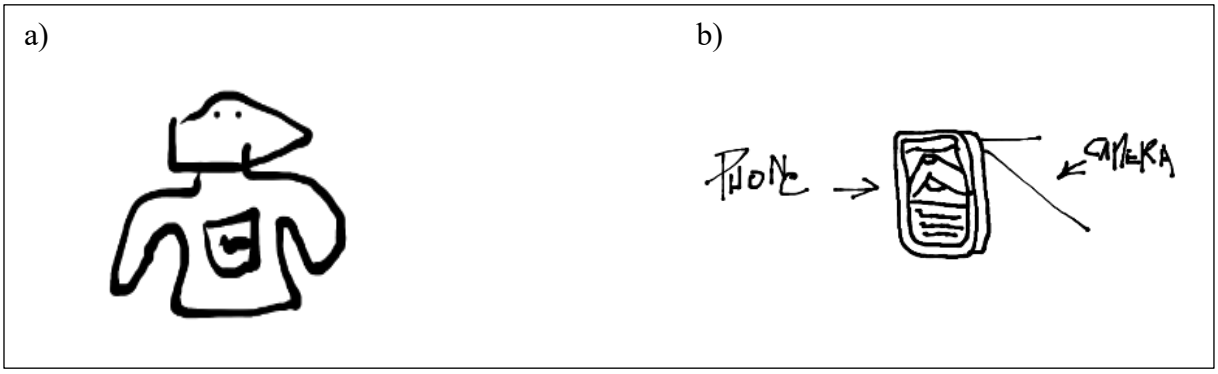
'Imagine you are a humanitarian volunteer aid for The Red Cross, an organization that responds to many disasters around the world. As a humanitarian volunteer, you are a critical resource for crisis response in areas that are affected by disasters like hurricanes, earthquakes, flash floods, wildfires, 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'.

2.3.5 Drawing Exercise

A drawing exercise was incorporated with the instruction: ‘Please draw a picture of what you think your AI-driven buddy 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’. This method was implemented in line with the study of Ososky et al. (2013a) in order to investigate participants’ mental models of robots prior to the exposure to the AI agent. It is argued that the drawing technique elicits “participants’ pre-conceived notions of what a health-care robot might look like [and is] predictive of participants affect toward the actual healthcare robot [and further reveals participants’ expected] descriptions of system form” (Ososky et al., 2013a, p. 1300). Thus, participants’ prior mental models regarding the expected outlook of the AI agent were assessed utilising the drawing exercise. Participants were provided with an empty rectangular box, in which they could draw their sketches (Figure 3).

Figure 3

Mental Model Sketch



Note. Exemplary sketches from two different participants, where a) represents an anthropomorphic and b) a virtual drawing of an expected AI buddy.

2.3.6 Scenario Part Two

The second part was formulated as follows:

‘In fact, the AI agent that will be joining you on your mission today is presented below. It is able diagnose medical problems and that will provide you with instructions on how to adequately help injured people. It has cameras and sensors to observe the environment and it can understand natural language and talk back’.

Within the second part of the scenario, there were four conditions, amongst which participants were divided randomly, namely the anthropomorphic condition, zoomorphic condition, mechanical condition, and virtual condition (see Figure 4). For each condition, a distinct AI agent was selected using a GIF file to demonstrate the agent’s outlook to the participants.

Figure 4

The Four AI Agent Embodiments

Anthropomorphic Condition	Zoomorphic Condition
	
Mechanical Condition	Virtual Condition
	

2.3.7 Mental Model Congruence

Participants' self-reported mental model congruence was assessed, which refers to the degree of correspondence between the participants' expected outlook of an AI agent and the actual AI agent they were exposed to. Participants' expected outlook of an AI agent was expressed by their drawing in the first step. After participants were assigned to an AI agent condition, they were asked to indicate the match between the drawing they made earlier and the given AI agent. Mental model congruence was measured on a 5-point Likert scale, ranging from 1 (*none at all*) to 5 (*a great deal*) (M = 2.09, SD = 1.42). The given item was: 'To what extent does this AI agent match the drawing that you made earlier?'. The self-reported mental model congruence measure was used for further analyses and to test the hypotheses. Participants' original drawings were not used for qualitative coding as previous research demonstrated that "drawings are very illustrative, but they are sometimes hard to interpret" (Dinet & Kitajima, 2011).

2.3.8 Perceived Trustworthiness Measured Before the Trust Violation

The Perceived trustworthiness measure was adopted from Schaefer (2016). Overall, perceived trustworthiness was measured two times, before and after a trust violation. The

perceived trustworthiness measure contained 14 items that were assessed using a rating scale ranging from 0% to 100% with 10% increments. Instructions were: 'The following questions are about the AI agent from the scenario. What % of the time will this AI agent be…'. An exemplary item was: 'predictable', which participants had to rate according to the percentage scale (see Appendix C). Another example for further questions that were stated regarding perceived trustworthiness was: 'What % of the time will this AI agent perform exactly as instructed'. The percentage rating scale was coded in a way that 0% was equal to 1, 10% were equal to 2, 20% were equal to 3, …, and 100% were equal to 11. The first time perceived trustworthiness was measured, a mean of 7.57 and a standard deviation of 1.46 was obtained, meaning that around 65% of the participants trusted the AI agent before the trust violation. The reliability was very good (α = .89). The validity was assessed by conducting Principal Component Analysis, which revealed a Kaiser-Meyer-Olkin value of .81 and a significant Bartlett sphericity test (p < .001). For a detailed view of inter-item correlations and the component matrix see Table C1 and Table C2 in the Appendix.

2.3.9 Perceived Anthropomorphism

Perceived anthropomorphism of the AI agent was measured as one dimension of the Godspeed scale amongst perceived likability and perceived intelligence (Bartneck et al., 2009). The overall Godspeed scale contained 15 word-pair items, of which four positively formulated items and one reverse formulated item measured perceived anthropomorphism. The items were assessed using a 5-point Likert scale, ranging from human characteristics to artificial characteristics. An exemplary word pair was: 'humanlike-machinelike', which participants had to rate accordingly (see Appendix D). The first time the reliability was assessed, Cronbach's Alpha was good (α = .60). However, the inter-item correlation matrix showed that the reverse formulated item had low and negative correlations with the other items (see Table D1 in the Appendix). Consequently, the reverse formulated item was excluded from the reliability analysis which resulted in a higher reliability score (α = .73). Further analyses were conducted without the problematic item, resulting in a mean of 2.11 and a standard deviation of .77. The validity was assessed by conducting Principal Component Analysis, which revealed a Kaiser-Meyer-Olkin value of .69 and a significant Bartlett sphericity test (p < .001). For a detailed view of the component matrix see Table D2 in the Appendix.

2.3.10 Scenario Part Three

The third part was written as follows:

'You and your AI agent have left basecamp. 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. When the AI agent completes its scan, it tells you whom to help first and provides you with precise instructions on how to provide adequate care to that injured individual.

In retrospect, the AI agent's assessment of who was in the most urgent need for help 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 first, all would have recovered completely'.

This part demonstrated the trust violation in order to measure perceived trustworthiness a second time after the AI agent made a mistake.

2.3.11 Violation Check

A violation check was created in order to assess whether participants indeed perceived the AI agent's mistake as a valid mistake. Therefore, it was asked: 'Has the AI agent made a mistake?'. Answer possibilities were provided on a 5-point Likert scale, ranging from 1 (*definitely not*) to 5 (*definitely yes*). Descriptive statistics for the violation check measure revealed a mean of 3.64 and a standard deviation of 1.06.

2.3.12 Perceived Trustworthiness Measured After the Trust Violation

Perceived trustworthiness of the agent was measured a second time after the trust violation. This measure was provided a second time in order to assess possible changes in participants' assigned trustworthiness towards the agent for exploratory purposes. The second time measured, the mean was 6.86 and the standard deviation was 1.69. The reliability was excellent ($\alpha = .92$). The validity was assessed by conducting Principal Component Analysis, which revealed a Kaiser-Meyer-Olkin value of .86 and a significant Bartlett sphericity test ($p < .001$). For a detailed view of inter-item correlations and the component matrix see Table C3 and Table C4 in the Appendix.

2.4 Data Analysis

The gathered data were processed and analysed using the Statistical Package for the Social Sciences (SPSS 25). In the first step, the data set was screened for missing data. In total, 109 responses were recorded of which 27 were unfinished and two were test runs, resulting in a sample size of $N = 80$ used for the analyses. In the following, negatively formulated items of the scales were reverse coded. Next to that, new variables were created by computing mean scores of several variables, including openness, perceived anthropomorphism, and the trust measures before and after the trust violation.

2.4.1 Preliminary Analyses

Preliminary analyses were conducted, such as calculating descriptive statistics and correlations between variables to get an overall impression of the data set. Furthermore, the mental model congruence variable was examined in more detail to achieve a better perspective on participants' self-reported congruence as a basis for the first and second hypotheses. Next to that, linear assumptions were checked for the variables used in the hypotheses. The linear assumptions check included examinations of normality, linearity, homoscedasticity, and the absence of multicollinearity. Moreover, group differences in perceived anthropomorphism between the four conditions were examined by utilizing planned comparison analyses. This was necessary to investigate how participants perceived the AI agents to further assess the third hypothesis.

2.4.2 Hypotheses Testing

In order to assess the first hypothesis, a linear regression analysis was used to examine the influence of mental model congruence on initially perceived trustworthiness. For the second hypothesis, a multiple linear model was tested to assess the hypothesized interaction effect between mental model congruence and openness with regards to initially perceived trustworthiness. The variables mental model congruence and openness were centred and an interaction term between them was created in order to avoid potentially problematic multicollinearity with the interaction term. To examine the third hypothesis, planned comparison analysis was used in order to calculate potential differences in initially perceived trustworthiness between conditions. In particular, contrast specification for the anthropomorphic condition was applied in order to compare each condition to the anthropomorphic one to assess potential differences in initially perceived trustworthiness.

2.4.3 Exploratory Analyses

In addition to the main analyses, exploratory analyses were conducted in order to gain further insights into variables of interest that were not assessed through the hypotheses. Due to the limited availability of time for this research and an extensive focus on the investigation of initially perceived trustworthiness, trust after the violation was assessed exploratorily. Therefore, the development of trust over time was examined in more detail using a paired-samples t-test in order to compare possible differences in the trust measures. Moreover, mental model congruence, openness, and perceived anthropomorphism were assessed as independent variables with regard to the dependent variable perceived trustworthiness measured after the trust violation. Here, use was made of a correlation matrix for the variables. Furthermore, repeated measured analysis of variance (ANOVA) was used to assess possible effects of the variables on the development of trust over time. Such time-dependent repeated measured

ANOVA of trust was necessary to correct for the development of trust over time and give valid conclusions about possible effects. Here, perceived trustworthiness over time was examined as the within-subjects variable and the independent variables as between-subjects effects.

3. Results

3.1 Preliminary Analyses

3.1.1 Descriptive Statistics and Correlations

Mental model congruence, openness, perceived anthropomorphism, and perceived trustworthiness measured before the trust violation were examined using a correlation matrix (Table 1). Results show that there are no significant correlations between the variables.

Table 1

Descriptive Statistics and Pearson Correlations Matrix for Variables

Variable	M	SD	1	2	3	4
1. Mental Model Congruence	2.09	1.42	—			
2. Openness	3.68	.52	-.07	—		
3. Perceived Anthropomorphism	2.11	.77	.14	.13	—	
4. Perceived Trustworthiness Measured Before the Trust Violation	7.56	1.46	.09	.14	.15	—

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

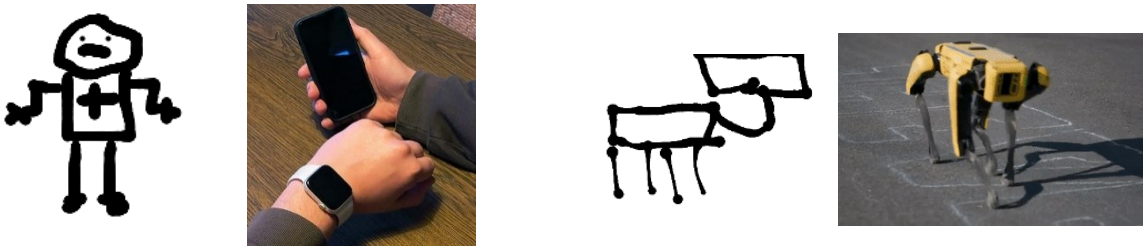
3.1.2 Mental Model Congruence

The distribution of participants' self-reported mental model congruence was examined in order to assess the overall match between participants' expectations and the given AI agent. Descriptive statistics revealed a mean of 2.09 and a standard deviation of 1.42. Considering the distribution of answers for the mental model congruence question, most participants (n = 40) indicated a mismatch between their drawing and the AI agent they were exposed to. Furthermore, 19 participants responded that the AI agent matched their drawing a little, and five participants answered that the AI agent matched their drawing moderately. Furthermore, six participants answered that the AI agent matched their drawing a lot, and 10 participants indicated that the AI agent matched their drawing a great deal. To illustrate how participants rated the degree of mental model congruence, Figure 5 demonstrates participants' drawings and their corresponding self-reported mental model congruence score.

Figure 5

Metal Model Sketch and Mental Model Congruence Score

a) Low mental model congruence b) High mental model congruence



Note. a) Low mental model congruence shows an exemplary drawing of a participant who was assigned to the virtual AI agent condition and indicated a mental model congruence score of 1 (*none at all*). b) High mental model congruence shows an exemplary drawing of a participant who was assigned to the zoomorphic AI agent condition and indicated a mental model congruence score of 5 (*a great deal*).

3.1.3 Linear Assumptions Check

The variables used for the hypotheses were checked for their compliance with linear assumptions. Analyses revealed that all assumptions were met (see Appendix E).

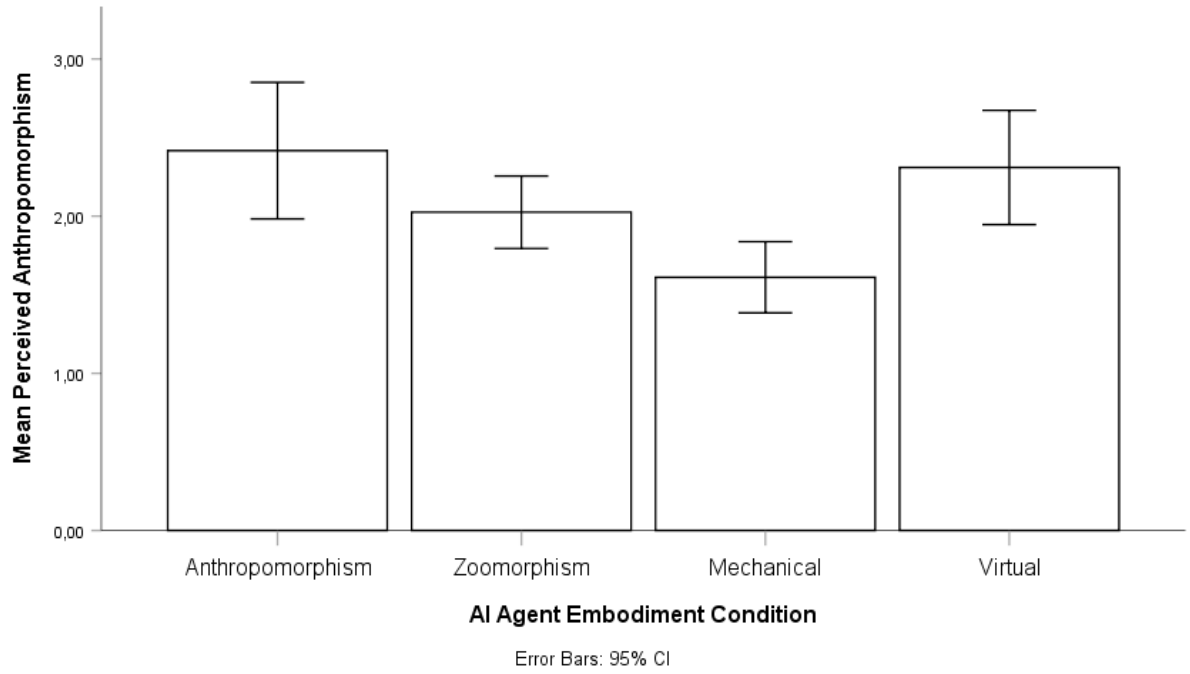
3.1.4 Perceived Anthropomorphism

Differences in perceived anthropomorphism were examined across the conditions in order to check whether participants perceived the different AI agents as intended. A significant between-subjects effect was found with regards to perceived anthropomorphism across the conditions, $F(3, 76) = 4.82$, $p = .004$. Furthermore, each condition was compared to the anthropomorphic condition by adding contrast specialization. This analysis revealed a significant mean difference of .81 between the anthropomorphic condition ($M = 2.42$, $SD = .95$) and mechanical condition ($M = 1.61$, $SD = .46$), meaning that participants in the anthropomorphic condition rated the AI agent as more anthropomorphic than participants in the mechanical condition ($p < .001$). However, no significant differences in perceived anthropomorphism were found between the anthropomorphic condition compared to the zoomorphic condition ($M = 2.03$, $SD = .49$, $p = .084$) and the virtual condition ($M = 2.31$, $SD = .80$, $p = .630$). Surprisingly, a comparison between the mechanical condition and virtual condition revealed a significant mean difference of .70, showing that participants rated the virtual condition as significantly more anthropomorphic than the mechanical condition ($p =$

.003). In Figure 6, the mean scores of perceived anthropomorphism are plotted per condition in order to illustrate the group differences.

Figure 6

Means of Perceived Anthropomorphism Across Conditions



3.2 Hypotheses Testing

3.2.1 First Hypothesis

The first hypothesis, namely *A mismatch between the prior mental model and assigned AI agent (mental model incongruence) leads to lower levels of initially perceived trustworthiness amongst participants, compared to a match between the prior mental model and assigned AI agent (mental model congruence)*, had to be rejected. The linear regression analysis showed that mental model congruence did not significantly predict perceived trustworthiness of the AI agent, $F(1, 76) = .65$, $p = .423$, $R^2 = .01$.

3.2.2 Second Hypothesis

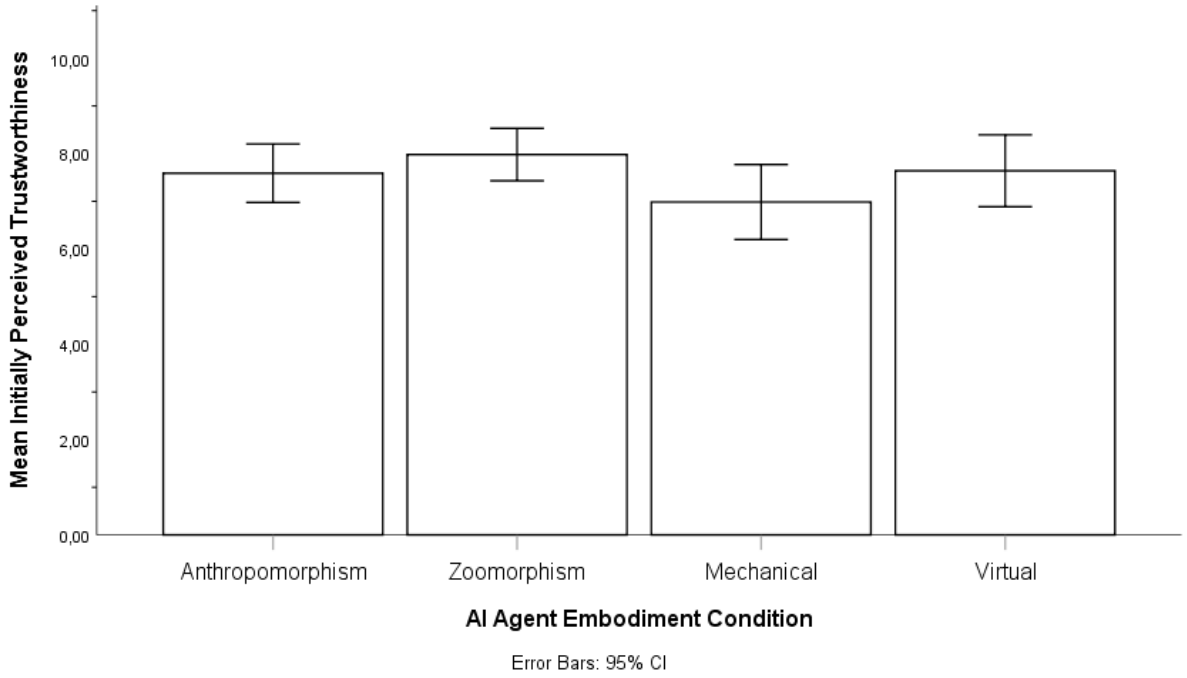
The second hypothesis, namely *The negative effect of mental model incongruence on initially perceived trustworthiness is weakened by openness,* was rejected. The linear regression analysis displayed that mental model congruence and openness did not significantly predict perceived trustworthiness, $F(2, 75) = 1.22$, $p = .301$, $R^2 = .03$. Next, the interaction term between mental model congruence and openness was added, which did not account for a significant proportion of the variance in perceived trustworthiness, $F(3, 74) = .86$, $p = .46$, $R^2 = .03$.

3.2.3 Third Hypothesis

No support was found for the third hypothesis, that *The anthropomorphic agent condition yields the highest level of initially perceived trustworthiness amongst participants, compared to the zoomorphic, mechanical, and virtual agent conditions.* Planned comparison analyses revealed that the group differences did not significantly differ, meaning that the level of perceived trustworthiness did not depend on the condition a participant was in, $F(3, 74) = 1.52$, $p = .215$. Despite the fact that the means across conditions did not significantly differ, the zoomorphic condition yielded the highest mean of perceived trustworthiness compared to the other conditions ($M = 7.98$). The means for the anthropomorphic, mechanical, and virtual conditions were 7.59, 6.98, and 7.64, respectively. In Figure 7, the means of initially perceived trustworthiness are displayed across conditions.

Figure 7

Means of Initially Perceived Trustworthiness Across Conditions



3.3 Exploratory Analyses

3.3.1 Development of Perceived Trustworthiness Over Time

According to the paired-samples t-test used to compare the level of trust before and after the trust violation, a significant difference between the two levels of trust was found, $t(76) = 4.93$, $p < .001$. In particular, initial trust measured before the violation ($M = 7.57$, $SD = 1.46$) decreased by .71 units after the violation took place, resulting in a mean of 6.86 and a standard deviation of 1.69 for trust after the violation.

3.3.2 Variables Influencing the Development of Perceived Trustworthiness Over Time

A correlation matrix revealed no significant correlations between perceived trustworthiness after the trust violation and mental model congruence, $r(78) = .059$, $p = .605$. Also, perceived trustworthiness after the trust violation did not significantly correlate with openness $r(78) = .213$, $p = .061$. Only perceived anthropomorphism was found to significantly correlate with perceived trustworthiness measured after the trust violation, $r(78) = .227$, $p = .046$. Thus, perceived anthropomorphism significantly predicted perceived trustworthiness after the trust violation. In order to assess the effect of anthropomorphism on trust in more detail, repeated measured analysis of variance (ANOVA) was conducted. Results revealed that perceived anthropomorphism did not significantly influence the development of trust over time, $F(13, 63) = 1.06$, $p = .407$. Despite the significant correlation found between the variables, more valid predictions can be made about the effect of perceived anthropomorphism on the development of trust over time considering the ANOVA. The significant correlation might have been accounted for by external factors which are corrected for by the repeated measured ANOVA. Hence it is more valid to conclude that perceived anthropomorphism does not have an effect on the development of trust over time.

Lastly, possible effects of the AI agent conditions on the development of trust over time were examined. The repeated measured ANOVA revealed no significant effect of AI agent condition on the development of trust over time, $F(3, 73) = .31$, $p = .822$. Thus, the AI agent condition a participant was in did not significantly influence the development of trust over time.

4. Discussion

The purpose of this study was to quantify and model the effect of mental models, openness, and anthropomorphism in an explorative way to assess their impact on the interaction with AI agent types, particularly with respect to trust. Therefore, a range of different embodied AI agents was used within an online experiment to answer the research question, namely, to what extent peoples' mental models, openness, and perceived anthropomorphism influence the level of trust towards AI agents. The main results revealed that neither mental model congruence nor the moderation by openness on mental model congruence was associated with perceived trustworthiness, meaning that limited assertions can be made about predictors of trust in AI. Furthermore, the level of perceived trustworthiness was not related to the AI agent embodiment condition (anthropomorphic, zoomorphic, mechanical, virtual). In particular, participants in the anthropomorphic AI agent condition did not rate the agent as more trustworthy than participants in the other conditions. Thus, limited conclusions can be given about the influence of anthropomorphism and AI agent embodiment on initial trust. Exploratory analyses showed that trust in the AI agent decreased after a single trust violation. Despite the

finding that perceived anthropomorphism correlated with perceived trustworthiness after the trust violation, more detailed analyses showed that anthropomorphism did not influence the effect of trust over time. Similarly, the AI agent condition did not have an effect on the development of trust over time.

4.1 First Hypothesis

The rejection of the first hypothesis, namely *A mismatch between the prior mental model and assigned AI agent (mental model incongruence) leads to lower levels of initially perceived trustworthiness amongst participants, compared to a match between the prior mental model and assigned AI agent (mental model congruence)* showed that participants' perceived trustworthiness of an AI agent did not depend on their prior mental model of an AI agent. This result can be interpreted in the light of the argument by Jones et al. (2011), who described that people partially form mental models based on their goals and motives for constructing the mental model, and further, their background knowledge. Accordingly, it is likely that participants in this study did not possess much background knowledge of AI agents, as shown by the variety of participants' drawings of expected agents. The lack of knowledge of AI technology is richly discussed in the literature, further demonstrating its importance regarding trust calibration (Phillips et al., 2011). It can be assumed that participants' lack of knowledge about AI agents, which ultimately implies an incomplete mental model, led to an unsteady rating of trust. This is further supported by Ososky et al. (2013a), as people use existing mental models to make predictions about a new system or object. In this context, this could imply that participants' limited knowledge of AI technology accounted for their unsure prediction of trust towards the AI agent. Consequently, the question that arises here is, how can people calibrate an appropriate level of trust towards an AI agent without sufficient knowledge about such a system? It might be interesting for future research to compare different groups of people, e.g. AI experts and people who are inexperienced with AI, in order to assess possible differences in mental models between the groups and their calibration of trust.

Several limitations to this study can be outlined considering the measurement of the first hypothesis. First, the assessment of mental models by utilizing the drawing exercise as used by Ososky et al. (2013a) must be critically reflected. It is impossible to directly observe mental models as they are complex mental structures within individuals. In particular, Kodama et al. (2017) mention that people have limited self-awareness to express their mental model accurately due to its complexity and often unconscious use. Next to that, participants' self-reported mental model congruence score was used for the analyses without considering more objective evaluations of mental model congruence by the researchers. Due to the limited time

availability and scope of this study, participants' self-reported mental model congruence score was not verified using objective comparisons, such as inter-rater reliability. Furthermore, vague or ambiguous instructions from the researchers might have caused participants to have difficulties in drawing. For instance, participants could misunderstand the definition of AI or the first written part of the scenario. Another issue might be the online drawing tool which could have caused the drawings to be less detailed than they would be on paper. Moreover, Kodama et al. (2017) note the mental model uncertainty principle as outlined by Richardson et al. (1994), meaning that asking people to draw their mental model might already lead to distortions in what is elicited. Thus, assessing mental models remains difficult and requires further research on methodology.

4.2 Second Hypothesis

The rejection of the second hypothesis, namely *The negative effect of mental model incongruence on initially perceived trustworthiness is weakened by openness* indicated that openness did not significantly influence and moderate participants' level of trust towards an AI agent. Previous research discussed openness as a predictor for facilitated robot interaction (Rossi et al., 2018), and acceptance of novelty (Jones et al., 2011). However, as this study demonstrated, openness was not associated with higher levels of trust, meaning that possible parameters might lie between robot interaction, acceptance and trust. Thus, it is important to differentiate between the constructs of trust in AI and the acceptance or interaction with it in literature. Further research could focus on the conceptualization of trust and acceptance and investigate intervening variables between those constructs in the context of human-agent teaming. Furthermore, advanced personality research might reveal additional insights into predictors for trust in AI agents.

4.3 Third Hypothesis

The rejection of the third hypothesis, namely that *The anthropomorphic agent condition yields the highest level of initially perceived trustworthiness amongst participants, compared to the zoomorphic, mechanical, and virtual agent conditions* did not offer support for the claim that anthropomorphic AI agents yield higher levels of trust, which is contrary to the findings of Kox et al. (2021), De Visser et al. (2016), Natarajan and Gombolay (2020), Kaplan et al. (2021), and Jensen et al. (2021). Overall, the manipulation of AI agent embodiment utilizing the four conditions was part of an exploration of how people respond to different design features. Despite assumed differences in initially perceived trustworthiness depending on the condition a participant was in, participants did not evaluate the AI agents differently on trust across conditions. Nonetheless, speculating that differences in initially perceived trustworthiness exist

across conditions, participants rated the zoomorphic AI agent as most trustworthy, followed by the virtual, anthropomorphic, and mechanical AI agents. Interpreting such results, it might be the case that the zoomorphic AI agent was appealing to participants as it might have evoked associations with familiar pets or rescue dogs in disaster contexts. Moreover, participants might have found the zoomorphic agent the most effective in fulfilling the shared task of helping people due to its robust movement on four 'legs', and thus, trusted this agent the most. Furthermore, it is plausible to assume that participants trusted the virtual AI agent in the second place due to their familiarity with smartphones and watches and the technology's integration into society. Moreover, it can be assumed that participants do not have a lot of prior experience with humanoid or mechanical AI agents, hence resulting in lower levels of perceived trustworthiness towards the anthropomorphic and mechanical agents.

Next to the assessment of trust, the level of perceived anthropomorphism across conditions yielded surprising results. Almost no differences in perceived anthropomorphism were found between the anthropomorphic and zoomorphic, as well as the anthropomorphic and virtual conditions. However, the virtual agent was found to be rated as more anthropomorphic than the mechanical agent. Finding explanations for these results, it can foremost be assumed that participants associated some level of anthropomorphism with each of the conditions. Considering previous findings in the literature, anthropomorphism can be triggered even by subtle anthropomorphic cues and is further dependent on individual and situational factors (Lemaignan et al., 2014; Sims et al., 2010). Participants might have associated anthropomorphism with the virtual agent, such as a smartphone's Siri voice assistant, making them rate the virtual agent as more anthropomorphic. Another possible factor that might have accounted for the unforeseen high rating of anthropomorphism for the virtual agent might be the content of the GIF used for this condition. This GIF did not only present the AI technology (smartphone and smartwatch) used for the fictional mission, but also a human holding and wearing this technology. Consequently, participants might have assigned significantly higher levels of perceived anthropomorphism towards the virtual condition compared to the mechanical condition. This can be seen as a limitation of this study, demonstrating that the elicitation of anthropomorphism must be considered thoughtfully.

Another limitation of the study might be the overall limited strength of the experimental scenarios in affecting participants' evaluations of the AI agent. The experiment consisted of written fictional scenarios and the AI agents were shown as GIFs, probably limiting participants' intensities of responses toward the agents. Studies that included other exposure techniques, such as virtual reality or physical in-person meetings with AI agents might elicit

more ecologically valid responses from participants regarding their perception of an AI agent, and consequently, assigned trust. Moreover, according to Lemaignan et al. (2014), Powers and Kiesler (2006), and Phillips et al. (2011), the robot's design features and how they appear to the human user, such as shape, behaviour, and the degree of communication, determine the user's perceived level of anthropomorphism. These findings support the notion that higher degrees of interaction with the AI agent than what was possible in this study might yield clearer differences in perceived anthropomorphism and trust between conditions. Thus, anthropomorphism confirmed itself as a multi-faceted concept which is likely to depend on multiple factors that determine the user's perception of the robotic system. Hence, it is suggested that future research focuses on more interactive possibilities with AI agents when examining anthropomorphism and its effect on trust.

4.4 Exploratory Analyses

Exploratory analyses revealed that perceived trustworthiness decreased after a single trust violation. Thus, findings show that participants lost a noticeable amount of trust after the AI agent made a mistake. Consequently, it can be assumed that in the fictional scenario of helping injured individuals in a disaster context, trust is an important component for people dependent on the help of an AI agent. Trust is especially important in situations of uncertainty and vulnerability (Lee & See, 2004). Therefore, a mistake made by an AI agent might have ultimately led to a decrease in trust as the individual's goal to help injured people is not completely achieved and leaves that person disappointed.

The fact that neither perceived anthropomorphism nor the AI agent conditions influenced the development of trust over time is not in line with the findings of previous research. It has been found that anthropomorphism leads to greater resistance to breakdowns in trust toward an AI agent (Kox et al., 2021; De Visser et al., 2016; Natarajan & Gombolay, 2020; Kaplan et al., 2021; Jensen et al., 2021) which could not be replicated in that study. Nonetheless, contradictory findings can be considered as important contributions to the research domain of human-agent interaction as this area is still at its beginning of investigations.

4.5 Design Recommendations

Lastly, this study's goal was to further give AI agent design recommendations for effective human-agent collaboration. As Phillips et al. (2011) mentioned, most humans hold inaccurate mental models of AI agents, which can be confirmed by this study considering the results of participants' versatile AI agent drawings. Such inaccurate mental models could endanger humans in critical situations who work with them. Moreover, exploratory analyses showed that a single trust violation of an AI agent is sufficient for trust to decrease in

individuals. Therefore, it is of priority to inform more people about the functioning and limitations of AI agents. In particular, such systems should be designed with transparency, meaning that their purposes and capabilities should be displayed to users. Only when the AI agent's essence and capability can be evaluated rationally, trust can be calibrated appropriately. Furthermore, the future of HATs should be established by utilising training units between AI agent experts/ developers, the AI agent, and its critical users (e.g. military units, firefighters). By spreading knowledge about AI agents, the risk of misuse, disuse, and sole reliance on superficial characteristics, such as anthropomorphism, might be reduced.

5. Conclusion

This study was aimed at investigating how people respond to different AI agent types in a social context to gain insights into predictors for effective human-agent interaction. Trust was investigated as a key element to successful interaction, indicating that the appropriate calibration of trust for AI agents remains a challenge. Nonetheless, mental models are assumed to play a role in trust calibration, meaning that people need to possess knowledge about AI agents in order to consciously and rationally assign trust in the initial formation stage and after a trust violation. Furthermore, this research can be used as food for thought in the research domain of human-agent interaction as it shows critical implications of the experimental design choices and the measurement of mental models, anthropomorphism, and trust. Lastly, recommendations were given regarding AI agent design and cooperation within a team, suggesting that the future of HATs not only depends on the design of the AI agent but also on AI-educated team members to ensure appropriate trust calibration and effective teaming.

References

Al-Diban S. (2012). Mental Models. *Encyclopedia of the Sciences of Learning*. https://doi-org.ezproxy2.utwente.nl/10.1007/978-1-4419-1428-6_586

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 2429-2437. https://doi.org/10.1609/aaai.v33i01.33012429

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*, 71-81. https://doi.org/10.1007/s12369-008-0001-3

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Dinet, J., & Kitajima, M. (2011). "Draw me the Web". Impact of mental model of the Web on information search performance of young users. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2044354.2044358

Gillath, O., Ai, T., Branicky, M., Keshmiri, S., Davison, R., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*. https://doi.org/10.1016/j.chb.2020.106607

Jensen, T., Khan, M. M. H., Fahim, M. A. A., & Albayram, Y. (2021). Trust and Anthropomorphism in Tandem: The Interrelated Nature of Automated Agent Appearance and Reliability in Trustworthiness Perceptions. *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*, 1470-1480. https://doi.org/10.1145/3461778.3462102

Jermutus E, Kneale D, Thomas J and Michie S. (2022). Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review. *Wellcome Open Research*. https://doi.org/10.12688/wellcomeopenres.17550.1

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). https://doi.org/10.5751/ES-03802-160146

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors*. https://doi.org/10.1177/00187208211013988

Kodama, C., St. Jean, B., Subramaniam, M., & Taylor, N. G. (2017). There's a creepy guy on the other end at Google!: engaging middle school students in a drawing activity to elicit their mental models of Google. *Information Retrieval Journal*, *20*(5), 403–432. https://doi.org/10.1007/s10791-017-9306-x

Kok, B. C., & Soh, H. (2020). Trust in Robots: Challenges and Opportunities. *Current Robotics Reports*, *1*(4), 297–309. https://doi.org/10.1007/s43154-020-00029-y

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, *35*(30), 1-20. https://doi.org/10.1007/s10458-021-09515-9

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors and Ergonomics Society, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lemaignan, S., Fink, J., Dillenbourg, P., & Braboszcz, C. (2014). The Cognitive Correlates of Anthropomorphism. *Proceedings of the 2014 Human-Robot Interaction Conference*, 1-6. http://infoscience.epfl.ch/record/196441

Lin, J., Panganiban, A. R., Matthews, G., Gibbins, K., Ankeney, E., See, C., … Long, M. (2022). Trust in the Danger Zone: Individual Differences in Confidence in Robot Threat Assessments. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.601523

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., & Wohleber, R. W. (2021). Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personality and Individual Differences*, *169*. https://doi.org/10.1016/j.paid.2020.109969

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, 33-42. https://doi.org/10.1145/3319502.3374839

Özer, Ö., & Zheng, Y. (2017). *Trust and Trustworthiness*. http://dx.doi.org/10.2139/ssrn.3046303

Ososky, S., Philips, E., Schuster, D., & Jentsch, F. (2013a). A picture is worth a thousand mental models: Evaluating human understanding of robot teammates. *Proceedings of*

the Human Factors and Ergonomics Society, 57(1), 1298-1302. https://doi.org/10.1177/1541931213571287

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. (2013b). Building appropriate trust in human-robot teams. *AAAI Spring Symposium - Technical Report*, 60-65. https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5784/6008

Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. *Unmanned Systems Technology XVI*, *9084*. https://doi.org/10.1117/12.2050622

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. *Proceedings of the Human Factors and Ergonomics Society, 55*(1), 1491-1495. https://doi.org/10.1177/1071181311551310

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *HRI 2006: Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, 218-225. https://doi.org/10.1145/1121241.1121280

Rossi, S., Santangelo, G., Staffa, M., Varrasi, S., Conti, D., & Di Nuovo, A. (2018). Psychometric Evaluation Supported by a Social Robot: Personality Factors and Technology Acceptance. *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, 802-807. https://doi.org/10.1109/ROMAN.2018.8525838

Satow, L. (2020). *B5T® Big-Five-Persönlichkeitstest: Test- und Skalendokumentation.* https://www.drsatow.de/tests/persoenlichkeitstest/

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

Sims, V. K., Chin, M. G., Yordon, R. E., Sushil, D. J., Barber, D. J., Owens, C. W., Smith, H. S., Dolezal, M. J., Shumaker, R., & Finkelstein, N. (2010). When function follows form: Anthropomorphism of artifact "faces." *Proceedings of the Human Factors and Ergonomics Society, 49*(3), 595-597. https://doi.org/10.1177/154193120504900381

Appendices

Appendix A

Informed Consent

Welcome to this research study!

We are interested in understanding how people see and understand Artificial Intelligence (AI). In this study, you will be presented with a scenario involving an AI agent and you will be asked a series of questions.

The fictional scenario involves a natural disaster. If you find this distressing, please be assured that your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice.

Please be assured that your responses will be kept completely confidential.

The study should take you around 30 minutes to complete.

If you have any questions about the study, you can contact the research team under h.s.grasel@student.utwente.nl.

By clicking the button below, you acknowledge that your participation in the study is voluntary, you are at least 18 years of age, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

- ○ I consent, begin the study
- ○ I do not consent, I do not wish to participate

Appendix B

Openness Questionnaire

Openness

The following questions are about you.

For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes you.

	Does not describe me	Describes me slightly well	Describes me moderately well	Describes me very well	Describes me extremely well
I always want to try new things	○	○	○	○	○
I am a curious person	○	○	○	○	○
I would prefer everything to stay as it is	○	○	○	○	○
I always enjoy learning new things	○	○	○	○	○
I have many ideas and a vast imagination	○	○	○	○	○
I read a lot about scientific topics, new discoveries or historical events	○	○	○	○	○
I like to discuss things	○	○	○	○	○

Table B1

Inter-Item Correlation Matrix for the Openness Questionnaire

Item	1	2	3R	4	5	6	7
Openness 1	—						
Openness 2	,284	—					
Openness 3R	-,038	,016	—				
Openness 4	,523	,337	,072	—			
Openness 5	,289	,185	,022	,254	—		
Openness 6	,020	,276	-,094	,184	,185	—	
Openness 7	,117	,305	,105	,389	,230	,298	—

Note. Openness 3R represents the reverse formulated item.

Table B2

Component Matrix for the Openness Questionnaire

	Component	
Item	1	2
Openness 1	,625	-,632
Openness 2	,650	,143
Openness 4	,768	-,273
Openness 5	,550	-,095
Openness 6	,471	,676
Openness 7	,631	,390

Note. The extraction method was Principal Component Analysis.

Appendix C

Perceived Trustworthiness Questionnaire

The following questions are about the AI agent from the scenario.

What % of the time will this AI agent be...

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
...dependable	○	○	○	○	○	○	○	○	○	○	○
...reliable	○	○	○	○	○	○	○	○	○	○	○
...unresponsive	○	○	○	○	○	○	○	○	○	○	○
...predictable	○	○	○	○	○	○	○	○	○	○	○

What % of the time will this AI agent...

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
...act consistently	○	○	○	○	○	○	○	○	○	○	○
...function successfully	○	○	○	○	○	○	○	○	○	○	○
...provide feedback	○	○	○	○	○	○	○	○	○	○	○
...malfunction	○	○	○	○	○	○	○	○	○	○	○
...have errors	○	○	○	○	○	○	○	○	○	○	○
...meet the needs of the mission/task	○	○	○	○	○	○	○	○	○	○	○
...provide appropriate information	○	○	○	○	○	○	○	○	○	○	○
...communicate with people	○	○	○	○	○	○	○	○	○	○	○
...perform exactly as instructed	○	○	○	○	○	○	○	○	○	○	○
...follow directions	○	○	○	○	○	○	○	○	○	○	○

Table C1

Inter-Item Correlation Matrix for the Perceived Trustworthiness 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 C2

Component Matrix for the Perceived Trustworthiness Questionnaire Measured Before the Trust Violation

Item	Component			
	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.

Table C3

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 the reverse formulated items.

Table C4

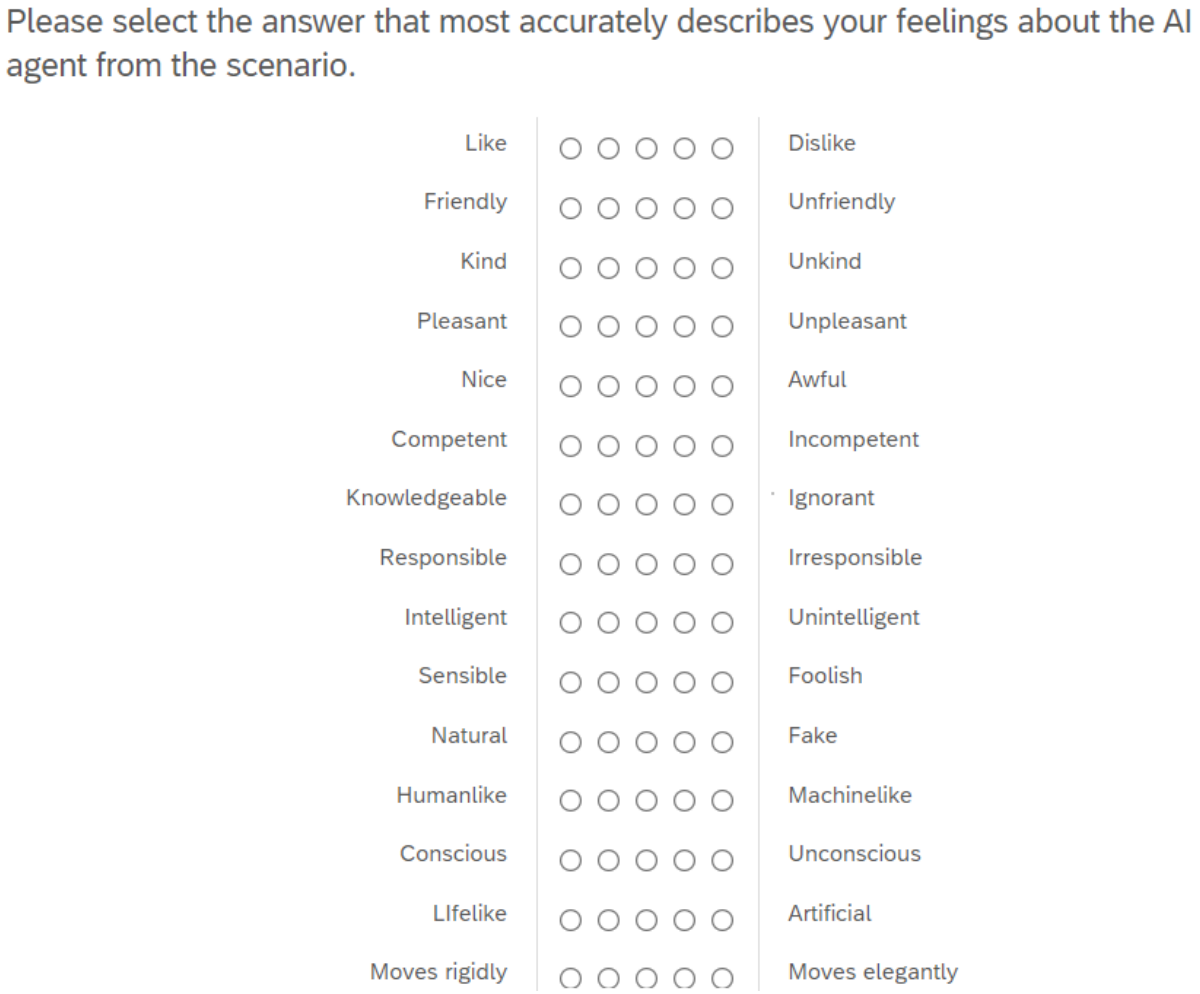
Component Matrix for the Perceived Trustworthiness Questionnaire Measured After the Trust Violation

	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.

Appendix D

Perceived Anthropomorphism Questionnaire



Note. Perceived anthropomorphism is measured using the five word-pairs at the end of the questionnaire, ranging from natural-fake to moves rigidly-moves elegantly

Table D1

Inter-Item Correlation Matrix for the Anthropomorphism Questionnaire

Item	1	2	3	4	5R
Anthropomorphism 1	1,000	,411	,149	,317	,037
Anthropomorphism 2	,411	1,000	,403	,639	-,072
Anthropomorphism 3	,149	,403	1,000	,565	,027
Anthropomorphism 4	,317	,639	,565	1,000	-,087
Anthropomorphism 5R	,037	-,072	,027	-,087	1,000

Note. Anthropomorphism 5R represents the reverse formulated item.

Table D2

Component Matrix for the Anthropomorphism Questionnaire

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

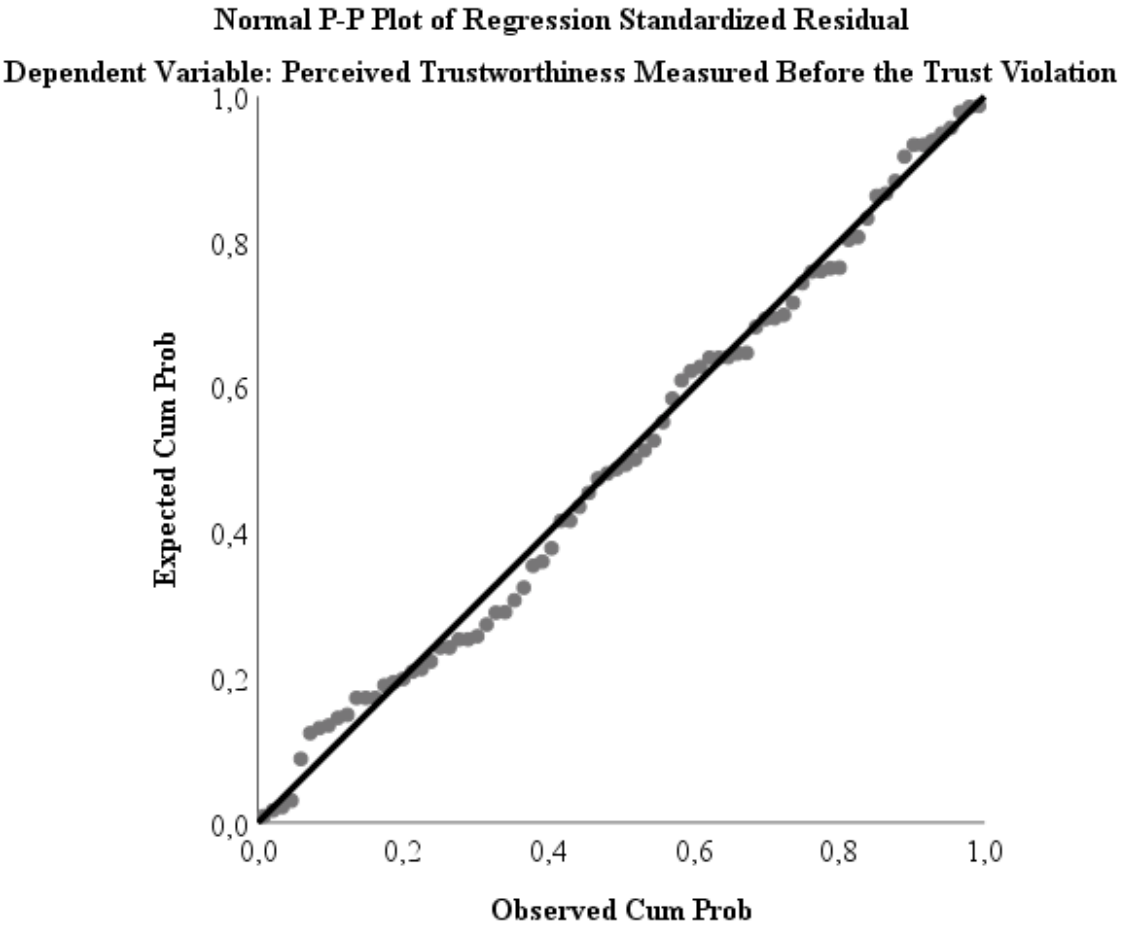
Note. The extraction method was Principal Component Analysis.

Appendix E

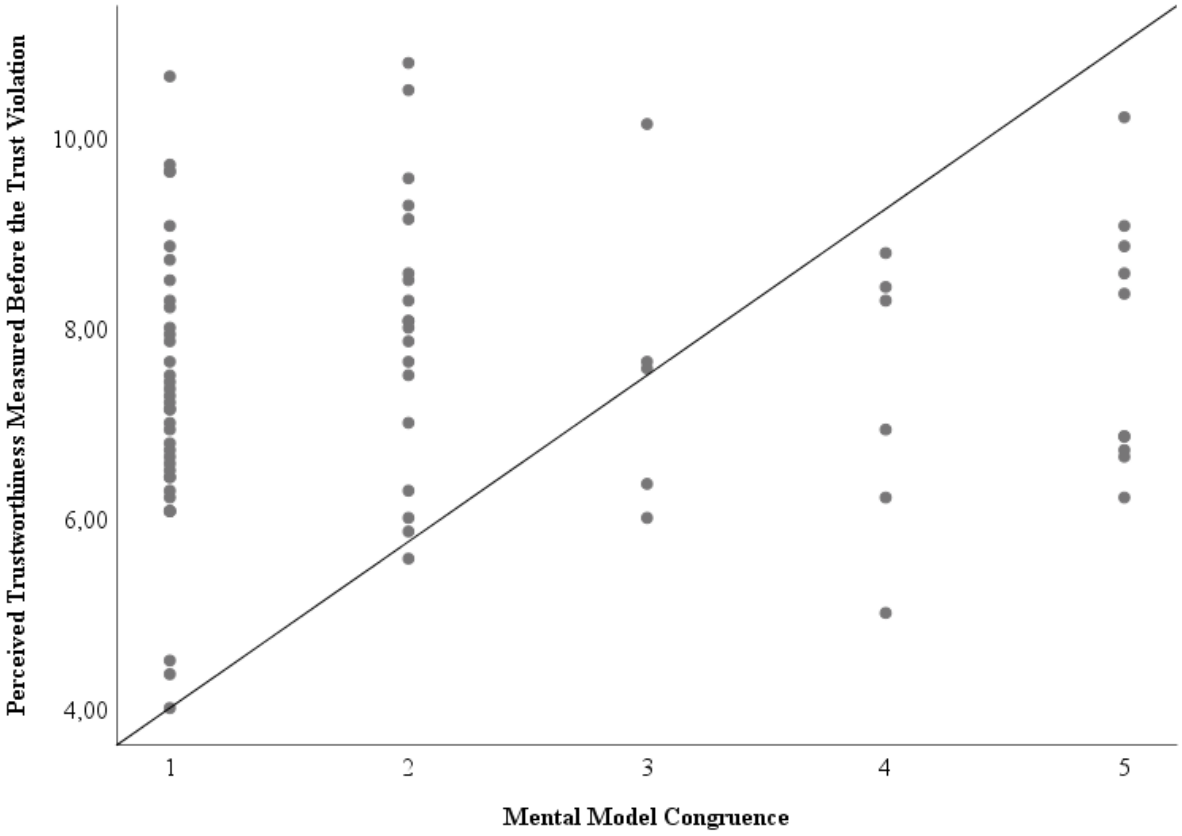
Linear Assumptions Check

H1: *A mismatch between the prior mental model and assigned AI agent (mental model incongruence) leads to lower levels of perceived trustworthiness amongst participants, compared to a match between the prior mental model and assigned AI agent (mental model congruence).*

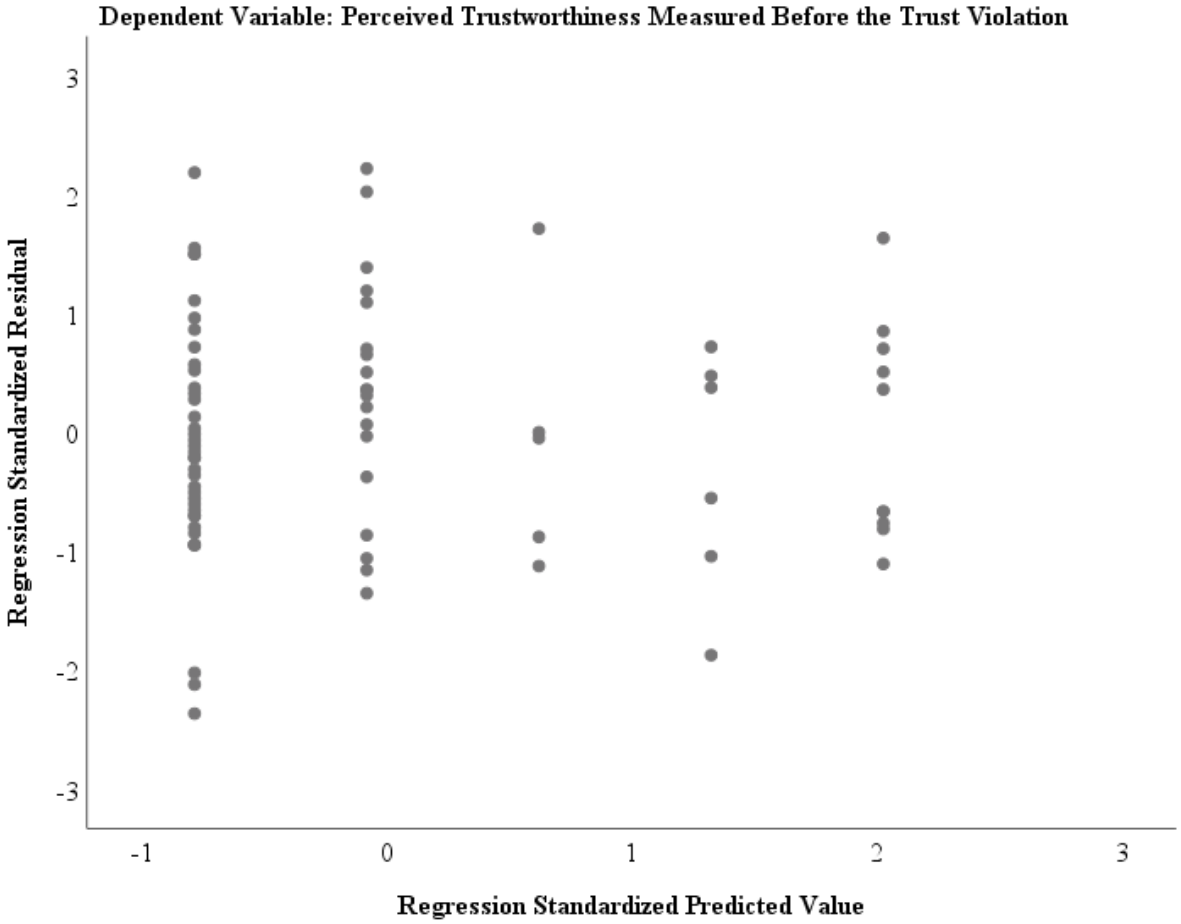
1.1 Normality Check for Mental Model Congruence



1.2 Linearity Check for Mental Model Congruence

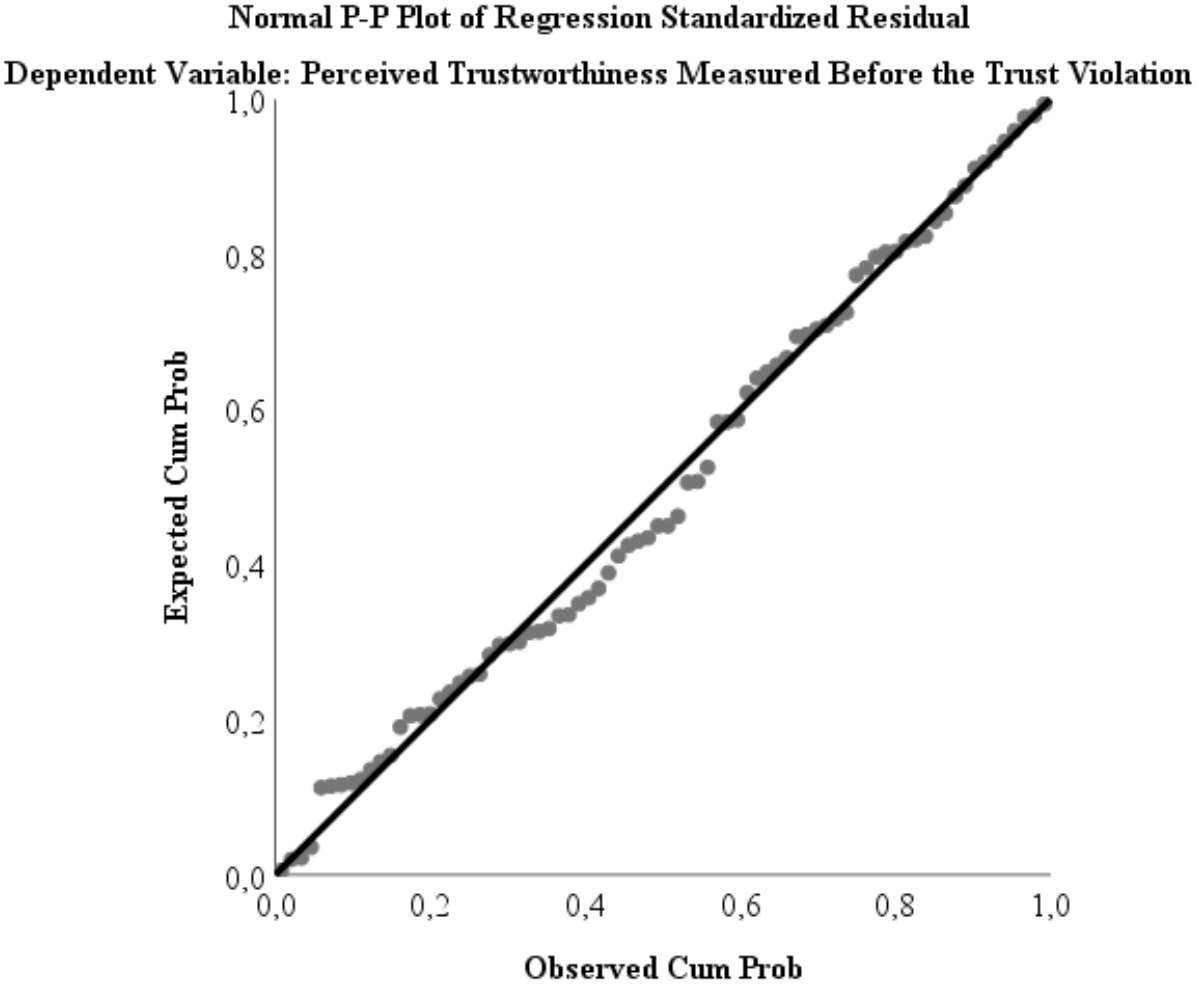


1.3 Homoscedasticity Check for Mental Model Congruence

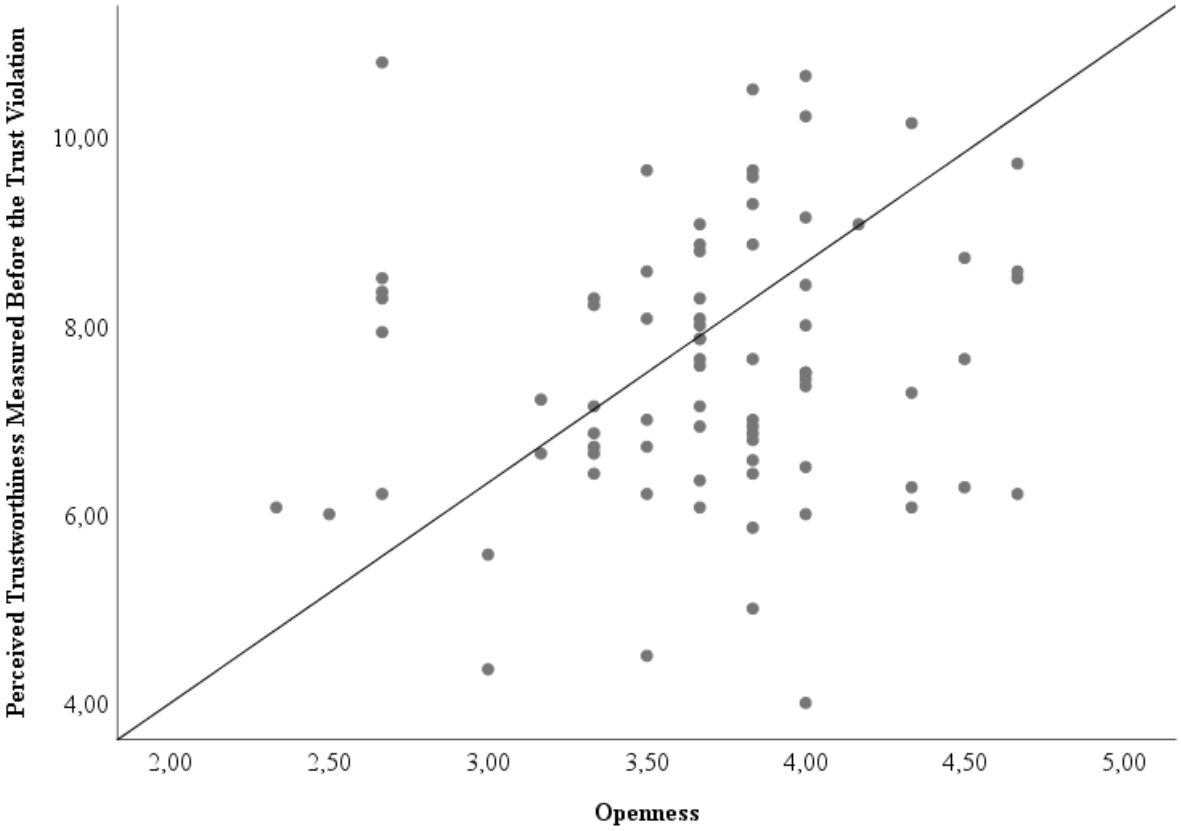


H2: *The negative effect of mental model incongruence on perceived trustworthiness is weakened by openness.*

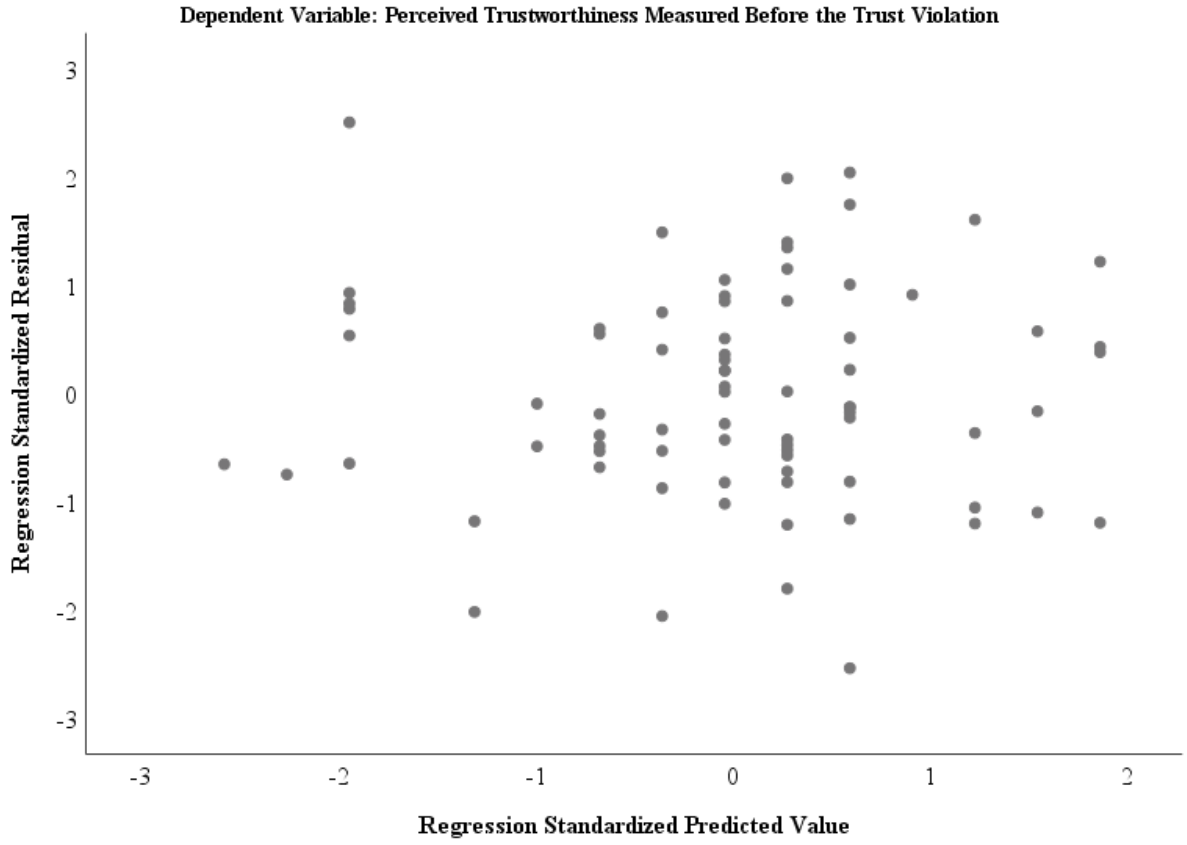
2.1 Normality Check for Openness



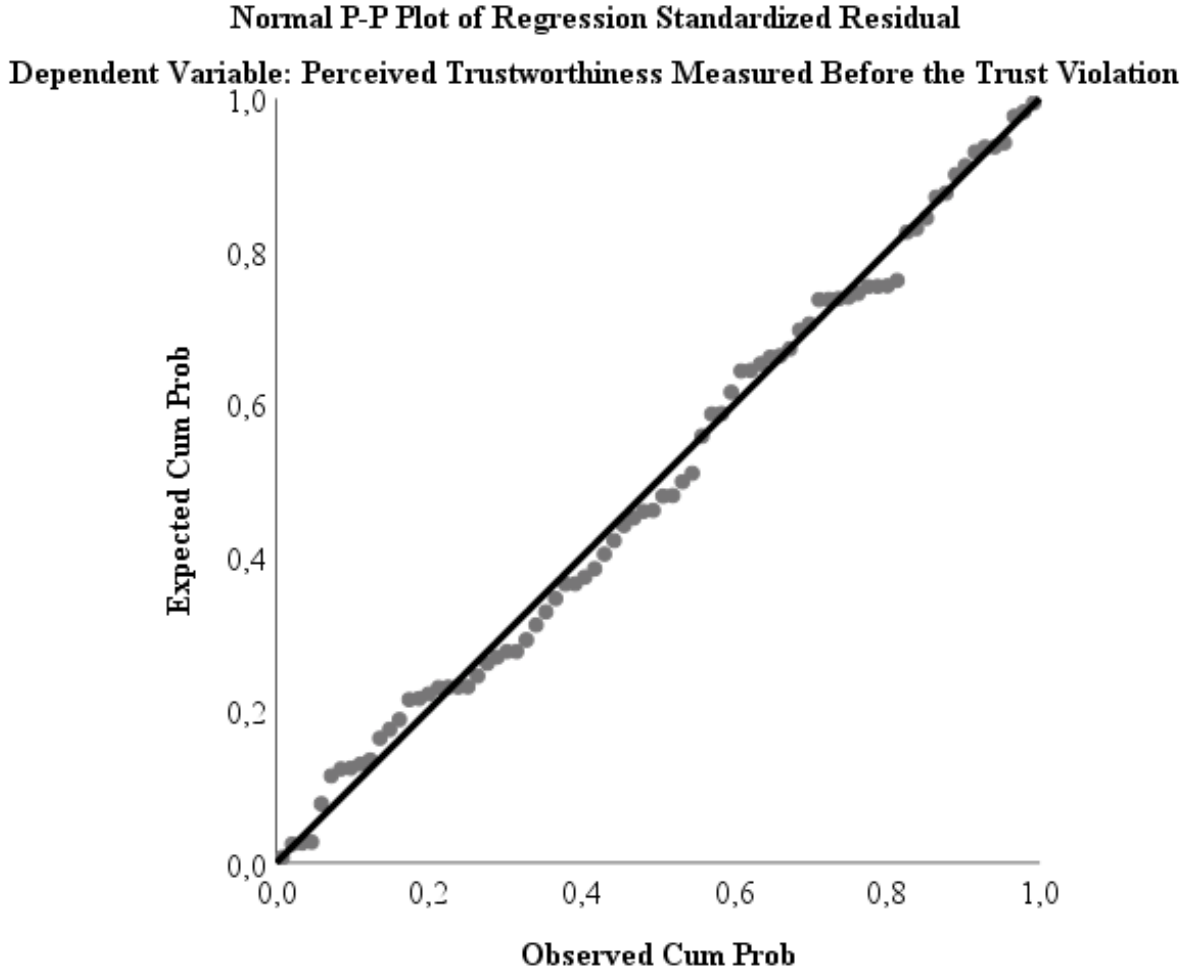
2.2 Linearity Check for Openness



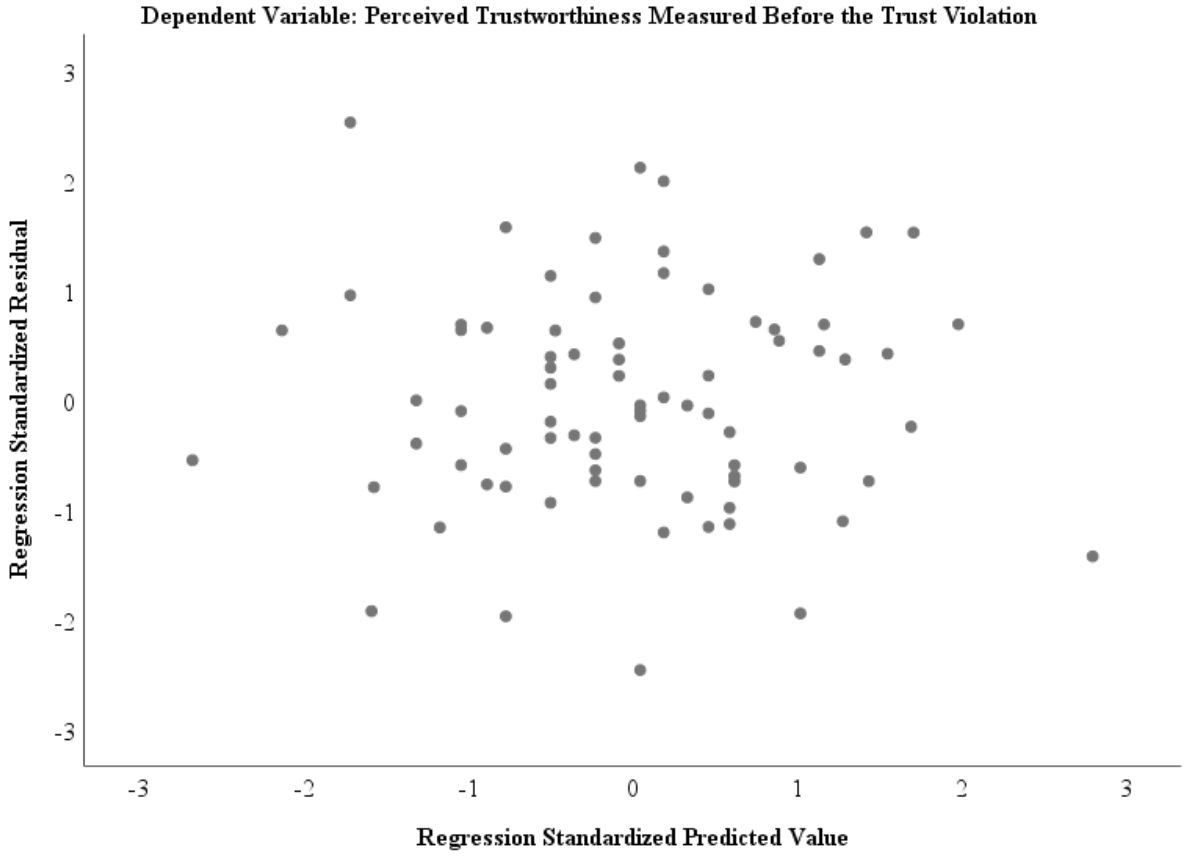
2.3 Homoscedasticity Check for Openness



2.4 Normality Check for Mental Model Congruence and Openness



2.5 Homoscedasticity Check for Mental Model Congruence and Openness

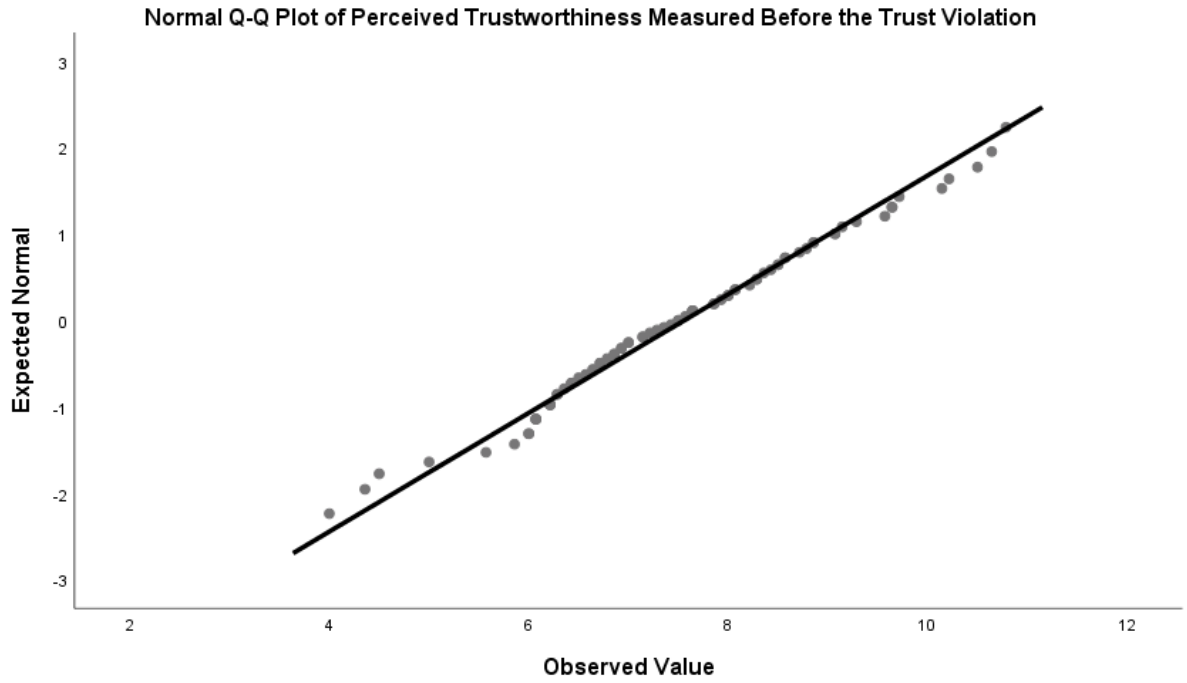


2.6 Multicollinearity Check for Mental Model Congruence and Openness

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
Constant	5,776	1,221		4,731	,000		
Openness	,422	,316	,152	1,337	,185	,993	1,007
Mental Model Congruence	,107	,117	,105	,920	,360	,993	1,007

H3: *The anthropomorphic agent condition yields the highest level of perceived trustworthiness amongst participants, compared to the zoomorphic, mechanical, and virtual agent conditions.*

3.1 Normality Check for Perceived Trustworthiness Before the Trust Violation



3.2 Homoscedasticity Check for Perceived Trustworthiness Before the Trust Violation

		Levene Statistic	df1	df2	Sig.
Perceived Trustworthiness Before the Trust Violation	Based on Mean	1,576	3	74	,202
	Based on Median	1,080	3	74	,363
	Based on Median and with adjusted df	1,080	3	66,120	,364
	Based on trimmed mean	1,570	3	74	,204