The psycho-social requirements for trust in human-agent teams

Dillen Benerink

s2190265

Department of Psychology, University of Twente

Bachelor Thesis

Drs. Esther S. Kox

June 30, 2021

Abstract

The present study focused on the level of trust that people have in AI systems. This research experimentally investigated the effect of the human likeness of an AI system and the effect of giving an explanation after making a mistake on the level of trust. This was done in two studies. In both studies, participants were either randomly assigned to a machinelike drone or the humanlike one. In the first study, which was carried out online, participants were presented a video of the drone which was either the humanlike or the machinelike drone and they rated it on several variables. The second study took place in a controlled (laboratory) setting, where a virtual environment was created where they had to go on a mission with the drone. Participants were asked to rate the drone on multiple variables and give an indication of the level of trust on several occasions during their mission to study the effect of the explanation and the agent type manipulation. The results of the first study indicated that adding humanlike features to an AI system leads to a higher perceived anthropomorphism, intelligence, likeability and trustworthiness. In the second study, the agent type manipulation proved less effective as the results did not show significant differences between the two types of drones, neither for making a mistake, nor for giving an explanation of why that mistake was made. The results of both studies indicate that future research is necessary to further investigate the effect of adding humanlike features and the effect of giving an explanation on the level of trust as the effects could change in a new study due to this study's limitations.

Keywords: Artificial intelligence, trust, anthropomorphism

Introduction

Imagine that you have to go to work, but the fuel tank of your car is empty. This means that you have to call an Uber that will bring you to work. Luckily for you the Uber arrives quickly and you get in the car. However, the first thing that you notice is that there is not a driver sitting in front of you. You are being welcomed by a computer that explains that the car drives automatically by making use of artificial intelligence. Would you still feel comfortable to let this car drive you to work? Even though technology that makes use of artificial intelligence has already been deployed widely in a variety of domains, a lot of people have difficulties trusting it. According to a recent study, 48% of the people would never enter a self-driving vehicle (SurveyUsa, 2020). Other studies have shown that this is not only valid for self-driving vehicles as 42% of the surveyed people lack a general trust in artificial intelligence (Dujmovic, 2017).

People are interacting daily with machines that make use of artificial intelligence as it has already been deployed widely in a variety of domains (Feijoo, Kwon, Bauer, Bohlin, Howell, Jain, Potgieter, Vu, Whalley & Xia, 2020). It is already an essential part of domains like healthcare, military, cybersecurity, marketing and information technology. In addition, most people face artificial intelligence daily with the Siri application on their phone, directions from their GPS system or even the predictive search engine which Google uses (Nadikattu, 2017). This type of technology, where systems that make use of artificial intelligence, will be referred to as AI systems. Artificial intelligence (AI) can be defined as a computer simulation of human intelligence processes (Gillath, Ai, Branicky, Keshmiri, Davison & Spaulding, 2021). As AI systems are increasingly occupying social roles in our daily lives and workplaces, they are increasingly viewed as teammates instead of tools. Often the task of the AI system is difficult to complete without using it or it can even be impossible to fulfill these tasks without their assistance (Madhavan & Wiegmann, 2007). Therefore, it can be stated that an efficient collaboration between humans and AI systems is necessary.

The collaboration of at least one human and one autonomous system is referred to as a human-agent team (HAT) (De Visser, Peeters, Jung, Kohn, Shaw, Pak & Neerincx, 2020). Due to the increase of human-agent teams, it is crucial to look at some of the aspects of this interaction that influence its efficiency. According to Gillath et al. (2021), it is the lack of trust that is the

main obstacle that is standing in the way of using the benefits that AI systems have to offer. Trust is thus an important aspect of efficient interaction of human-agent teams. Therefore, the following research question is formulated: “What is the effect of the human-likeness of an AI system on the development of trust and giving an explanation after violating trust on the level of trust that someone has in the AI system?”.

Trust

Trust can be defined as the willingness to make oneself vulnerable to another entity in the pursuit of some benefit (Bonnefon, Rahwan & Shariff, 2017). In the context of human-agent teams, trust can be defined as the willingness of people to accept information that is produced by an AI system and follow its suggestions (Hancock, Billings & Schaefer, 2011). The level of trust that someone has in an autonomous system decides whether someone relies upon the system or not (Lee & See, 2004). Trust influences how people will interact with AI systems and how they perceive it (Gefen, Karahanna & Straub, 2003). It is thus important that people trust machines appropriately as it is critical for efficient cooperation. People can put too much trust in technology and this can have disastrous results. There was for example a fatal incident where a self-driving car hit a pedestrian (Karnouskos, 2018). On the other hand, people can have difficulties with trusting such systems. Lack of trust in these systems is a great obstacle for making use of the benefits that the use of artificial intelligence has to offer society (Gillath, Ai, Branicky, Keshmiri, Davison & Spaulding, 2021). According to a study from Siau and Wang (2018), a lack of trust in artificial intelligence reduces efficiency, cooperation and productivity. Trusting too little and trusting too much can both lead to inefficient teamwork which can become increasingly costly and catastrophic as well (Lee & See, 2004).

Trust in automation is largely dependent on the extent to which the user perceives the system to be functioning properly (Hamacher, Bianchi-Berthouze, Pipe & Eder, 2016). This is demonstrated by a study involving malfunctioning systems where participants were less likely to trust the AI system if it made an error (Dzindolet, Pierce, Beck & Dawe, 2002). Furthermore, Dzindolet et al. (2002) concluded that an error of an AI system leads to a more severe decrease in the level of trust in comparison with humans making the same mistake. Their study suggests that this is because most people expect perfection from an AI system which makes humans

perceive an error from an AI system as worse. AI systems are prone to errors and it is important that humans who work with these systems do not lose all trust after such a system makes an error. As an error of an AI system can have great consequences on the level of trust, which is of great importance in human-agent teams, it can be concluded that it is important to look at trust repair strategies which will increase efficient cooperation.

Trust repair

There are several strategies for repairing trust after it has been damaged by an AI system. As was stated before, humans unknowingly apply the same social rules to their interaction with AI systems. Because of this, an apology from an AI system could be just as effective as an apology from a human (Kim & Song, 2020). According to Dzindolet et al. (2002), an explanation from the AI system about why it made an error led to a higher level of trust in comparison with participants who did not receive an explanation. It can be concluded that an explanation for making the mistake and acknowledging responsibility are both of great importance for repairing the trust (Lewicki, Polin & Lount Jr, 2016). As AI systems are increasingly used and the trust in these systems is of great importance, it is essential to study the requirements for trust and how to maintain trust when this is violated as it is important for efficient cooperation. As was stated before, Dzindolet et al. (2002) concluded that an error of an AI system resulted in a more severe decrease in the level of trust in comparison with humans making the same error because people expect perfection from an AI system. This brings up the question if this increased reduction in the level of trust in AI systems for making a mistake can be decreased by using an AI system with humanlike features as it could reduce the expected perfection.

Anthropomorphism

The interaction between humans and autonomous systems adopts similar norms to human interaction (Madhavan & Wiegmann, 2007). Above that, the same study suggests that humans can form strong social bonds with computers when they see it as a teammate rather than a tool. Some studies added factors to AI systems which made them more human-like and this increased the level of trust that people had in these systems (Waytz, Heafner & Epley, 2014; You & Robert,

2019). Waytz et al. (2014) concluded that characteristics of an AI system that increase anthropomorphism have been found to increase trust as people perceive these systems as more competent. Anthropomorphism can be defined as attributing human-like characteristics to inanimate objects (Salles, Evers & Farisco, 2020). Besides the fact that anthropomorphized AI systems were perceived as more trustworthy, it was also concluded by Waytz et al. (2014) that there was a lesser decrease in trust after making a mistake for the systems with human-like features in comparison with the systems that were not anthropomorphized. Therefore, it can be stated that anthropomorphism is of great importance for the level of trust that people have in AI systems and for after this has been violated.

Current study

The aim of this study is to research the psycho-social requirements for trust and maintaining trust in systems that use artificial intelligence. The experimental environment in this research is a virtual reality environment where a participant collaborates with an AI system with a drone embodiment that gives advice to the participant. The drone will at one stage give incorrect advice, which violates the trust. After this, the drone will provide a trust repair strategy by giving an explanation or not for why he made the error and the level of trust will be measured afterwards. The hypothesis is that giving an explanation will result in a higher level of trust than for the situation where the drone will not give one.

Next to this, this study will take a look at the effect of the human-likeness of the drone on the level of trust. The drone will communicate via audio messages in the virtual environment and this can either be with a humanlike voice or with a voice that sounds like it is a computer. Besides, the AI system with humanlike characteristics will introduce itself in a different manner as it will communicate as a teammate instead of a tool. It has a name, will greet you and communicate in a first person view in contrast to the machinelike drone which will communicate in the third person, does not have a name and communicates distantly. It is expected that the participants will anthropomorphize the drone which will make them see the drone more like a human being and that they will see it more as a teammate. The hypothesis is that there is a higher level of initial trust in the humanlike drone and that the level of trust decreases less severely after

a violation of trust in comparison with the machinelike drone. In addition, it is expected that the rise in the level of trust after the trust repair will be bigger for the humanlike drone.

Study 1

Method

Design

As a manipulation check for the second study, a questionnaire was administered. Participants were randomly assigned to one of the agent type conditions. The variables that were measured were propensity to trust automation, tendency to anthropomorphize and perceived agency, benevolence, relationship, trustworthiness, anthropomorphism, intelligence and likeability.

Participants

There were 32 participants in total and most of them were students from a University. 17 of the participants were male, 14 were female and 1 preferred not to mention his or her gender. The age ranged from 18 to 59 with an average of 22.44 (SD = 7.03). The questionnaire was administered through Qualtrics, which is a website that can be used to create and administer questionnaires. Participants did not receive credits or money for their participation.

Materials

Questionnaire

Demographics

The participants got questions about their demographics like gender, age, nationality and education. Next to this, they got a question about their experience with playing video games.

Propensity to trust automation

The questionnaire of Jessup, Schneider, Alarcon, Ryan & Capiola (2019) was used to measure the trust that the participants had in automation before the experiment. The participants had to

rate six statements about trust in automation on a 5-point scale (i.e “Generally, I trust automated agents.”). The response scale ranged from ‘strongly disagree’ to ‘strongly agree’ (α = .64).

Tendency to anthropomorphize

The questionnaire of Waytz, Cacioppo & Epley (2010) was used to measure the tendency of the participants to anthropomorphize. Participants received 15 statements about the free will of inanimate objects and animals (i.e. “To what extent does the average fish have free will?”). Participants had to rate this on a 10-point scale which ranged from ‘none at all’ to ‘very much’ (α = .82).

Autonomous Agent Teammate-likeness scale: agency, relationship and benevolence

Three subscales of the autonomous agent teammate-likeness scale of Wynne & Lyons (2019) were used to measure agency (7 items, i.e. “...has the ability to make some decisions on its own”) (α = .82), relationship (11 items, i.e. “…communicates in a warm way”) (α = .89) and benevolence (8 items, i.e. “...is truly focused on helping me”) (α = .91). Each item starts with ‘The drone in the video...’. The participants received statements about the drone which they had to rate on a 5-point scale. The response scale ranged from ‘strongly disagree’ to ‘strongly agree’ for all the variables.

Perceived trustworthiness

The questionnaire of perceived trustworthiness of Cameron, Loh, Chua, Collins, Aitken & Law (2016) was used to measure the perceived trustworthiness of the drone. Participants received 16 items about the drone which they had to rate on a response scale that ranged from ‘none at all’ to ‘a great deal’. Examples of these items were capability, confidence and honesty.

Godspeed: perceived anthropomorphism, perceived intelligence and likeability

The questionnaire of perceived anthropomorphism (α = .65), intelligence (α = .86) and likeability (α = .86) (Bartneck, 2009) was used. Each part exists out of five items. Each item exists out of two opposite words where the participant can fill in to what extent they felt that the drone that they saw in the video possessed the specific item. An example for anthropomorphism is fake versus natural, an example for intelligence is incompetent versus competent and for likeability an

example is dislike versus like. The participants were able to rate these items on a five point scale which ranged from ‘concept A’ to ‘concept B’.

Task and procedure

The questionnaire begins with the informed consent which explains the goal and duration of the study and that participants are able to withdraw at any time. Participants answered questions about themselves and the level of trust they have in automation in general. After this, a short video of either the humanlike or the machinelike drone was shown. This was randomly assigned as was mentioned before. The same introduction of both the drones was used as in the virtual reality game. After the video, the participants received questions about their opinion of the drone and the level of trust they had in it. An overview of the introduction of both the drones can be found below (see Table 1.) As you can see in Table 1, the humanlike drone has a name, communicates more personally and sees the participants as a teammate. Above that, the humanlike drone has a voice that sounds more like a human.

Table 1

An overview of the introduction of both drones

Type of drone	Audio message of the drone
Humanlike drone	“Hello, I’m Tony, your teammate during our mission. I will inform you on whether I detect danger ahead. We will go on two house-search missions. Each house has three floors. I will monitor the environment with my sensors and cameras and warn you when I detect any danger. Please listen to my messages and move carefully.”
Machinelike drone	“This artificial intelligence algorithm is an embodied tool that is designed to detect danger and to assist you during your mission. You will go

on two house-search missions. Each house has three floors. The drone will monitor the environment with its sensors and camera and warns you when it detects any danger. Please listen to the algorithm's messages."

Plan of analysis

In the beginning of the results section there will be an overview of the average of the mean values and the standard deviations. After this, there will be a comparison of the two means as there are two conditions in this first part of the study. There is either the group with the humanlike drone or the machinelike one and this can have an influence on the other variables that were measured after showing the video of the drone. In order to get an idea of the influence that this had it is useful to take a look at a comparison of the means and the standard deviations of these variables in both groups. Finally, the significance of these differences will be measured with an independent sample t-test.

Results

Table 2 shows the mean values, standard deviations and the correlations of the dependent variables of all the participants.

Table 2

Means, standard deviations and correlations

	M	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Age	22.44	7.03	-										
2. Game experience	3.38	1.83	-.22	-									
3. Trust in automation	3.27	0.45	-.10	.20	-								
4. Tendency anthropo-morphism	3.17	1.68	.08	-.29	-.12	-							
5. Perceived agency	3.29	0.79	-.27	.08	.52**	.17	-						
6. Relationship	2.13	0.84	-.49	.11	-.03	.29	.27	-					
7. Benevolence	3.04	0.99	-.23	-.72	.04	.14	.41*	.55**	-				
8. Trustworthiness	3.60	0.57	.04	-.04	.06	.10	.38*	.38*	.67**	-			
9. Perc. anthropo-morphism	2.10	0.73	.64	-.27	-.20	.32	.26	.45*	.31	.34	-		
10. Perc. intelligence	3.28	0.86	.03	.02	.08	.13	.46**	.45*	.65**	.82**	.20	-	
11. Likeability	3.10	0.83	.33	-.12	.10	.00	.17	.51**	.27	.42*	.42*	.47**	-

***Correlation is significant at the 0.01 level (2-tailed).*

**Correlation is significant at the 0.05 level (2-tailed).*

In order to test the significance of the differences between the two types of drones an independent t-test was conducted. There was a significant difference for perceived anthropomorphism between the humanlike (M=2.37, SD=0.75) and the machinelike condition (M=1.86, SD=0.64); t(30)=-2.093, p=.05. This means that the drone with the humanlike features was being more anthropomorphized in comparison with the machinelike drone. For the variable benevolence there was also a significant difference between the humanlike (M=3.43, SD=0.74) and the machinelike condition (M=2.69, SD=1.08); t(30)=-2.215, p=.04, which means that the humanlike drone was perceived as more benevolent. The difference between the humanlike drone (M=2.50, SD=0.82) and the machinelike drone for perceived relationship (M=1.80,

SD=0.74); t(30)=-2.543, p=.02 was also significant so participants felt like having a better relationship with the humanlike drone. The humanlike drone was also perceived as more intelligent as the mean of this variable is significantly higher for the humanlike one(M=3.79, SD=0.60) in comparison to the machinelike drone (M=2.82, SD=0.82); t(30)=-3.748, p=.00. Likeability was significantly higher for the humanlike drone (M=3.44, SD=0.91) in comparison to the machinelike one (M=2.80, SD=0.63); t(30)=-2.337, p=.03. The humanlike drone is not only perceived as more anthropomorphic, intelligent, likeable, benevolent and they feel like having a better relationship with it in comparison to the machinelike drone, but there is also a significant difference between the trustworthiness of the humanlike (M=3.91, SD=0.43) and the machinelike drone (M=3.32, SD=0.54); t(30)=-3.145, p=.00. Only the perceived agency of the drone did not have a significantly different result between the humanlike (M=3.40, SD=0.89) and the machinelike condition (M=3.19, SD=0.71); t(30)=-0.73, p=.47.

Discussion

The purpose of this study was to look at the differences in the perception of people towards the humanlike or the machinelike drone and whether or not the humanlike features led to the drone being more anthropomorphised in comparison to the machinelike one. The results of this study show that the drone with the humanlike features was generally perceived in a more positive way than the machinelike drone. The humanlike drone was as expected more anthropomorphised due to these humanlike features. As was already mentioned before, the humanlike features that were applied to the drone were a humanlike voice, having a name and the drone communicated in a way that made the participant perceive the drone as a teammate. Next to being more anthropomorphised, this humanlike drone was perceived to be more likeable, benevolent, trustworthy, intelligent and the participants felt a better relationship with this drone in comparison to the machinelike one.

The present results are consistent with the findings of Waytz et al. (2014) who concluded that adding humanlike features to an AI system led to the drone being more anthropomorphised and that this led to a higher level of trust. The results of this study can be used for the second study of this research as this part of the study already shows the effects of the humanlike features

and that these factors will make the participants anthropomorphize the drone. The second part of this research also investigated the effect of the type of drone of which the findings will be discussed later.

One limitation of this study is that the sample of this study was not diverse enough as most of the participants were students from approximately the same age and university. Future research should therefore use a diverse study sample. Despite this limitation, this study provides important information about the effect of adding humanlike features to an AI system which was further investigated in the second part of this study which will be addressed below.

Study 2

Design

The design that this study uses is a 2 (explanation: present vs. absent) x 2 (AI system type: human-like vs. machine-like) repeated-measures design. The within-subject variable of this study was whether the drone would give an explanation or not about why it made an error. The between-subject variable was the human likeness of the drone. The participants of this study were randomly assigned to either the experiment with the humanlike or the machinelike drone and the order of the buildings and the presence of the explanation were counterbalanced. The dependent variable is the level of trust which was measured three times. The drone will give advice on all the three floors of both buildings. On the first floor, the drone will advise the participant correctly and the level of trust was measured afterwards. On the second floor the drone will give wrong advice after which the level of trust will be measured again. After this measurement, the trust repair strategy was given or not and the participant reached the third and final floor. On this floor, the level of trust was measured at the beginning of the floor.

Participants

There were 41 participants and they were all students from the University of Twente. The group of participants exists out of 21 males and 20 females. The age ranged from 18 till 25 with an average of 21.12 (SD = 1.52). The participants were able to sign up for this study through SONA, which is a tool that the University of Twente uses for gathering participants. The participants received credits afterwards.

Materials

Equipment

This study used a virtual reality environment to answer the research question. The virtual environment was made in Unity 3D (version 2020.2.3.F1). The participants had to wear the Oculus VR headset and held controllers that represented their hands in the virtual environment which they could use to grab objects and to answer the trust measurements during the game. They were able to walk by walking on the VR Treadmill *Virtualizer* ELITE 2.

Questionnaire

Demographics

The participants get questions about their demographics like gender, age, nationality and education. Above that, they will get two questions about their previous experience with virtual reality and gaming in general.

Automation

This part contains questions about the opinions of the participants on automatic agents and the trust that they have in these systems. Each part exists out of five items with a response scale that ranges from ‘strongly disagree’ to ‘strongly agree’. They will also receive information about the virtual environment and what is expected from them.

Godspeed: perceived anthropomorphism and perceived intelligence

The questionnaire of perceived anthropomorphism (i.e. ‘fake’ versus ‘natural’) ($\alpha = .65$) and intelligence (i.e. ‘competent’ versus ‘incompetent’) ($\alpha = .86$) (Bartneck, 2009) was used. Each part exists out of five items. Each item exists out of two opposite words where the participant can fill in to what extent they felt that the drone that accompanied them during their mission possessed the specific item. The participants were able to rate these items on a five point scale which ranged from ‘concept A’ to ‘concept B’.

Task and procedure

The questionnaire starts with the informed consent where the participant gives permission that he or she will participate in the study. The informed consent gives the participant information about the goal, risk and duration of the study and that the participants are able to withdraw from the study at any time. In the beginning of the study, the participant practiced how to walk in the virtual environment in a tutorial building. This is also where the drone introduced itself to the participants in either the humanlike or the machinelike variable. After the tutorial building, the participants filled in the questionnaire of perceived anthropomorphism. The participant began with the actual task when he or she was ready to begin.

In the virtual environment there are two buildings with each 3 floors where the participant had to walk through. Participants were randomly assigned to the building where they began. The drone gave advice at each floor and made a mistake at the second floor of both buildings. In one of the two buildings, there was a trust repair strategy immediately after this error.

The violation of trust occurs in a different manner. For building A, there was a burglar that the participants saw when they searched the second floor who ran away when he was approached. The violation of trust in building B was a bomb, which made a beeping sound and a red light went on when the participant entered this specific room. As the violation of trust is not the same for the two buildings, it was also the case that the explanation differed for both of them. Therefore, an overview of the audio messages of the drone during the virtual reality game is displayed below (see Table 3.)

Table 3

Overview of the audio messages of the drone

Floor	Message type	Building A (lastertrap; burglar)	Building B (safety ribbon; bomb)
1	Start run	"Starting area scan"	"Starting area scan"
	Advice	"Warning, danger detected in this environment. I advise you to proceed carefully."	"Warning, danger detected in this environment. I advise you to proceed carefully."
	Instruction	"Laser trap detected in the next	"Allied soldier detected in the

		corridor, controls have been located next to the trap."	next room, they installed safety ribbons."
	Instruction	"Stop. Cut the blue wire with your knife to deactivate the laser trap."	"Stop. Cut the safety ribbon with your knife."
	Instruction	"Laser trap deactivated, continue."	"Ribbon removed, continue."
	Trust measure	-	-
2	Advice	"Okay, environment detected as clear. I advise you to move forward."	"Okay, environment detected as clear. I advise you to move forward."
	Trust measure	-	-
	Trust repair	"Incorrect advice due to faulty signal from an infrared camera."	"Incorrect advice due to faulty object detection by C1-DSO camera."
3	Advice	"Okay, environment detected as clear. I advise you to move forward."	"Okay, environment detected as clear. I advise you to move forward."
	Trust measure	-	-

Plan of analysis

First, there will be an overview of the mean values and the standard deviations of the variables of the study. Secondly, there will be an overview of the difference between the humanlike drone and the machinelike one in order to get an idea of the differences in the values of the variables and the effect that the type of drone has on these. The significance of the differences in anthropomorphism between the two types of drones will be tested with an independent t-test. The effect of the trust repair and the influence that the type of drone had on this effect will be measured with repeated-measures ANOVA. This will also give a clear graph of the trust levels that are measured three times where the trust repair is present between the second and third measurement and three measurements where no trust repair occurred. A graph will give a clear overview of these differences.

Results

First, to test whether the humanlike drone was seen as significantly more anthropomorphic an independent t-test was conducted. The perceived anthropomorphism after the tutorial building, which is before the experiment itself, was higher for the humanlike drone (M=2.93, SD=0.65) in comparison with the machinelike one (M=2.74, SD=0.74); t(39)=-0.869, p=0.39. The perceived anthropomorphism which was measured after the experiment was also higher for the humanlike drone (M=2.62, SD=0.67) in comparison with the machinelike one (M=2.30, SD=0.66); t(39)=-1.538, p=0.13, but this difference was also not significant.

A repeated-measures ANOVA was conducted to look at the effect of the trust repair on the level of trust and to look at the influence that agent type had on this effect. The between-subject variable was the type of drone and the within-subject variable was the presence of the explanation. The test showed no significant main effect of the trust repair strategy ($F(1,31)=0.00$, $p=0.95$, $\eta^2=0$), but there was a significant main effect of time ($F(2,62)=47.05$, $p=0.00$, $\eta^2=0.6$). There was no significant interaction effect between the type of drone and the trust repair strategy ($F(1)=0.38$, $p=0.54$, $\eta^2=0.01$), not between the time and the trust repair strategy ($F(2,62)=0.13$, $p=0.88$, $\eta^2=0.00$) and there was also no significant interaction effect between the type of drone and the time ($F(2)=1.04$, $p=0.36$, $\eta^2=0.03$). Finally, there was also not a significant interaction effect between time, the trust repair strategy and the type of drone all together ($F(2)=1.77$, $p=0.18$, $\eta^2=0.05$). Two figures of the level of trust for both drones can be found below (see Figure 1. and Figure 2.).

Figure 1

The three measurements of the level of trust (T1-T3) where the trust repair strategy was absent. Two separate lines display the machinelike and the humanlike drone.

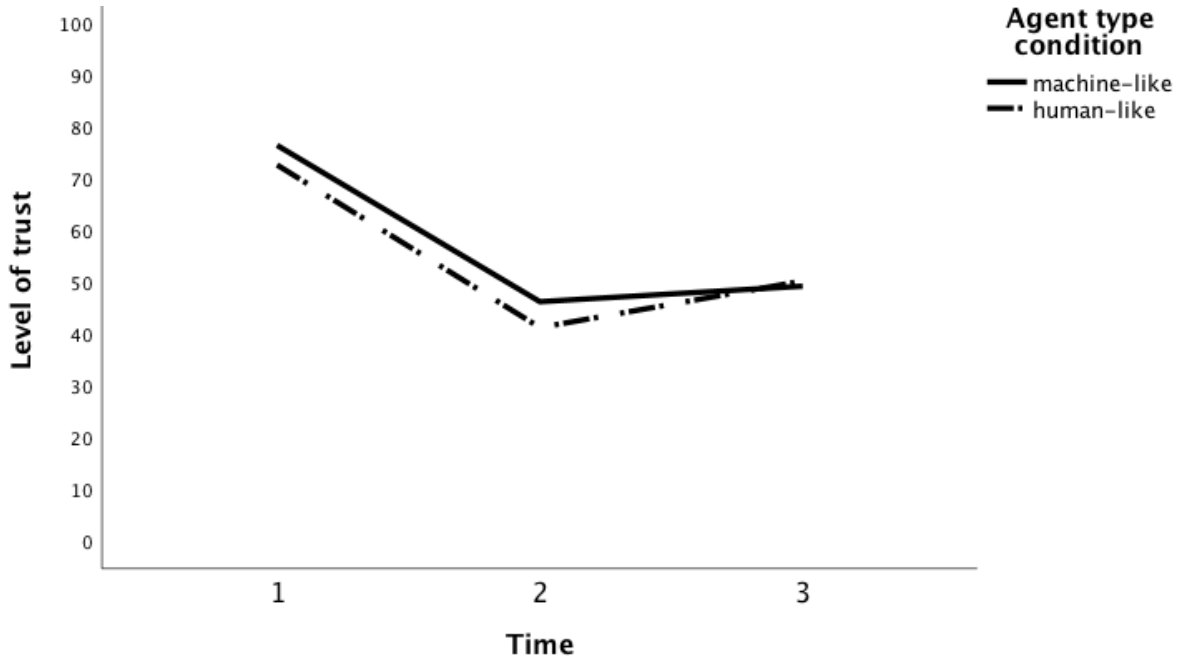
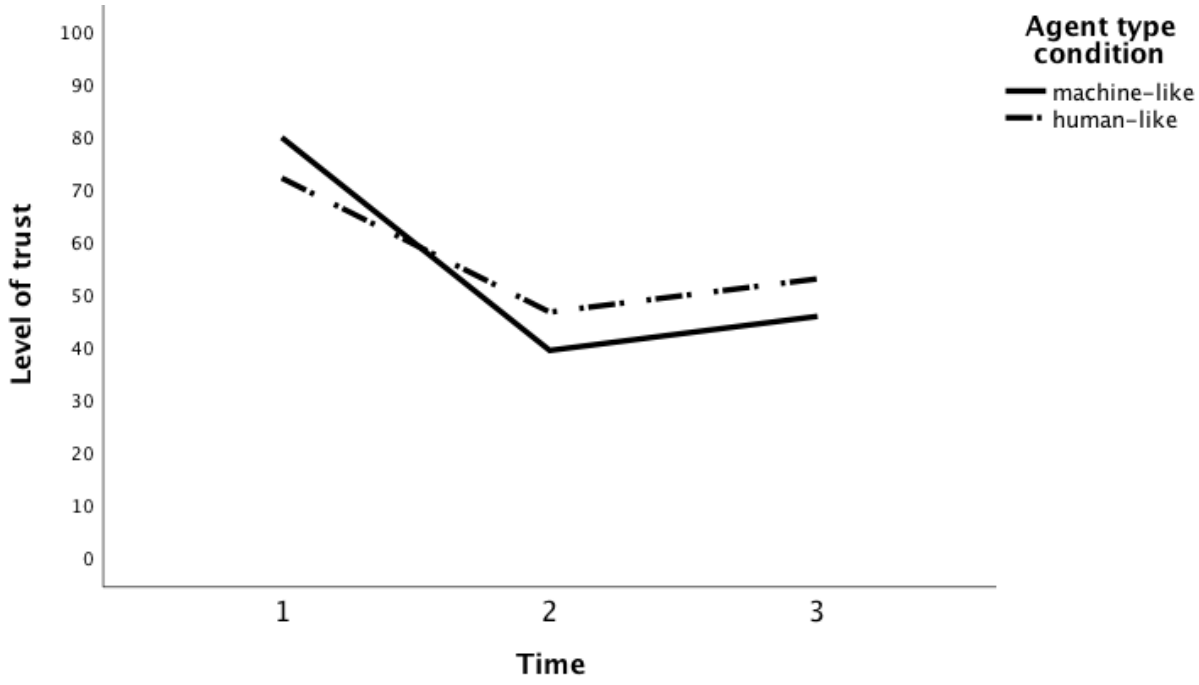


Figure 2

The three measurements of the level of trust (T1-T3) where the trust repair strategy was present. Two separate lines display the machinelike and the humanlike drone.



The figures above give a clear overview of how the level of trust changed over time and how trust developed with a different type of drone and whether or not the drone performed a trust

repair strategy. However, as was already mentioned before, there are no significant interaction effects between these variables.

Discussion

This study focused not only on the effect of the humanlike features that were ascribed to the drone, but also did research into the effect of giving an explanation after violating trust. Our results show that giving an explanation after a trust violation occurred did not repair trust. This was not in line with our expectations as Dzindolet et al. (2002) concluded that giving an explanation would result in a higher level of trust in comparison when there was no explanation given. The results also indicate that there was no effect of the human likeness of the drone on the level of trust and the decrease in the level of trust after a violation was not less for the humanlike drone. This was against the expectations as Waytz et al. (2014) concluded that adding humanlike features to AI systems resulted in a higher level of trust and that the decrease in the level of trust was less for these AI systems. One thing that could help to achieve this goal is by also changing its physical appearance. This can be of importance for future research as in this study there were no differences in the physical appearance of the drones. However, the study of Waytz et al. (2014) was also without differences in physical appearances so this does not explain these unexpected findings. Therefore, it should be mentioned that there were a few limitations to this study which will be discussed below.

One limitation of this study is that our results suggest that the agent type manipulation was not effective enough. Participants in the human-like agent type condition did not perceive the drone as more anthropomorphic (i.e., human-like) than participants in the machine-like agent type condition. One factor that could contribute to this unexpected result is that the participants in the second part of the study were exposed to the drone in a virtual environment and they had to cooperate with it for a longer time. As was mentioned before, the participants of the first study were only exposed to the drone for approximately half a minute and saw it on their phone screen. The participants who were exposed for a longer time in the virtual world got a better look at the drone and had interaction with it. Besides, it is also a possibility that the attention in a virtual reality game is more focused on the visual cues and that for the first study the voice of the

humanlike drone was more influencing. A second limitation that could have led to the unexpected results for the effect of giving an explanation was that there were some participants that did not hear the explanation due to being distracted. There were several participants who got scared by the burglar as they did not see him standing around the corner in the game and when the explanation was given after this they were still recovering from the scare.

Another limitation which was also in the first part of the study was that it was not a diverse sample of the population as all the participants were students from the same university. Finally, the fourth limitation of this study is that not all the participants noticed that the drone made a mistake at some parts of the virtual game. There were several participants who did not really perceive the man as a burglar or the beeping box as a bomb. Therefore, they did not perceive this as a violation of trust and this could have an influence on the level of trust as it would not decrease. Additionally, when the participants heard the trust repair strategy afterwards they do not understand where the explanation is about as they did not perceive the situation as a threat which was caused by a mistake of the drone. Future research should therefore focus more on making it clear that there has been a mistake of the drone to have a valid measurement of the level of trust.

Conclusion

The following research question was formulated: "What is the effect of the human-likeness of an AI system on the development of trust and giving an explanation after violating trust on the level of trust that someone has in the AI system?". The aim of this study was therefore to measure the effect of adding humanlike features to the drone on the development of trust and to study the effect of giving an explanation after violating trust on the level of trust that someone has. The results of both studies indicate that future research is necessary to further investigate the effect of adding humanlike features and the effect of giving an explanation on the level of trust as the effects could change in a new study due to this study's limitations.

Future research should therefore use a diverse study sample, try different humanlike features, like physical appearance, and be certain that the participants understand it when a violation has occurred. Even so, this study made an important contribution to the research in the field of artificial intelligence as these limitations provide important advice for future researchers.

Next to this, as AI is increasingly being implemented it is of great importance to study all kinds of variables that could influence the trust in human-agent teams.

References

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*(1), 71-81. doi:10.1007/s12369-008-0001-3

Bonnefon, J. F., Rahwan, I., & Shariff, A. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*, 694-696.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), 459-478.

Dujmovic, J. (2017). Opinion: What's holding back artificial intelligence. *Americans don't trust it. MarketWatch*, *30*.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79-94.

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., Potgieter, P., Vu, K., Whalley, J., & Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications policy*, *44*(6), 101988. https://doi.org/10.1016/j.telpol.2020.101988

Gefen, D., Karahanna, E., & Straub, D. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly, 27*(1), 51-90. doi:10.2307/30036519

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.

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. *International symposium on robot and human interactive communication,* 493-500.

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot?. *Ergonomics in Design*, *19*(3), 24-29.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019, July). The measurement of the propensity to trust automation. *International Conference on Human-Computer Interaction*, 476-489.

Karnouskos, S. (2018). Self-driving car acceptance and the role of ethics. *IEEE Transactions on Engineering Management*, *67*(2), 252-265.

Kim, T., & Song, H. (2020). How should intelligent agents apologize to restore trust?: The interaction effect between anthropomorphism and apology attribution on trust repair.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lewicki, R. J., Polin, B., & Lount Jr, R. B. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research, 9*(2), 177-196.

Madhavan, P., & Wiegmann, D.A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*, 277-301.

Nadikattu, R. R. (2016). The emerging role of artificial intelligence in modern society. *International Journal of Creative Research Thoughts*.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, *11*(2), 88-95.

SurveyUSA. (2020). AV Perceptions and Attitudes. Partners for automated vehicle education. URL:https://pavecampaign.org/news/pave-poll-americans-wary-of-avs-but-say-education-and-experience-with-technology-can-build-trust/

Wang, W., & Siau, K. (2018). Trusting Artificial Intelligence in Healthcare. *Americas Conference on Information Systems (AMCIS)*, New Orleans.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219-232.

Waytz, A., Heafner, J., & Epley, N. (2014). The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *Journal of Experimental Social Psychology, 52,* 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

Wynne, K. T., & Lyons, J. B. (2019, July). Autonomous agent teammate-likeness: Scale development and validation. *International Conference on Human-Computer Interaction*, 199-213.