4-25-2021

Factors Influencing Trust in Automation

Hannah Marie Smitherman

Factors Influencing Trust in Automation

by

Hannah Marie Smitherman

An Honors Capstone

submitted in partial fulfillment of the requirements

for the Honors Diploma

to

The Honors College

of

The University of Alabama in Huntsville

4/23/2021

Honors Capstone Director: Dr. Nathan Tenhundfeld

Assistant Professor of Psychology

Hannah Smitherman 4/23/2021
Student Date

Nathan L Tenhundfeld 4/25/2021
Director Date

Jodi L Price Digitally signed by Jodi L Price Date: 2021.04.26 09:31:59 -05'00'
Department Chair Date

William Wilkerson Digitally signed by William Wilkerson Date: 2021.04.26 13:06:12 -05'00'
Honors College Dean Date

Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

Honors Thesis Copyright Permission

This form must be signed by the student and submitted as a bound part of the thesis.

In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

Hannah Smitherman

Student Name (printed)

Hannah Smitherman

Student Signature

4/23/2021

Date

Table of Contents

Abstract

This paper outlines the current findings regarding factors affecting users' trust in automation. This is broken into categories reflecting trust in different types of automation. The included topics in this paper are trust in decision aids, trust in autonomous vehicles, trust in robots, models of trust, and other topics. Despite focusing on different types of automation, many of the findings of these studies are similar. For example, more transparency in decision making (regardless of type of automation) tends to lead to increase trust. Many of the studies discussed here introduce new factors influencing trust (such as pedigree or animation), while many others study established factors (such as transparency or anthropomorphism).

Factors Influencing Trust in Decision Aids

Aoki (2021) studied how assuring users that they are part of the decision loop can affect initial trust in a decision aid. To explore this, a survey was created where participants were asked about their trust in an automated care plan manager. Participants in this study were either long term care users, caregivers of long-term care users, or family members of long-term care users. The survey started by briefing participants on the use of an automated care manager in long term care situations. The participants then rated their initial trust in the system. Following this, participants were told how the care manager would be used. At the end of this statement, participants were either told nothing or were told that a human care manager will create a plan based on the AI's suggestion. Next, participants were either told nothing or that results of using the AI could cause either a reduction in care managers' workloads, streamlined care planning, care planning based on scientific analysis, an improvement in care users' independence, or a reduction in nursing care and medical expenses. Participants then rated their trust in the system and described whether they would prefer an AI's or a human's care plan. Aoki found that assuring that humans would still be included in the decision loop was associated with slightly higher trust. Users receiving this assurance were more likely than those without to trust an AI care plan more than a human's.

Brauner et al. (2019) examined the effects that errors in decision support systems (DSSs) have on users. To study this, they had participants play a 'Quality Intelligence' business simulation game that lasted for 18 in-game months, where the participants had to understand the state of production and control the investments in procurement, inspection of goods, and inspection of production quality. Participants would be equipped with a DSS that would recommend the amount of supplies to order; they would, in random order, experience a DSS that

performed at near optimal levels, and a DSS that performed well for 6 months and then made suggestions that were 50% below the correct recommendation. Participants throughout the study responded to measures regarding demographics, perceptions and feelings towards automation and technology, and feelings towards the DSS. Compliance with the DSS and order changes were also measured. Among other findings, they found that correctness of the system did influence trust. However, this effect was not shown until the game's second round. So, correctness may affect trust, but it is possible that it takes time to realize a system's defects.

Chavaillaz et al. (2019) studied the differences between experts' and novices' performance, operation, and trust when completing an X-ray baggage screening task alongside a diagnostic aid. Participants were categorized as either Novices or Experts, depending on their experience and skill with X-ray screening. The participants were then assigned to either the Diagnostic Aid group or the No Diagnostic Aid group. In the Diagnostic Aid group, participants freely chose between 3 different automation levels. During the task, participants searched X-ray images of luggage for guns and knives. The aid was set at a reliability of 75%; the system would always report if there was a target, but it would sometimes have false positives or cue a non-targeted item when a weapon was present. Participants completed 64 trials, half of which contained a weapon. Several dependent variables were measured regarding performance, use of the aid, participants' subjective states, and trust. They found that novices performed better with automation than without, but still not as well as experts, whose performance was unaffected by the aid. Experts also had higher compliance, reliance, and self-confidence and lower fatigue, frustration, mental load, and time pressure than novices. For novices, trust was found to be correlated with compliance. They suggested that experts are better at estimating diagnostic aid performance and thus more likely to follow correct recommendations than novices.

Jensen et al. (2019) examined the effects of reliability and blame on trust in automated aids. During the study, participants were asked to complete a target identification task where they were tasked with labelling 20 images as either dangerous or not dangerous. This task was explained as searching for dangerous or criminal activity in an area. During this task, the participants would interact with an aid with a set level of Reliability and a set Blame mode. Reliability could either be High or Low and Blame could either be Internal ("I am sorry that X images assigned to me were counted as misidentifications. I was not able to process those images."), Pseudo-External ("I am sorry that X images assigned to me were counted as misidentifications. The developers were not able to account for processing those images."), or External ("I am sorry that X images assigned to me were counted as misidentifications. A third-party algorithm that I used was not able to process those images."). As the task went on, participants were able to choose how many images were left to the automated aid; they were incentivized to trust the automation naturally, as the scores, of which high scores could receive a bonus, for the task included both the automation and the participants' performances. They found that High Reliability systems elicited greater behavioral and subjective trust. They also found that systems that frame errors as the developers' responsibility elicited less trust. Their findings indicate that Blame may be more important to user perceptions than Reliability.

Keller & Rice (2009) examined whether multiple aids will be trusted as one system or trusted as separate systems. To examine this, participants were asked to complete a simulated flight pursuit task where they used a joystick to pursue a moving aircraft. At the bottom of the screen were two gauges which the participants were asked to monitor. After the needles moved into a designated unsafe range, a system failure was said to have occurred. Participants were asked to identify these failures. The manipulations included Gauge (Gauge 1, where each gauge

had an aid that might be imperfect, and Gauge 2, where each gauge had an aid that was always perfect) and Automation Reliability (High Reliability, where both gauges' aids were 100% reliable, Moderate Reliability, where one gauge's aid was 85% reliable and the other 100%, and Low Reliability, where one gauge's aid was 70% reliable and the other 100%. They found that, when paired together, decision aids are trusted as one system rather than multiple. This means that when one system performs perfectly and another is imperfect, they will both be trusted as much as the imperfect system if they have been paired together.

Pan & Steed (2016) studied people's trust in expertise between avatar, video, and robot agents. In their study, the agent served as an advisor to participants in a task where participants could win chocolates based on their performance answering difficult general knowledge questions. Participants would have two different advisors (either robot, avatar, or video) with two different levels of expertise (non-expert or expert). Only one advisor could be asked for advice per question. Following the experiment, participants were asked to respond to several measures regarding trust, as well as an open-ended question where they were asked to describe which advisor they tried to rely on. Performance and behavioral measures had also been recorded during the experiment. They found that participants could discriminate between experts and non-experts and thus were more likely to choose expert advice in all scenarios. They did find a bias to trust video agents over avatars, possibly due to the presence of nonverbal communication cues. The avatar was preferred less than the video and robot agents, possibly due to seeming synthetic or unreal. The robot agent was preferred over the avatar, but a similar relationship was not found between the robot and the video agent.

Pearson, Geden, and Mayhorn (2019) examined effects of Advisor (Human or Automation) source type bias and Pedigree (Low or High) in a decision-making task where a

Human and Automated Advisor advised participants in choosing the safest route for a military convoy. Pedigree levels were established in a previous study. Participants completed the task 8 times; in trials 1-3 and 5-7 the Advisors recommended the same route and in trials 4 and 8 they recommended different routes. In a trial, participants were first shown a Human and an Automation Advisor profile of random pedigrees in random order. The profiles were read aloud with either a male human voice (Human) or a male text to speech generator (Automation). Participants were then shown a map with three possible routes and all obstacles labeled. While the participant looked at the map, both decision aids gave a recommendation via text and audio. Participants then made their choices. They found that for Low Pedigree, Automated Advisors had higher perceived pedigree, and that for High Pedigree, Human Advisors had higher perceived pedigree. Significant differences in trust for different Pedigree groups were found. When Pedigree levels between Advisors were different, the Higher Pedigree Advisor was more trusted. When both advisors had High Pedigree, the Human was more trusted.

Pearson et al. (2016) studied the ways that perceived risk and workload can affect trust between human and automated advisors. In this study, participants were asked to choose the safest route for a military convoy given differing advice from human and automated advisors. Presentation of information differed between participants, with some being presented with both information sources at the same time, some with the automated tool presented first, and some with the human source presented first. The map they were shown illustrated past improvised explosive device locations and known enemy territory, which were to be considered when choosing the safest of the three possible routes. Following this task, participants were asked to respond to items regarding trust, workload, and perceived risk. Findings of this study indicate

that people tend to trust other humans more in situations of higher perceived risk. Findings also indicated that people tend to trust automation less when workload is perceived to be higher.

Rovira, Pak, & McLaughlin (2017). studied how differences in working memory impact users' trust in automated systems. To explore this, participants were first asked to complete a working memory task and then completed a task where they had to identify the most dangerous enemy target and select a friendly unit to join in combat, given several detailed criteria. A secondary task to increase workload was also included; participants had to respond by clicking on a button every time they were "contacted". Task Load was either Low (3 friendly, 3 enemy units) or High (6 friendly, 6 enemy units) and Degree of Automation was either Manual (no automation), Information (a list of all possible engagement combinations in alphabetical order), Low-Decision (a list of all possible engagement combinations listed from best to worst), or Medium-Decision (a list the 3 best possible engagement combinations listed from best to worst). Participants experienced each Task Load and each Degree of Automation in counterbalanced order. They found, with low levels of automation, performance is better for individuals with higher working memory capacity. They also found that working memory was significantly negatively correlated with trust; thus, users with lower working memory would likely have higher trust in the system.

Schmidt, Biessmann, & Teubner (2020) discussed the effects that AI transparency can have on trust. To study this, they asked participants to complete a classification task where they must label 50 movie reviews as either positive or negative; negative and positive reviews were split evenly. During this task, participants were given a decision aid which would recommend which category reviews belonged in. The decision aid was set to perform at 80%, about the capability of humans. Each participant experienced an aid that either had Highlights or No

Highlights and that either had a Confidence Score or No Confidence Score. For Highlights, the three most relevant words in the review would be highlighted by the aid. For Confidence Score, a percentage was given by the aid, representing how confident it was in a choice. They found that trust is affected by the system's correctness. They also found that transparency can negatively affect trust.

Stokes et al. (2010) studied the effects of mood on trust in an automated decision task consisting of choosing the safest route for a convoy. They induced positive or negative moods randomly in participants using the The International Affect Picture System. The Positive and Negative Affect Schedule was used to measure success of the induction; induction was successful for positive affect and marginally successful for negative affect. Participants then completed the convoy task. In this task, 3 sources of information were provided: a map displaying hostile areas, sensor data on the thermal, audio, and magnetic activity, and an automated-decision aid that used real time data to display newly projected hostile areas on the map and suggested a route at the end of the trial. Participants completed this trial 9 times. Of the trials, 3 had low vulnerability (high agreement between map display, sensor data, and automated aid) and 6 with high vulnerability (low agreement between map display, sensor data, and automated aid). They found that mood affected the first experimental trial, where participants were more trusting of the automated decision aid if they had been induced for positive affect. After this, mood no longer seemed to affect trust. This implies that mood is important for developing first impressions of trust in automation, but that mood will not alter trust in automation over extended periods of time.

Yuksel, Collisson, and Czerwinski (2017) sought to understand the effects of agent reliability and attractiveness on trust. Agents were manipulated to be either highly attractive or

neutrally attractive (no unattractive agents were used because it was deemed unlikely that designers would intentionally create an unattractive agent) and had either high (83%) reliability or low (50%) reliability. All participants reported being heterosexual and were thus paired with agents of the opposite gender. For the task, participants were given timed general knowledge multiple choice questions. These were designed to be difficult enough that participants may need help with the answers. They were allowed to use a search engine to try to find out answers. Incorrect answers resulted in losing points and would be repeated until they got the right answer or timed out. Timing out resulted in losing a point, and thus less possible gratuity. During some questions, the agent would appear, and the participant would be able to see their recommended answer. They found that users believed highly attractive agents to be more trustworthy and more accurate, even over more reliable but less attractive agents. This implies that agent attractiveness is just as important, maybe even more important, than agent reliability is for trust.

In Experiment 1, Zhang, Liao, and Bellamy (2020) explored Trust Calibration, through use of Confidence Information, as a method of increasing trust in automated decision aids. For this study, participants were equipped with an AI system that provided different information based upon the experimental conditions and were asked to predict whether a person's annual income would exceed $50k, given demographic and job information. Manipulations included Confidence Information (Show, where the system displayed a message regarding its confidence, or Not Show, where it displayed no such messages), AI Prediction (Show, where the system displayed its prediction, or Not Show, where it displayed no such prediction), and Model (Full, where participants and system had access to all information, or Partial, where participants had access to more information). They found that Showing Confidence Information improved trust calibration regardless of if AI Prediction was Shown and led to higher trust when Showing high

Confidence. Trust Calibration did not affect decision performance. Model also seemed to have no effects.

In Experiment 2, Zhang, Liao, and Bellamy (2020) explored Trust Calibration, through use of prediction specific Explanations, as a method of increasing trust in automated decision aids. This time, AI Prediction was always Shown, and the Full Model was always used. The only manipulation was Explanation (Present or Absent), where Present Explanation consisted of a graph the system would present to show likelihood of an income above $50k based on each demographic category, as well as a base chance. They found that Explanation did not seem to affect participant trust. Decision performance seemed to decrease when an explanation was present.

Factors Influencing Trust in Autonomous Vehicles

Ajenaghughrure, da Costa Sousa, and Lamas (2020) examined how risk level (no, low, high, very high) affects a driver's trust in an automated vehicle using a simulated driving experience. Trust was measured by the number of times participants used the joystick to take control, as well as by electrodermal activity signals, and a human computer trust model questionnaire. Participants were told that the experiment was a game, were unaware of the risk levels, and were told to earn 75 points (that would be subtracted from for crashes or mistrust) to earn a gift card. They found that trust in the vehicle significantly varied before and after interacting with it; trust decreased after interaction, though not significantly. Trust does significantly vary between risk conditions. Although insignificant, trust was slightly lower in very high risk than in high risk situations, slightly lower in low risk than in no risk situations, and much lower in very high and high risk than in low and no risk situations. Users took control more frequently in the very high, high, and low risk categories than in the no risk category,

suggesting that risk is perceived as present or absent. There was a significant correlation between psychophysiological response and varying trust.

Antrobus, Burnett, and Large (2018) examined differences in trust between a traditional graphical user interface (GUI) and natural language interface (NLI). The GUI was a mainly visual interface, which users could interact with using short vocal cues such as “yes” or “no”. The NLI was a mainly language-based interface, which guided drivers though scenarios by providing news updates, comfort interactions, email and calendar organization, and assistance with driving-related issues. For this experiment, participants were asked to ride in a simulated highly automated vehicle, described as “the taxi of the future”. Participants completed 3 drives with the system, after each of which they completed items regarding trust in automation and system acceptance as well as an interview. They found that the NLI was seen as more useful, and capable, and satisfying. However, there were no significant differences found in trust between the two interfaces.

Forster, Naujoks, and Neukum (2017) studied the effects of speech on trust in an automated car. In this study, participants rode in a (simulated) automated vehicle on a course which lasted about 10 minutes and would take them through 4 different scenarios which would require driver takeover. In the Generic condition, participants rode in a vehicle that would play unspecific warning tones during these takeover scenarios. In the Speech condition, participants received the warning tones form the General condition as well as semantic speech output. During the drive, participants were asked to complete a distractor task of reading selected articles from a news magazine. Measures regarding trust, anthropomorphism, usability, and acceptance were collected. They found that the Speech condition vehicle was seen as more trusted, anthropomorphic, usable, and acceptable.

Gold et al. (2015) examined the effects that take-overs have on driver trust in automated vehicles. To study this, they had participants ride in a simulated highly automated vehicle for around 15 - 20 mins. In this simulation, there were 3 take-over scenarios where a stranded vehicle blocked the lane. Before and after the simulation, participants were asked to respond to items regarding attitudes towards highly automated driving. During the simulation, eye tracking data was gathered. They found that older people were more trusting of the vehicle and had higher perceptions of safety and intention to use the vehicle. Experience with the vehicle decreased feelings of safety and driver benefits from using the automation, but it did cause increased trust.

Ha et al. (2020) explored the relationships between an autonomous vehicle's explanation types, perceived situational risk, and trust in autonomous vehicles. Participants were asked to participate in a driving simulation where Situational Risk (clear day & slow speed, clear day & fast speed, snowy night & slow speed, snowy night & fast speed) and Explanation Type (no explanations, simple explanations, attributional explanations) were manipulated. Participants experienced the simulation for 10 minutes, with 1 manipulation of Situational Risk and 1 manipulation of Explanation Type. Ha et al. found that Explanation Type had a significant effect on trust, and that the effect was moderated by the perceived Situational Risk. This relationship was such that high levels of perceived Situational Risk, attributional Explanation Type had the lowest trust ratings and no Explanation Type had the highest trust ratings. This was reversed for low levels of perceived Situational Risk; for low levels, no Explanation Type had the lowest trust ratings and attributional Explanation Type had the highest trust ratings. This implies that, as risk increases, more information is desired in order to trust an automated vehicle.

Kunze et al. (2019) studied the effects of communicating uncertainty on trust in an automated vehicle. In this study, participants were asked to ride for approximately 20 mins in a

simulated automated vehicle. Participants either rode in a vehicle with an uncertainty display, made using a stylized heart rate monitor to represent stress and uncertainty in a situation, or in a vehicle with no uncertainty display. All participants rode in the vehicle through four different levels of fog density. For vehicles with an uncertainty display, the system heart rate changed linearly with the visibility range; in other words, as visibility decreased, heart rate increased. During the ride, participants would be asked to take over control of the vehicle once. Participants were also asked to complete a distractor visual search task during the ride. Their findings indicate that the heart rate monitor led to more appropriate trust calibration and that the monitor providing information in increments allowed for better knowledge of system capabilities. They also verified that users pay less attention to a system the more they trust it.

Lee et al. (2020) studied differences in trust of automated vehicles between different demographic characteristics. To examine this, participants rode in a simulated automated vehicle. The drive took place on a highway for all 4 trails. For trials 1, 2, and 4, the vehicle performed perfectly and handled all traffic on its own. During trial 3, the vehicle asked the driver, short notice, to intervene and take control. After completing the trials, participants were asked to fill out questionnaires regarding demographic information and trust in automation. They were also asked to participate in a brief interview regarding their impressions of automation. Men and women were found to have no significant differences in age, driving experience, or driving mileage per week. Women drivers were found to be more trusting than men. Driving frequency was not found to affect trust. Student drivers were found to have higher levels of distrust and non-student drivers more frequently gave higher trust ratings. Trust was also measured with regard to purpose (why the system was designed), process (how the system functions), and performance (how the system is operating); trust was found to be moderate across these

measures. However, for all demographics, ratings for the purpose dimension were significantly lower than for the process and performance dimensions.

Lee et al. (2016) studied factors causing distrust in users who rode in a prototype automated car on a real road for one hour a day over six days. During this ethnographic experiment, participants were asked to respond to a survey regarding their expectations of automated cars, experience with driving, and technology acceptance. A week following this survey, participants began the experiment, where they rode in the level 2 prototype automated car on a real road for one hour a day over six days. During the experiment, weather conditions varied. The same path was followed each day to allow participants to get familiar with the route. During each ride, a technician drove from the driver's seat, participants sat in the front passenger seat, and a researcher sat in the back seat to observe. For each trial, participants were also asked to complete varying secondary tasks while the car was in automated mode. Lee et al. identified three areas which seemed to cause distrust in users: Performance, Process, and Purpose. Performance consisted of functional incompetence (when the vehicle performed below expectations), lack of control (ability to take control of the vehicle), and lack of confidence (doubts that the car could overcome situational obstacles). Process consisted of lack of information (feeling of being uninformed about the car's actions), unpredictability (anxiety felt in unpredictable driving situations), and machine-likeness (machine-like movements of the vehicle). Purpose consisted of fiduciary irresponsibility (lack of clear responsibility in an automated vehicle), value incongruence (misalignment of passenger's and vehicle's values), and disloyalty (feeling of the car acting against the user's wishes).

Lee et al. (2015) examined the effects of appearance and different levels of autonomy on perceptions of an autonomous vehicle. The proposed research model suggested that Appearance

and Autonomy affect Social Presence, which in turn affects Perceptions of the vehicle (including cognitive and affective trust). In this study, autonomous vehicles were presented as either High Automation (vehicle can stop and start on its own) or Low Automation (participants must stop and start the vehicle) and as either Human-Like in Appearance (a NAO robot driving agent) or Gadget-Like in Appearance (An iPhone driving agent). In the Low Autonomy condition, participants were given an iPad to control the starts and stops of the vehicle. During the experiment, participants watched the "autonomous vehicle" (created from a children's remote-control car) drive three predetermined courses. On the third course, a "pedestrian" NAO crossed unexpectedly in front of the vehicle. In the Low Autonomy condition, the driving agent would warn the participant and tell them to stop the car. In the High Autonomy condition, the driving agent would warn the participant and stop the vehicle themself. They found that Human-Like Appearance and High Autonomy vehicles produced greater social presence, intelligence, safety, and trust. Appearance was found to only affect affective trust, while Autonomy affected affective and cognitive trust. They also found evidence supporting the idea that perceived social presence is a mediator of perceived intelligence, safety, and trust.

Li et al. (2020) study the relationship between BFI personality traits and trust in automation. For this study, participants rode in a simulated automated vehicle. They were instructed to complete unrelated mathematical tasks on a tablet while they felt that the vehicle was driving safely. They were informed that the vehicle was not perfect and may need them to take control at some points. These two points were scheduled accidents that could only be avoided if the participants took control in time. Personality was measured using the Chinese BFI, trust was measured using a modified pre-existing scale, and gaze behavior was monitored using Tobii Pro Glasses. This study found that individuals higher in openness were associated with less

trust in the vehicle. It is suggested that this may be due to high openness individuals' desire for "intellectually challenging activities".

Löcken et al. (2020) examined the effects of an ambient light display on trust in automated vehicles. In their study, each participant would experience each type of display: No Information (no light display), Conflicting Objects (a bright red bar displaying obstacles), and Trajectory and Conflicting Objects (a bright red bar displaying obstacles and a white bar displaying the trajectory). A Latin Square design was used to limit carry-over effects from each scenario. After each ride, participants completed a user experience questionnaire and a predeveloped trust scale. Following the drives, participants were asked to draw a UX curve describing the user experience of their ride. They were also given the option to write down the changes that occurred during the ride. Lastly, participants were asked to complete a semi-structured interview regarding the positive and negative aspects of the three displays and what additional information would be helpful. The Trajectory and Conflicting Objects display was rated as the most trustworthy and was also most frequently mentioned as having the best user experience in the interviews.

Ma et al. (2021) examined the relationship between different levels of visual feedback and trust in autonomous vehicles. To explore this relationship, they created a driving simulation containing 10 real-world driving scenarios where feedback was either No Feedback, Moderate Feedback (displaying traffic signs and what the vehicle can see), or High Feedback (displaying traffic signs, what the vehicle can see, and what the vehicle will do). The order of driving scenarios was constant, while the type of Feedback was randomized. They found that participants trusted the High Feedback interface the most, and that there were no significant differences in trust between the No and Moderate Feedback groups.

Mackay et al. (2019) examined how different types of feedback (no feedback [No], feedback on surrounding vehicles [Sensors], and feedback on surrounding vehicles and decision making [Decision]) affect trust in an autonomous vehicle. To explore this idea, participants were placed in a driving simulator and had electrodes attached to them to measure heart rate. The simulation took them on a 12 min. highway drive with several different event scenarios. A visual search where participants indicated the presence or absence of an upwards arrow in a grid of differently oriented arrows was used to indirectly measure trust in the system; this was completed at 3 min. and at 6 min. without notifying participants. Since this task required participants to shift their attention from the road, higher performance was associated with higher trust. Higher heart rates were associated with stress and, thus, lower trust. Trust was also measured using a post-task questionnaire. Across all feedback conditions, no significant differences in heart rate were found. More visual search answers were missing for Decision than the No and Sensors groups, possibly due to distraction from the amount of information. There were no significant differences in the percentage of correct answers between feedback types, but there was a performance increase after initial completion of the task. All feedback types were seen as trustworthy and safe and the intentions of the system were understood. There were some differences found between feedback levels in the amount that participants felt they understood the actions of the system, however. Overall, the findings indicate that autonomous vehicles of various feedback levels were seen as safe and trustworthy but giving too much information in feedback may be a hindrance.

Niu, Terken, and Eggen (2018) examined the effects of anthropomorphizing information on trust in autonomous vehicles. For this study, participants first rode in a simulated autonomous vehicle which presented no information to establish a baseline for the vehicle's performance.

Participants were then split into groups. The Symbolic group showed participants symbols representing the vehicle's actions. The Anthropomorphic group was the same, with the addition of animated eyes representing the vehicle's actions. Following the drive, participants completed a questionnaire which included items regarding perceived anthropomorphism, likeability, and trust. Interviews were conducted to collect qualitative data. Information style had significant effects on perceived anthropomorphism and trust (but not likeability), where the Anthropomorphic Information style was preferred. For trust, likeability, and perceived anthropomorphism, No and Symbolic Information were the same, but there were significant differences between No and Anthropomorphic Information. Additionally, perceived anthropomorphism was positively correlated with Trust. In the interviews, participants noted that the vehicle should provide information about road hazards, rather than just acting upon them. Others mentioned that the eyes should have been paying attention to the road. Overall, Anthropomorphism seems to increase trust in users, but it should likely be accompanied by more detailed information.

Oliveira et al. (2020) studied the effects of different types of interfaces on trust in a self-driving vehicle. In this study, each participant rode in a simulated automated pod-style vehicle where they each experienced several different interfaces. The interfaces included were Baseline, Third-Person Animation, Camera Feed Overlaid with Information, and Augmented reality (AR) Windscreen. During the task, participants were asked to tell the vehicle ‘‘OK Pod, take me to Tesco". The Pod would then take them through a simulated town where they would avoid several "hazards" along the way. The AR Windscreen was the most trusted interface, and the only to significantly differ from Baseline. This interface displayed the vehicles actions on the same screen as the rest of the visual information and highlighted hazards as they appeared.

Ruijten, Terken, and Chandramouli (2018) examined how conversational interfaces (that follow Gricean maxims of communication) and confidence can affect user trust. Participants were placed in a driving simulator where the simulation took them through 5 min. of urban driving followed by 5 min. of highway driving. Some participants were assigned to the Graphical User Interface (GUI) and others to the Conversational Interface (CUI); these were identical with the exception of the CUI providing spoken messages regarding the drive. Participants were given the choice of a male or female voice. All participants experienced both the high and low confidence conditions in a randomized order. Trust, perceived intelligence, likability, and anthropomorphism were measured using preexisting scales. The CUI was found to be trusted more, perceived as more intelligent, perceived as more human-like, and found to be more likeable than the GUI. Similarly, the high confidence interface was found to be trusted more, perceived as more intelligent, perceived as more human-like, and found to be more likeable than the low confidence interface.

Schwarz, Gaspar, and Brown (2019) analyzed differences in trust based on capability (more capable and less capable), order (more capable first or less capable first), age (18–25 or 25–55), and gender (male or female). Participants rode in a simulated automated vehicle for 2 30 min drives. During these drives, the same events occurred in differing orders. The start and stop locations of the drive also differed. Participant swerve asked to complete a distractor task of completing trivia questions during the drive and were told they would receive a payment bonus if they made above a certain score. Trust was measured during the drive via an interface that asked participants to rate their current comfort in the vehicle. They found that trust calibration achieved during the first drive affected trust on the second drive, especially with older people, of which women were more affected than men. When the more capable system was the second used,

younger people more easily recalibrated trust and spent less time looking at the road; older people became more vigilant. When the less capable system was the second used, people did not seem to lose their trust, which had already been calibrated.

Sun et al. (2020) explored how personalizing an autonomous vehicle (AV) to users' driving behaviors can affect users' trust in the system. The system was designed to personalize according to user data on driving speed, rate of acceleration, and event-specific behaviors. Each participant drove in 3 simulated scenarios: (1) a few seconds before the traffic lights changed from green to yellow at an intersection; (2) when a car was about to overtake from behind; and (3) when a truck ahead was moving slowly (~30 km/h). Participants manually operated the vehicle to establish a baseline. For experimental drives, the AV was programmed to drive like the person, with personalized speed, rate of acceleration, and event-specific behavior. Participants were asked to fill out a questionnaire concerning their perceived trust, comfort, and situational awareness. The personalized AV was found to be more trustworthy and comfortable to users because the system was considered more human-like and intelligent in their ability to drive and make decisions; users also felt they could understand the AV's actions when it acted like they did. The AV was also seen as reliable because it could meet users' expectations for driving. An insignificant difference was found in situational awareness; AV users were slightly less aware.

Verberne, Ham, and Midden (2015) investigated the effects that an autonomous agent's Similarity (Similar or Dissimilar) to its user would have on a user's trust. To manipulate Face Similarity, a digital face was created using images of participants, merged with a default male digital face, for the (male) autonomous agent; each participant thus produced a different version of the agent. For the Similar condition, participants were shown the agent's face that had been

created using their own face. For the Dissimilar, participants were shown an agent's face that had been created using a different participant's face. Head Movements were also manipulated; for the Similar condition, the agent would mimic the participant's head movements with a 4 s delay, and for the Dissimilar, it would use head movement data from the same participant used for the Dissimilar Face condition. To manipulate Driving Goals, participants were asked to rank the importance of comfort, energy, efficiency, and speed; for the Similar condition, the agent would share the participant's rankings, and for the Dissimilar, the agent's rankings would be reverse ranked from the participant's. The agent took participants through 13 driving scenarios. Participants responded to items regarding trust, liking, and perceived similarity to the agent, as well as manipulation checks. Trust was also measured indirectly using monetary trust games. They found that Similarity was positively correlated with trust and liking and that the Similar agent was more trusted. Trust was higher for the Similar agent regardless of experience with the system. These results held true for the driving simulator, but not the trust games.

Walker et al. (2018) examined differences in trust based off of experience with a level 2 automated car. Prior to the study, none of the participants had any experience with a level 2 automated vehicle. Participants were asked to drive and be passengers in a level 2 automated vehicle. The drive took about 20 minutes and was driven in various environments, including a motorway, an urban road, and a rural road. An expert rode in the front passenger seat; they explained the vehicle and its features to participants. Participants were asked to fill out a questionnaire which included items regarding age, gender, travelling profile, and attitudes toward new technology. Another part of the questionnaire included measures regarding trust toward level 2 cars in different scenarios. Participants filled out this questionnaire 3 times: before the drive, immediately following the drive, and two weeks following the drive. Overall, they found

that trust calibration improved, with both increases and decreases in some areas of trust, after experiencing the automated car. Participants seemed to generally overestimate the car's abilities to begin with, however. There were also no profound differences in trust immediately after the drive and two weeks after the drive.

Waytz, Heafner, and Epley (2014) examined how anthropomorphism, achieved through voice, gender, and name, can affect trust in an autonomous vehicle. This study was conducted using a driving simulator. Participants were assigned to either the Normal (non-automated vehicle that participants must control), Agentic (automated vehicle that can control its own steering and speed), or Anthropomorphic (same as Agentic, but the car was given the name Iris, given a female gender, and given a voice) condition. Experimenters read a script describing the vehicles' features and when to use them. Participants were then asked to drive 2 approximately 6 min. Courses where they would experience an accident caused by another driver. Physiological measures were used to measure heart rate change. In addition, a questionnaire was used to assess anthropomorphism, liking, trust, and blame for accidents. They found that the Anthropomorphic vehicle had the most perceived anthropomorphism and perceived and behavioral trustworthiness, followed by the Agentic vehicle, and lastly, the Normal vehicle. The Anthropomorphic and Agentic vehicles were equally liked and had equal self-reported trust, both more than the Normal vehicle. Anthropomorphism was found to be a mediator for overall trust. Participants blamed the accident on the car the most in the Agentic vehicle, followed by the Anthropomorphic vehicle, and lastly, the Normal vehicle. This implies a relationship between independent agency and responsibility, but it is unclear why the Anthropomorphic was less blamed than the Agentic.

Zhang, Yang, and Robert (2020) studied the relationship between expectations of autonomous vehicles and trust. In their experiment, Weather (Sunny or Snowy) and Driving

Behaviors (Normal Driving or Aggressive Driving) were manipulated. First, participants responded to items regarding expectations. After this, participants were asked to watch a video and respond to items regarding perceived performance and trust. This was done 4 times, once for each manipulation. They found that trust was highest when the vehicle performed better than expected. They also found that the impact of meeting expectations on trust can be affected by Weather and Driving Behavior.

Factors Influencing Trust in Robots

Babel et al. (2021) studied the effects of talk initiative, robot gaze behavior, and dialog content on trust in social robots. During this experiment, participants were given a script that described how they should interact with the robot. Talk initiative could be led by the human (where the script contained questions for the participant to ask the robot) or robot (where the robot asked the questions and the participant's script told them how to answer). The robot's gaze behavior could either be directed (looking at the participant without moving its head) or random (making slight random head movements). Dialog content would be at one time task-oriented (a conversation consisting of planning a trip with the robot as a travel agent) and at another time small talk (a conversation consisting of hobbies, travel, and food). They found that participants tended to trust the robot more when the robot initiated the talk, rather than when the human did. Acceptance, reliability, and performance were rated higher for the robot when completing the service task, rather than small talk. Robot performance was also rated better when the robot-initiated conversation during the service task. They also found that, in the service task, the robot was more accepted when it initiated the talk; the opposite was true of the small talk conversation. The robot was viewed as more anthropomorphic and was more accepted if it had a directed gaze during small talk.

Behrens et al. (2018) examined the influence of the gender of a robot's voice on trust. In their first study, they presented participants with a still image of a NAO robot accompanied by a randomly chosen text-to-speech generated (male or female) voice. Participants were then asked to fill out a questionnaire regarding their perceptions of the robot and their willingness to share information with the robot. In this study, they found that the male voiced robot was perceived as friendlier, more trustworthy, and having a more fitting voice. Participants were also more likely to ask the male voiced robot for help and more willing to share information with it.

In Behrens et al.'s (2018) second study, participants interacted directly with a NAO robot (half with a male voiced robot, half with a female voiced robot), which introduced itself, asked the participant if they felt good, offered the participant a seat, made small talk by asking them to share something they were looking forward to, asked participants to help retrieve a paper with a written URL (participants were told the URL was a website for advanced communication with NAO and they could decline to do this twice), continued small talking by asking participants if they would share something embarrassing, requested that they create and store a username and password on the website, and played an alarm informing the participant that time was up and they needed to leave the room. In this study, out of six participants, all were willing to share something they looked forward to, four shared something embarrassing, and two shared login credentials. Overall, sharing was equal between voices, however one participant mislabeled the gender and three said the robot was genderless. There also seemed to be validity issues with the login trust scenario.

Bernotat, Eyssel, and Sachse (2019) examined the relationship between robot gender (manipulated using robot torso's waist-to-hip ratio and shoulder width) and social judgements, such as trust. Their first study was used to ensure accuracy of gender depictions and equal robot

likeness and machine likeness; all of these were found to be at appropriate levels. In their second study, they measured how much male and female stereotyped tasks would require close HRI. They found that female stereotyped tasks were rated as needing closer HRI, which may have affected the results of their third, and main, study. In their main study, participants were told they were assisting in the evaluation of a new robot prototype. Participants were shown either the male or the female robot and rated their perceptions of them. They found that their male robot was perceived as male, and the female robot as female. Both robots were equally machine and human like and typical for a robot. The female robot was perceived as more communal, more capable of female stereotyped tasks, more trustworthy (cognitive trust), and more trustworthy (affective trust) than the male robot. Both robots were perceived as equally agentic and equally capable of male stereotyped tasks. However, these findings may have been affected by the desire to respond in a socially desirable manner or by benevolent sexist attitudes.

Brink and Wellman (2020) evaluated whether stating incorrect information would affect the trust a young child has in that robot. In their first study, there were four trials where two robots would give conflicting answers when asked to name an object that the children would recognize. The children were asked to indicate which robot was correct. After these trials, a single trial was completed where the children were asked to indicate which robot was not good at the task. Following this were four more trials, where the children were asked which robot they wanted help from in finding out the correct names of unfamiliar objects. Then, the two robots would again give conflicting answers when naming objects and the children would be asked which robot was correct. Following the trials, the children were asked questions regarding the robots' psychological agency and perceptual experience. They found that children could identify the inaccurate robot and were more likely to ask for information from and agree with the

accurate robot. Children in general viewed the robots as having agency and perceptual experience. The more children viewed the robots as having agency, the more likely they were to trust, endorse, and ask for information from the accurate robot.

In Brink and Wellman's (2020) second study, the robots were replaced with inanimate objects which provided no agency cues; everything else was identical. This time, the children were still able to identify the inaccurate machine, but they did not consistently trust, endorse, or ask the accurate machine for information. Children performed significantly better during the first study and agency was found to significantly predict performance. These results imply that perceived agency is an important factor in children's ability to trust robots.

Bryant, Borenstein, and Howard (2020) studied the effects of robot gender and gender role match on trust. In this study, gender was manipulated through the voice and name of a robot. The robot was presented as either male, female, or gender neutral and would be shown in a video where it introduces itself to participants. Participants were asked to rate how well the robot would perform in a variety of professions for which gender associations were established previously. Participants were also asked to respond to items regarding trust. They found that the gender of the robot did not significantly affect trust in the robot to complete job tasks.

Calvo et al. (2020) studied whether attempts at persuasion would affect childrens' trust in social robots. To study this, a child was paired with a Furhat robot and given the task of creating a story character. Children could choose the context (classroom, park), main character (adult, child, animal), character features (hair color, skin color, clothes, emotional expression/activity). During this task, the robot would either be Persuasive or Neutral. The Persuasive robot made explicit statements regarding the choices the child should make and made clear statements of approval (if the child chose what the robot wanted) or disapproval (if the child did not choose

what the robot wanted). The Neutral robot made statements that explained the task at hand or asked the child what their preferences were. At the end of the task, the robot asked the child to fill out a questionnaire where they responded to items regarding demographics, trust in the robot's advice and in the robot's goodness, likability of the robot, and enjoyability of the task. They found no differences in trust or cooperation between the Persuasive and Neutral robots. Their findings suggest that children may perceive the robot as a stranger or a peer. The robot was perceived as a stranger more often in the Persuasive condition than the Neutral condition. Children had previously been shown to dislike persuasion attempts from parents or mentors, so their viewing of robots as strangers or peers may have caused higher trust.

Fischer, Weigelin, and Bodenhagen (2018) examined the effects of transparency and robot adaptability on people's trust when a robot takes their blood pressure. To study this, they designed an experiment where a robot's behavior is manipulated in four ways: transparent only, adaptive only, transparent and adaptive, or normal. In transparent conditions, the robot gives several statements describing its actions. In the adaptive conditions, the robot moves forward, turns to face the participant, and then continues at a slower speed. In the adaptive conditions the robot also asks whether the arm position is comfortable, and the robot will adjust the arm according to the participants' specifications. The normal condition had neither the transparent nor adaptive manipulations. Prior to the experiment, participants filled out a questionnaire with items regarding demographic information, experience with robots, and experience with blood pressure measurement. Following the experiment, participants were asked to respond to items regarding their perceptions of the robot, their feelings of anxiety, agitation, and comfort, and to what degree they would prefer a robot to measure their blood pressure. They found that

transparency increased participant's feelings of trust, predictability, and control over the robot. Adaptability was found to have a weak effect on trust.

Gallimore et al. (2019) examined the relationship between a person's gender and their trust in a security robot. In this study participants were shown a video of a security robot equipped with a non-lethal weapon. The robot would first ask people to see an ID, then it would analyze the ID and instruct the person to proceed if their ID was accepted. The first two people gained approval and moved through. The third person was not granted access and was told to step away. The person became confused and stepped closer, at which point the robot said it was authorized to use force. The robot flashed a high intensity strobe light at the person who then covered their eyes and moved away. They found that women were more trusting of the robot than men. Females also tended to view the robot as more machine-like whereas men viewed it as more human-like. Men and women perceived the robot as having equal integrity. They also found that the robot would be equally accepted by men and women in a military setting; they preferred the robot in a military, rather than public setting.

Geiskkovitch et al. (2019) evaluated whether stating incorrect information would affect the trust a young child has in that robot. In the history phase, children were presented with two robots that would label familiar objects (such as a ball), one labelling correctly and one incorrectly. In the same label phase, the robots labelled objects that would be unfamiliar to children, using made up names. The robots would use the same name for different objects. The children were then asked to select which object had been properly labelled. In the contrast label phase, the robots used different labels for different objects. After hearing these labels, the children were asked to present the researcher with the object they did not know the name of. In the clean up task phase, children were told it was time to clean up some papers. The robots gave

the children their instructions on how to clean up. In the same label phase, children were found to trust the previously correct robot more. The contrast label and clean up task phases found no significant results. Overall, the robots' mistakes did hinder trust when the information the robots provided matched the task the child was to complete.

Ghazali et al. (2019) studied the effects of interactive social cues on reactance, liking, trusting, and compliance. Social Cues were manipulated through movement of the robot and the robot's praise of participants. Social Cues were either Interactive (random head movements and random social praise), Low (mimicking participant's head movement), or high (mimicking participant's head movement and appropriately timed praise). The participants were asked to complete 3 different tasks with each manipulation of Social Cues. The second 2 tasks introduced the robots making persuasive statements regarding the tasks. They found that reactance was lower when the robot mimicked head movements and offered more praise. They also found that a robot that praises can increase trust.

Hancock et al. (2011) conducted a meta-analysis of studies regarding elements that affect trust in robots. Studies included: empirical, direct measurements of trust as a dependent variable, trust with regards to a robot, human interaction or viewing of robots through physical, virtual, or augmented means, and enough information to determine effect size. The studies were then classified as either robot-related factors, human-related factors, or environment-related factors affecting trust. Among the correlational data, they identified a moderate relationship between trust and all factors identified as influencing HRI. They found that human-related and environment-related factors only had a small relationship with trust and are thus not incredibly important in the development of trust in a robot. Robot-related factors, however, had a moderate relationship with trust. Further analyzing robot-related factors, robot performance, rather than

robot attributes, was more strongly related with trust. Among the experimental data, they identified a large relationship between trust and all factors identified as influencing HRI. Robot-related factors affected trust the most, followed by a moderate effect for environment-related factors, and a very small effect for human-related factors. Again, performance, rather than attributes, was more strongly related with trust. Overall, robot-related factors, especially performance, seem to be most crucial in trust development.

Herse et al. (2018) studied the effects of embodiment and preference elicitation (the ability for users to disclose their preferences to the system) on trust in robots. During this experiment, participants interacted with either an Embodied (a robot with a touchscreen interface) or Disembodied (touchscreen interface without robot) agent. The agent would either Elicit Preference (asked for user's preferences) or Not Elicit Preference (no regard to user's preferences). The scenario consisted of choosing a restaurant for a confederate, using the agent's GUI designed for choosing restaurants. Risk was a factor in a pilot study, where Japanese food was the stated preference, but was not an available choice. They found that when there was no risk in decision making, embodiment of the agent made no differences in trust. They also found that when risk was present during decision making, there was a greater amount of trust in the system that did have preference elicitation.

Kraus, Wagner, and Minker (2020) studied the effects of proactive dialogue on user trust in robots. They asked participants to complete a DIY project planning task (an easy task of building a wooden nesting box and a more difficult task of assembling a wall candle holder made from copper tubes) where they had to make decisions on how to perform individual task steps. During this task, they were accompanied by a robot teammate which would have one of several levels of proactivity: None (robot must be explicitly asked for help), Notification (robot notified

participants that it had a solution, but participants had to ask to hear the solution), Suggestion (robot directly provided a solution and its reasoning), or Intervention (robot bypassed the user and chose a solution on its own). The robot could interfere in one of two ways: Fixed (robot offered proactive dialogue at fixed points) or Insecurity Measure (robot offered proactive dialogue after 12 sec of user inactivity). They found that the Notification and Suggestion robots were the most trusted. The Intervention robot was less trusted, possibly due to the user feeling unincluded in the decision-making process. Timing of proactive dialogue had no observable effect on trust.

Law, de Leeuw, and Long (2020) examined the effects of a non-humanoid robot's movements on trust. To study this, they created a scenario where a Cosmo robot would come out of hiding, making either positive (squinting, wiggling, rocking, moving gently) or negative (glaring, retreating into hiding, shaking its head, aggressive shaking, snapping movements) body movements, and be introduced to participants by the researchers. Participants were told to pick out 10 pieces of candy and write an offer on a piece of paper where Cosmo could not see. Participants were asked if they thought Cosmo would accept the offer and why; they were also asked to rate Cosmo's emotional valence. Following this, Cosmo was told the offer, to which it randomly responded yes or no. Participants were given the choice of playing the game once more; they were allowed to change their offer or keep it the same. Following this, participants were asked to complete a survey with questions regarding their previous exposure to robots and their comfort with computers. They found that Cosmo's movements did not affect trust, though this could have been due to poor measures of trust.

Law, Malle, and Scheutz (2021) examined the effects that observing a robot touching a human can have on trust in a robot. In their first experiment, they showed participants a video of

a short interaction between a human and a robot. During this interaction, the robot would stand by the human as the human entered data into a computer saying things like “okay” or “alright”. Following this, the human would turn to face the robot and the robot would then respond with either a positive, neutral, or negative statement regarding the human’s performance. The robot would then either touch the human’s shoulder or not touch them at all. The robot and the human’s gender were also varied in this video. In this study, they found that the robot that touched the human was seen as more trustworthy, especially for sincere & ethical (moral) trust. The touching robots were also seen as more comforting. Robots with a positive or neutral attitude were similarly seen as more trustworthy. The touching robots, as well as the negative attitude robots, were seen as more inappropriate than their counterparts.

In Law, Malle, and Scheutz’s (2021) second experiment, they saw an image of the robot shaking hands with the human, followed by the neutral attitude robot from the first experiment. They found in this experiment that touch made the robot appear more comforting, surprising, and inappropriate. They suggest comfort as a mediator between touch and trust.

In Law, Malle, and Scheutz’s (2021) third experiment, they used the same neutral attitude robot video again. It was randomized whether or not they saw the handshake picture preceding the video. This time, they introduced function, which could either be baseline, customer focused, or performance focused. In this experiment, they found that the touching robot was trusted less; this was however, not the case when participants saw the handshake picture first. Touch again made the robot appear more surprising and inappropriate. When they saw the handshake first, the robot was seen as more comforting, in addition to appearing more surprising and inappropriate. Inappropriateness was recommended as a mediator between touch and trust.

Van Maris et al. (2017) studied the differences in trust between a physical robot and a virtual agent. To study this, participants were assigned to either a physical robot (in person NAO robot) or a virtual agent (virtual representation of a NAO robot). They would interact with their assigned agent 10 times over a period of 6 weeks. During these interactions, the agent and participant would work together to complete a blank map. The team would be presented with a map where the agent would ask the participant for missing information, such as the names of countries or capital cities. If the participant answered incorrectly, the robot would pause and not say anything, allowing the participant to correct their mistake. After so long, the robot would offer the correct answer. A trust game was played after the first and last sessions with the agent. They found no significant differences in trust between the robot and virtual agents. They also found that, regardless of the agent's embodiment, trust significantly increased over time.

Park and Lee (2014) studied whether a social robot's skin temperature affects users' perceptions of the robot. For this study, they utilized a Pleo robot, which is designed to look like a dinosaur. The robot could move its head and tail and walk on four legs. Heating and cooling rays were installed underneath the robot's skin. During the experiment, participants were asked to watch randomly chosen clips from both a sad and a scary movie (these movies were chosen because previous studies indicated that they would increase physical interaction between participants and others). Participants were told that they could "hug, handle, and interact with" Pleo as they would a pet while watching the clips. Participants were randomly paired with either a cool (9◦C), intermediate (18◦C), or warm (32◦C) Pleo. Following the movie clips, participants were asked to respond to items regarding their perceptions of the robot, including intention to own Pleo, perceived friendship, anthropomorphism, perceived emotional stability, immersive tendency, and social presence. They found that skin temperature greatly influenced perceptions

of Pleo. The warm skinned Pleo saw increased perceived friendship, intention to own, perceived emotional stability, and social presence.

Schneider and Kummert (2020) examined differences in trust between Adaptable (human is in control of changing robot's behavior to suit the human's needs) and Adaptive (robot is in control of changing its own behavior to suit the human's needs) robots. During the experiment, a NAO robot (either Adaptive or Adaptable) guided participants through a series of exercises. The Adaptive robot utilized an algorithm that allowed it to learn user preferences and adapt the exercise schedule based upon those preferences. The Adaptable robot did not have the preference learning algorithm and thus had to ask participants to manually select each exercise. They found that users were more trusting of an Adaptive robot.

Sebo et al. (2018) studied the effects of vulnerable speech on trust in robots. In this study, participants worked in teams of three along with a robot that was either Neutral (Makes neutral comments and does not admit to mistakes) or Vulnerable (Makes vulnerable comments, including admitting to mistakes). These human-robot teams worked together on a collaborative railroad route construction game. During this game, each team member constructs one section of the railroad. The goal of the game is for everyone to construct their section of the railroad and for the railroad to use as little pieces as possible. In two rounds, failure was ensured by not providing or removing key pieces from the railroad. They found that a robot's Vulnerable behaviors influenced trust related behaviors (such as explaining errors, consoling other team members, and shared laughing) with the robot, but also with human team members.

Sebo, Krishnamurthi, and Scassellati (2019) studied the effects of robots using different trust repair strategies on trust. In their study, participants and robots compete in a game where they use spaceships to shoot asteroids. During the game, players can be assigned a power up

where they are given the choice between getting points for each asteroid on the screen (which would be more beneficial) or immobilizing the opponent (which would mostly be useful for frustrating the opponent). During the game, the robot would promise not to use the immobilization strategy and would violate the participants' trust by using it anyways. The robot could respond to the trust violation in one of four ways: competence-apology (robot says it mistakenly chose immobilization and apologizes), competence-denial (robot says it mistakenly chose immobilization and denies having immobilized the participant), integrity-apology (robot expresses excitement over immobilizing the participant but then apologizes), and integrity-denial (robot first expresses excitement over immobilizing the participant but then denies having immobilized the participant). They found that trust was higher for robots that apologize rather than deny a competence violation. Trust was also higher for robots that deny rather than apologize for an integrity trust violation. Participants who reciprocally promised not to immobilize the robots were more likely than those who didn't to not choose the immobilization power up. They found that participants in the integrity-denial condition immobilized the robot following the robot's trust violation at a rate of two times or greater the percentage of participants in the other conditions, implying that it may have a strong adverse effect on trust.

Song and Luximon (2021) examined the effects of face shape and facial width-to-height ratio (fWHR) on trust in robots. In their experiment, face shape could either be round or rectangular and fWHR could either be high (3:2), medium (1:1), or low (2:3). All other aspects of the robot's appearance were controlled for. Each participant was exposed to only one face shape and one fWHR. Participants were exposed to the stimuli, asked to complete a manipulation check, and then respond to items regarding trust and purchase intentions. They found that fWHR directly affects purchase intentions. It can also indirectly affect purchase intentions, with trust

acting as a moderator. They found no significant differences between round and rectangular face shapes. They found that, for human robot interaction, robots with high fWHR had higher perceived trustworthiness and those with low fWHR had lower perceived trustworthiness. The opposite was true for robots in interpersonal settings, rather than in human robot interaction. They also found that fWHR also had a significant impact on each construct comprising trust (ability, benevolence, and integrity).

Stanton and Stevens (2014) studied the effect of robot eye gaze and lifelike body movements on trust. In this study, participant-robot teams played the shell game visual tracking task. All participants experienced the Eye Gaze (on and off) conditions, where the robot would make eye contact, and the Task Difficulty (four levels from easy to very hard) conditions, where cup movement speed changed. Participants also experienced the Breathing (on or off) condition, where the robot either displayed rhythmic breathing motions or no motion, and the Eye Tracking (on or off) condition, where the robot's head would either follow a cup or not move. After each game, the robot would ask the participant to select the proper cup. The participants were able to ask the robot for help and the robot could agree or disagree. As Task Difficulty increased, they found that trust and frequency of help requests increased. With Gaze on, trust and help requests increased for highest Task Difficulty trials but decreased on the easier Task Difficulty trials. With Gaze on, correct answers increased on easier Task Difficulty trials, but decreased on harder Task Difficulty trials. Participants answered quicker with Gaze on, which increased with Task Difficulty. There were no significant findings for Eye Tracking or Breathing on trust.

Steain, Stanton, and Stevens (2019) studied the effects of ingroup/outgroup dynamics on human perceptions of robots. In their first experiment, participants played the shell game with two robots that they were told ran using different algorithms (one developed by engineers, one

by psychologists). The participants were first year psychology students, so they expected an ingroup bias towards the robot designed by psychologists. During the game, one or both robots would sometimes disagree with participants. Trust was measured by rate of change to a robot's given answer. They found no significant differences in trust between ingroup and outgroup robots (rather, a majority rules effect occurred), however, participants gave the ingroup robot more favorable ratings, maintained closer interpersonal distances to it,, and more frequently chose it, given a choice.

In Steain, Stanton, and Stevens' (2019) second experiment, the method remained the same, with the exceptions of one robot being presented at a time and warmth and competence being introduced as manipulations. Warmth was manipulated using robot head positioning, eye gaze, body leaning, and limb positioning; competence was manipulated using rate of correct responses. A Black Sheep Effect was found; low competence ingroup robots were rated lower on perceptions of anthropomorphism, intelligence, vision system performance and were more physically distanced from participants than low-competence outgroup robots. Thus, the deviant ingroup member was judged more harshly than a deviant outgroup member. Social judgement in robots was also found to rely more heavily on competence, rather than warmth; this is dissimilar to human interaction.

Van Straten et al. (2018) explored whether children distinguish between technological and interpersonal trust in a robot. During their experiment, children between the ages of 7 and 11 interacted with a Nao robot in a session that unfolded in 4 stages. Stage 1 consisted of the experimenter introducing the child to the robot. Stage 2 consisted of small talk between the robot and the child. Stage 3 consisted of a game (interspersed by more small talk) where the robot would make a series of assertions for the child to guess if they were true or false; after guessing,

the robot would give the child the correct answer and an explanation. Stage 4 consisted of the robot and child saying goodbye, followed by the child being led to an interview. Their findings suggest that children may differentiate between technological and interpersonal trust in robots. They describe their findings as implying that children may not view technological trust as preceding interpersonal trust and that the concept of trust as a whole may be elusive to children.

Stuck and Rogers (2018) studied the characteristics needed for older adults to trust robot caretakers. Participants were older than 65, had mild cognitive impairments, and received at least 4 days of care per week. The participants were interviewed regarding what conditions would need to be met for them to trust a robot caregiver and what would negatively impact their trust in a robot caregiver for 4 scenarios: bathing, medication assistance, transfer, and household tasks. Participants also responded to post-interview questionnaires including a robot self-efficacy scale (confidence in operating robots), a robot familiarity and usage questionnaire (familiarity with various robots), and a trust preference checklist (preference of human or robot caretaker for various tasks). Stuck and Rogers found that older adults had little to no experience and low self-efficacy with robots. They also preferred human caretakers. They did find that older adults with higher self-efficacy with robots were more likely to prefer robot caretakers. Professional skills (general capability, precision, consistency of performance, safety, predictability, and gentleness) were the most mentioned traits important for trusting robot care providers, followed by communication (task specific communication, engaging and responsiveness in communication, and ability of robot to understand and communicate clearly), and, lastly, personal traits (material of robot, appearance of robot, compatibility of robot, congruence of robot values with older adult's values, benevolence/kindness, and dress).

Ullman and Malle (2017) studied whether having the human in the loop would affect trust in a robot. During this experiment, participants and robots completed a simple movement task, where if the robot detects an obstacle, it generates a new path. Participants experienced one of two types of interaction: autonomous (robot executed the new path plan on its own) or involved (robot waited for a button press from the participant to execute the new plan). After this phase of the experiment, participants were presented with different use contexts and their perceptions towards robots in these use contexts were measured. They experienced two types of use contexts: social (robot works as a security officer in an airport and has to decide whether to give a suspicious person a pat-down.) and nonsocial (robot works as an engineer in a nuclear reactor facility and has to decide whether to shut down the reactor when it overheats). They found that trust was higher in the involved condition, when the human had a say in the implementation of the robot's plan. When looking at use cases for future robots, participants showed greater trust in the involved condition for social robots, as well as greater trust in general for social robots.

Vattheuer et al. (2020) studied how making a robot more humanoid can affect users' trust in the robot. During their experiment, they asked participants to work with an agent to answer questions regarding two sets of ten general math needed to pay bills at restaurants. They were tasked with reading the questions off a piece of paper to the agents and answering the questions vocally for the agents. The agent participants interacted with would either be a tablet or a humanoid robot. The first time participants interacted with the agent, it would perform perfectly. The second time, the agent would have three obvious mistakes throughout the trial. The Multi-Dimensional Measure of Trust scale was administered after each question and answer trial with the agent. This trust scale measured capacity trust, which had a reliable subscale (reliable,

predictable, someone you can count on, consistent) and a capable subscale (capable, skilled, competent, meticulous). It also measured moral trust, which had an ethical subscale (ethical, respectable, principled, has integrity) and a sincere subscale (sincere, genuine, candid, authentic). Their findings suggest that people are more trusting of a humanoid robot than a tablet for both capacity and moral trust. The only form of trust unaffected by the agent making an error was moral trust of a humanoid robot.

Volante et al. (2019) considered the effects of communication and social conformity on trust in robots. Participants completed the Interpersonal Trust Questionnaire (ITQ), mini International Personality Item Pool (IPIP), Negative Attitudes Towards Robots Scale (NARS), and Propensity to Trust Machines scale. Participants observed a simulation of a robot searching for trapped people in an apartment building. They were assigned to one of four stimuli manipulations using the following variables: Robot Communication (On or Off), whether the robot informs participants when it finds the search target, and Social Group (Positive or Negative), whether the chat dialogue was positive or negative regarding the robot's performance. Participants completed the Trust Perception Scale specific to Human-Robot Interaction (TPS-HRI) after viewing the stimuli. They found that social conformity did play a large role in trusting the robot. Social Group had a significant effect on Trust where Participants in the Positive Social Group were significantly more trusting. They found no effect of Robot Communication on Trust and no interactions between Robot Communication and Social Group. They also found that propensity to trust was a poor predictor of actual levels of trust.

Wijnen, Coenen, and Grzyb (2017) studied the effects of robot dishonesty on trust. To test this, they had participants complete a collaborative tower building task with a NAO robot. The participants and robots would build a tower together, with the participants following the

robot's instructions. Following the building of the tower, the robot would knock over the tower. The honest robot would admit its faults and apologize. The lying robot would deny any fault of its own and instead blame the participants for the tower being knocked over. The participants would then fill out a questionnaire regarding their assignment and then play a trust game with both robots in random order. In the trust game, the participants were given tokens which they could share with the robot. They were told that the robot would share any earned tokens with the participant. They found that lying significantly lessened trust in the robot, and perceived friendliness, kindness, and responsibility of the robot. Both robots were seen as equally competent, however.

Ye, Feigh, and Howard (2020) examined the relationship between human embodiment (when the body is used in cognitive processing) and trust in human-robot interactions. To examine this, an experiment was conducted where participants would be asked to memorize Greek letters with the help of a Pepper robot. Pepper was first introduced to participants by researchers; the participants were then asked to respond to items regarding demographic, age, gender, trust bias, experience with robots, and Pepper's likeability. Pepper would perform a motion representing a Greek letter. Participants would then either click to see the next motion using a touch screen or repeat the motion themselves (a Kinect was used to monitor when participants finished the motion). After 20 word-motion pairs, participants completed a quiz on the word-motion pairs as well as items regarding trust in Pepper and Pepper's likeability. They found that users with low initial trust gained more trust with human embodiment than with touch screen interaction.

Zhu and Williams (2020) examined the relationship between proactive explanations of actions and trust in robots. Their manipulations consisted of Proactive Announcement (PA),

where a robot informed participants of its behavior before acting, Proactive Explanation (PE), where a robot informed participants of its behavior and why it chose that behavior before acting, and No Explanations (NE), where a robot gave no proactive explanations. For this experiment, participants interacted with robots from these manipulations in a resource management task, where participants spent different types of resources while exploring an environment. During this task, the user could instruct the robot to collect a particular resource or the robot could decide independently to collect a different type of resource, based on the assessed need of that resource as needed to explore the current environment. They found that, through objective measurements, proactive explanations lead to greater trust in robots. They also proposed that the proactive explanations could have affected users' mental models of the robots, implying that it would change varying dimensions of trust.

Models of Trust

Boubin, Rusnock, and Bindewald (2017) set out to expand definitions of reliance and compliance, demonstrate a method for inferring trust by quantifying compliance and reliance, and create a model for compliance and reliance for automated systems. To begin this, they designed an experiment where participants completed a space navigation task where they could fly ships through space while trying to land on planets. The task included no fly zones, which would deduct points from the participants' scores, as well as bonus zones, which would add points. During the space navigation task, participants would be equipped with an automated agent of one of three types. The first type, Similar Automation, used an algorithm to provide a path for the ship based upon the participant's previous paths. The second type, Dissimilar Automation, used a trigonometric function, with no regard to the participant's play style, to create a sinusoidal curved path for the ship. The third type, Line Automation, drew straight line

paths that went directly to the planet. These automated aids took control of the ship's route if a participant had not created a route within two seconds. They found that agent type significantly affected compliance and reliance, with Similar Automation having the highest compliance and reliance rates, followed by Line Automation, and lastly, Dissimilar Automation. They also found that task load significantly affected compliance and reliance rates. Based upon results from this experiment, they describe a model where user trust is directly influenced by automation predictability and performance. In this model, trust directly affects compliance, reliance, and acceptance. The overall model explains factors that relate to user trust and stress.

Celmer, Branaghan, and Chiou (2018) recommended a framework where the brand of an autonomous vehicle has an effect on trust through several variables. This was based on the idea that each brand will have its own reputation and approach to autonomous vehicles. Brand Personality was described as being composed of the Sincerity, Excitement, Competence, Sophistication, and Ruggedness displayed by the brand. They suggested that the System, composed of Brand Personality and the Automation itself, informs the Intent (Why is this being developed? Will it benefit the user?), Method (How will it work? What is the user's experience? How will it approach difficulties and communicate with the user?), Competence (Will it achieve the user's goals? Does the brand have the expertise to create it?), and History (Was the outcome achieved well in the past? How does the brand's identity/reputation fit in the industry?) of the vehicle. The Intent, Method, Competence, and History of the vehicle were then said to inform Performance Expectation, also conceptualized as prospective trust (trust expectations prior to the user experiencing the system's performance).

Ferrario, Loi, and Viganò (2020) proposed an incremental model of trust, where the triple T = (*simple trust*, *reflective trust*, *paradigmatic trust*). The element of this triple, i.e. simple

trust, reflective trust, and paradigmatic trust can all be represented by a 5-tuple. This 5-tuple can be explained as (interacting agent #1, interacting agent #2, action to be performed by interacting agent #2, goal of relevance to interacting agent #1, context). They explain that interacting agent #2 will be performing an action, trying to accomplish the goal of interacting agent #1, all within a certain context. They provided different situations for which the 5-tuple would represent simple trust, reflective trust, or paradigmatic trust.

O'Connor, Heavin, and Kupper (2021) proposed a theoretical framework for trusting intentions towards healthcare robots. This framework takes into account multiple contextual factors, which include individual characteristics (such as age or gender), personality traits (such as the Big Five traits), health related attitudes or beliefs (based on the Health Belief Model), explanation competency (reasoning, granularity, and transparency), and patient-physician relationship (authoritative or participatory). These contextual factors are described as directly influencing both anthropomorphic features, which is also influenced by human-likeness (benevolence, integrity, and ability) and system-like features, which is also influenced by system-likeness (reliability, functionality, and helpfulness). Anthropomorphic features and system-like features then go on to directly affect trusting intentions towards robots, in a relationship that is moderated by perceived risk.

Ogawa, Park, and Umemuro (2019) proposed a model for trust development in social robots. In this model, they suggest that there are four phases of time (anticipating, connecting, interpreting, and reflecting), and that it takes increasing self-esteem to advance to each stage. They describe general trust (general trust in humans) and category trust (trust in a category of people) as being developed in the anticipating phase, individual trust (trust in a specific

individual) beginning development in the connecting phase, and social relationship trust (trust based on a relationship) being developed in the interpreting and reflecting phases.

To validate this model, Ogawa, Park, and Umemuro (2019) ran an experiment to illustrate how categories of trust develop over phases of human-robot interaction. At the beginning of the experiment, participants were asked to respond to items regarding attitude towards robots, general trust in robots, and category trust in social robots, self-esteem, and demographics. Representing the anticipation phase, participants were asked to watch a video introducing a NAO robot. Representing the connecting phase, participants were shown a NAO robot which greeted them. Representing the interpreting phase, participants had a conversation with a NAO robot. After each phase of communication, participants responded to items regarding general trust, category trust, individual trust, social relationship trust, and self-esteem. They found a significant correlation between self-esteem and general trust. They found support for general and category trust developing in the early anticipation phase, but not for individual trust developing in the connecting and interpreting phases. They found insignificant increases in social relationship trust through the interpreting phase.

Thiebes, Lins, and Sunyaev (2020) proposed a framework for trustworthy AI. One part of their framework is the principles of trustworthy AI, which includes beneficence, non-maleficence, autonomy, justice, and explicability. These principals form a matrix with the stages of data processing: input data, model data, and output data. At the intersections of the principals and the stages, tensions can be created, where the situation does not reflect the principle. The input data stage has the possibility of having less low quality data (creating tension with beneficence), possibility of malicious data or invasions of privacy (creating tension with non-maleficence), or possibility of biased data (creating tension with justice). The model data stage

has the possibility of being affected by lack of availability of AI models (creating tension with beneficence), possibility of invasion of privacy (creating tension with non-maleficence), possibility of users being unable to quantify an AI's uncertainties (creating tension with autonomy), possibility of bias in model design (creating tension with justice), or possibility of lack of transparency (creating tension with explicability). The output data stage has the possibility of invasion of privacy (creating tension with non-maleficence) or possibility of discrimination based on data (creating tension with justice). All of these interactions go on to inform the future of trustworthy AI.

Other Topics on Trust in Automation

Antifakos et al. (2005) studied the effects of displaying confidence information on trust in a context-aware mobile phone. In their first experiment, they displayed a set of scenarios for which they ranked criticality. They accomplished this by asking participants "With which modality would you like to be notified?", "How critical is it to you, that you are notified correctly in this situation?", and "How critical is it to your environment, that you are notified correctly in this situation?" for each scenario. In their second experiment, they presented participants with situations that categorized in experiment one. These situations were either low criticality (sitting in a tram, looking at shop windows, etc.), medium criticality (driving a car, buying chewing gum, etc.), or high criticality (attending a lecture, studying at a university library, etc.). Participants would experience all types of situations and confidence levels and would experience both a system showing and not showing confidence information. The confidence in this experiment is relating to the system notifying the participants in the most appropriate way regarding the situation. Their findings suggest that people are more trusting of

the system displaying confidence information and that people are more trusting of more confident systems.

Aoki (2020) examined the public's initial trust in AI chatbots used by the government. In their experiment, participants were introduced to a chatbot in a text snippet that described the chat bot's area of inquiry (general information, parenting support, tax consultation, or waste separation) and purpose (No statement of purpose, reduced burden on staff, more time for staff to perform other tasks, uniformity in response quality, or 24-hour, 365-day, timely responses). After reading about the chat bot, participants were asked "To what extent do you think you can trust the chatbot's response to your enquiry?" and "Between the human staff and the chatbot, which do you think you can trust more?"; for these questions, they responded on a scale of 1 to 100. They found that the area of enquiry can affect the public's initial trust in chatbots. They also found that communicating some purpose for the chatbot is better than communicating no purpose at all. They found no differences between different types of purpose, however.

Balas and Pacella (2017) studied differences in trust between real and artificial faces. In their first study, 40 images of real human faces were used for the Real face category. For the Artificial face category, the 40 images of real faces were given to a software that converted them into synthetic versions of the real faces. During the study phase, participants were presented with 10 Real and 10 Artificial faces of different individuals for which participants were asked to rate the trustworthiness of the face. During the next phase, participants completed a recognition task where novel faces, 10 Real and 10 Artificial, were introduced. From this study, they found that computer-generated faces were generally rated as less trustworthy. Participants had a better memory for real faces than artificial faces, but participants performed at above chance levels for recognizing both types of faces.

In Balas and Pacella's (2017) second study, the 5 real faces that yielded the Highest and the 5 real faces that yielded the Lowest Trustworthiness in Experiment 1 were used, as well as the artificial counterparts to each of these faces. During each trial of this experiment, participants were presented with either 2 Real or 2 Artificial faces, one from the High Trustworthy and one from the Low Trustworthy group. Participants were asked to identify which face appeared more trustworthy. Participants were significantly better at recognizing the High Trustworthy faces for the Real condition than the Artificial condition.

Gillath et al. (2020) examined the relationship between users' attachment styles and their trust in AI. In their first study, they specifically examined attachment style (attachment anxiety and avoidance) and trust in AI. They had participants respond to items regarding adult attachment style, neuroticism, self-esteem, experience with AI, familiarity with AI, and trust in AI. They found that as familiarity with AI increased, so did trust. They also found that higher anxiety scores were associated with lower trust scores.

In their second study, Gillath et al. (2020) set out to establish whether anxious attachment style causes low trust or not. To examine this, they primed participants for either an anxious, secure, or avoidance attachment style using previously developed methods. Participants were asked to recall a relationship they had that modeled their assigned attachment type and were then asked to describe that relationship for approximately 3 minutes. In this study, they found that older people were less trusting of AI, that people more familiar with AI were more trusting, and that anxious attachment was again associated with lower trust.

In their third study, Gillath et al. (2020) aimed to establish whether attachment security increased trust in AI or not. They also aimed to demonstrate that this effect is not due to an increase in positive affect gained by experiencing secure attachment. In this study, participants

were either primed for secure attachment, primed for positive affect, or not primed at all. Here they found that older people were less trusting of AI, that people more familiar with AI were more trusting, and that secure attachment increases trust in a way that cannot be attributed to positive affect.

Kraus et al. (2019) studied the effects of automation reliability, brand reputation, and need for cognition (personality trait entailing enjoyment and engagement in effortful cognitive activities) on trust in automated vehicles. They utilized a pilot study to establish brand reputation for real world automotive companies. For their main study, participants were randomly assigned to a condition with one level of reliability (low or high) and one level of brand reputation (below average, average, or above average). Participants were evaluated for need for cognition as a quasi-independent variable. Participants read a description of an automated vehicle that could take full control of driving but may need user interference in the case of situations past the vehicle's technical limitations. Participants were then asked to respond to items regarding attitudes towards automation. Participants then read a report describing the vehicle's brand and its reliability according to a fictional function test. Participants' trust in the vehicles was then measured. Following this, participants watched a video of someone driving the vehicle they had read about. These videos included the driver pressing a button to activate the automation, an automated take-over and take-over request, a system-initiated take-over request where the driver takes over control, and the driver deactivating the system. After watching the videos, the participants' trust was measured once more and the participants' need for cognition was measured. More reliable vehicles and vehicle brands with better reputation were initially trusted more than their counterparts; however, only the effect of reliability persisted over time. High need for cognition individuals were more likely to align their trust based upon reliability

information. They also found that materialism, regulatory focus, the perfect automation scheme, and high expectations were positively correlated with trust.

Large and Burnett (2014) explored how changing the voice of an in-vehicle navigation system (IVNS) can affect trust and attention. To study this, they designed an experiment where participants were asked to drive (in a simulator) to a real location in England. During this drive, participants were equipped with an IVNS and road signs. The IVNS would have one of two voices. The High Trustworthy Voice was named "Tim" and was the default male navigation voice for TomTom navigation devices; in a previous study "Tim" had been rated highest for trustworthiness. The Low Trustworthy Voice was American celebrity "Snoop Dogg", who was available as a downloadable character voice for navigation systems; in a previous study "Snoop Dogg" had been rated lowest for trustworthiness. Towards the end of the drive, road signs and the IVNS would indicate conflicting directions. Results indicated that people readily identified significantly different personality traits for "Tim" and "Snoop Dogg". Participants were more trusting of "Tim".

Pak et al. (2017) studied trust across user groups (young adults, military, older adults), four domains (consumer electronics, banking, transportation, health), two automation types (information, decision), and two levels of reliability (low, high). Participants were asked to respond to measures regarding competency with automation. They were then presented with text illustrating 16 different scenarios involving automation. Through these scenarios, they were exposed to every domain, automation type, and level of reliability. Participants then described their trust in the automation for each scenario. They found that trust relied on the interaction of domain, automation type, reliability, and user group. They also found that students accurately calibrated trust, while older adults and military personnel did not. Military personnel consistently

trusted decision automation less than information automation. Students trusted decision and information automation equally. Older adults were found to more strongly trust decision automation in banking and to show unusually high levels of trust in transportation automation regardless of type of automation.

In this review, Roff and Danks (2018) discussed many facets of trust in automated weapons systems. They asserted that trust is not a binary state, as it is often discussed, and is thus much more complicated. They discussed that trust can be based upon an understanding that a system will cooperate, regardless of understanding how or why the system cooperates. This trust is based on the predictability and reliability of a system. They then described a deeper kind of trust, based on interpersonal relationships and dependencies. They describe this as stemming from a trustor and trustee sharing a mental model of the world, i.e., the trustor knows how the trustee will act and why they act that way. They described both of these types of trust as being necessary in a military setting, describing that one must have trust while also verifying that the system is capable. They proposed several challenges to developing trust in autonomous weapons systems. One such challenge is a lack of predictability, which could impact trust development, liability, and more. As a system becomes more predictable, however, it may come to a point where it is seen as less autonomous. Human teammates may also suffer from automation bias, where they may overtrust a system due to thinking it is more capable than it really is. Poor communication with human teammates (for example, if communications are jammed) may also cause decreases in trust. Some believe that autonomous weapons systems may be more moral than humans, but if a system begins to share a human teammate's mental model, it may act as morally as the human does. They described an inverse relationship between efficiency and trust for these systems; as the systems become more competent, they could also potentially alienate

users because they act too differently. They suggested multiple routes to fostering trust in autonomous weapons systems, including relying on transitive trust and incorporating every stakeholder throughout the design process. However, each of these recommendations come with some heavy drawbacks or are either not feasible.

Smith, Allaham, and Wiese (2016) studied the differences in trust between different agents and task types. During this experiment, there were three different agent types: Human (an undergraduate student sitting at home and performing the task with the participant), Avatar (A human-like cartoon character, described as an online support tool, and Computer (A matrix arrangement of red, green and blue lights, described as a desktop computer). This experiment also contained two different types of tasks: Social (determining emotion based off an image of eyes) and Analytical (solving a math problem). Participants experienced each type of task with each type of agent. The agents would offer their answers to each question; trust was measured by compliance with the agent. They found that trust relied on the combined influence of task and agent type, rather than both separately influencing trust. The Human agent was more trusted on the Social task, while the Avatar and Computer agents were more trusted on the analytical task. They suggest that these results are caused by people's expected expertise of the agents and trying to find which agent best suits the situational context.

Waggoner et al. (2019) examined the effects of big-data related terminology on trust in public policy automation. For their study, they had participants define algorithms in their own words and rank the order of importance of features of algorithms. They then presented participants with pairs of algorithms (modeled after those that would be used by a judge in determining criminal sentencing) to choose between. The participants were shown six randomly ordered design features: human role in the algorithm design, location from which the algorithm's

data was collected, number of factors, size of training data, source of algorithm designer; and transparency of the algorithm's code. Participants were next asked to indicate whether they trusted three different algorithms to make criminal justice decisions in their states. They found that, when ranking importance of algorithm features, two big data heuristics, size of the training data and number of input features, were most commonly ranked highest. Their conjoint experiment showed that, with marginal mean values, algorithms with the most features and the largest training data sizes are most likely to predict favorability. This experiment also showed, with average marginal component effects, that training data size and number of defendant features were the most likely to affect algorithm selection. These results show that people consistently prefer the algorithms with profiles containing big data related features. When asking participants if they trusted three different algorithms, tracking data indicates that the big data related factors were not the only elements considered. Essentially, this study found that big data related terminology can enhance trust in algorithms, but it is not the only factor that is taken into consideration.

Yamani, Long, and Itoh (2020) discussed how the COVID-19 Pandemic may affect trust in automation, specifically in users naïve to automation, based on findings from previous research. They describe that, during the pandemic, many have been encouraged and/or required to adopt automated technologies. Thus, many naïve users have been experiencing many different automated technologies for the first time. They describe that active monitoring and analysis of system behaviors is crucial to forming trust at appropriate levels in automated systems. This is because naïve users may initially trust systems based on faith. This faith would likely be based on superficial information gathered from things like brand familiarity or hearsay and may not be accurate to the system. Users can also base faith in interactions, but these still may not be reliable

due to the small amount of interactions that a new user would have. Furthermore, many new users, let alone experts, do not have good understandings of how automated systems work. By including active monitoring and analysis of system behaviors, the system will become more transparent, and thus, allow users to calibrate their trust to a system's actual operation, rather than unstable opinions. They suggested that new users forced into using automation may stop trusting a system if they do not understand how and why it works that way. They recommended that manufacturers focus on giving users clear, accurate information about their technologies.

Reference List

Ajenaghughrure, Ighoyota Ben., Sonia Claudia da Costa Sousa, and David Lamas. "Risk and Trust in Artificial Intelligence Technologies: A Case Study of Autonomous Vehicles." In *2020 13th International Conference on Human System Interaction (HSI)*, 118–23. Tokyo, Japan: IEEE, 2020. https://doi.org/10.1109/HSI49210.2020.9142686.

Antifakos, Stavros, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. "Towards Improving Trust in Context-Aware Systems by Displaying System Confidence." In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services - MobileHCI '05*, 9. Salzburg, Austria: ACM Press, 2005. https://doi.org/10.1145/1085777.1085780.

Antrobus, Vicki, Gary Burnett, and David R. Large. "'Trust Me –I'm AutoCAB': Using Natural Language Interfaces to Improve the Trust and Acceptance of Level 4/5 Autonomous Vehicles." In *Proceedings of the 6th Humanist Conference*. The Hague, Netherlands, 2018. https://www.docdroid.net/wNIfQsg/8-antrobus-pdf.

Aoki, Naomi. "An Experimental Study of Public Trust in AI Chatbots in the Public Sector." *Government Information Quarterly* 37, no. 4 (October 2020): 101490. https://doi.org/10.1016/j.giq.2020.101490.

Aoki, Naomi. "The Importance of the Assurance That 'Humans Are Still in the Decision Loop' for Public Trust in Artificial Intelligence: Evidence from an Online Experiment." *Computers in Human Behavior* 114 (January 2021): 106572. https://doi.org/10.1016/j.chb.2020.106572.

Babel, Franziska, Johannes Kraus, Linda Miller, Matthias Kraus, Nicolas Wagner, Wolfgang Minker, and Martin Baumann. "Small Talk with a Robot? The Impact of Dialog Content, Talk Initiative, and Gaze Behavior of a Social Robot on Trust, Acceptance, and Proximity." *International Journal of Social Robotics*, January 6, 2021. https://doi.org/10.1007/s12369-020-00730-0.

Balas, Benjamin, and Jonathan Pacella. "Trustworthiness Perception Is Disrupted in Artificial Faces." *Computers in Human Behavior* 77 (December 2017): 240–48. https://doi.org/10.1016/j.chb.2017.08.045.

Behrens, Sofie Ingeman, Anne Katrine Kongsgaard Egsvang, Michael Hansen, and Anton Mikkonen Møllegård-Schroll. "Gendered Robot Voices and Their Influence on Trust." In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 63–64. Chicago IL USA: ACM, 2018. https://doi.org/10.1145/3173386.3177009.

Bernotat, Jasmin, Friederike Eyssel, and Janik Sachse. "The (Fe)Male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots." *International Journal of Social Robotics*, May 25, 2019. https://doi.org/10.1007/s12369-019-00562-7.

Boubin, Jayson G., Christina F. Rusnock, and Jason M. Bindewald. "Quantifying Compliance and Reliance Trust Behaviors to Influence Trust in Human-Automation Teams." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, no. 1 (September 2017): 750–54. https://doi.org/10.1177/1541931213601672.

Brauner, Philipp, Ralf Philipsen, André Calero Valdez, and Martina Ziefle. "What Happens When Decision Support Systems Fail? — The Importance of Usability on Performance in Erroneous Systems." *Behaviour & Information Technology* 38, no. 12 (December 2, 2019): 1225–42. https://doi.org/10.1080/0144929X.2019.1581258.

Brink, Kimberly A., and Henry M. Wellman. "Robot Teachers for Children? Young Children Trust Robots Depending on Their Perceived Accuracy and Agency." *Developmental Psychology* 56, no. 7 (July 2020): 1268–77. https://doi.org/10.1037/dev0000884.

Bryant, De'Aira, Jason Borenstein, and Ayanna Howard. "Why Should We Gender?: The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency." In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 13–21. Cambridge United Kingdom: ACM, 2020. https://doi.org/10.1145/3319502.3374778.

Calvo, Natalia, Maha Elgarf, Giulia Perugia, Christopher Peters, and Ginevra Castellano. "Can a Social Robot Be Persuasive Without Losing Children's Trust?" In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 157–59. Cambridge United Kingdom: ACM, 2020. https://doi.org/10.1145/3371382.3378272.

Celmer, Natalie, Russell Branaghan, and Erin Chiou. "Trust in Branded Autonomous Vehicles & Performance Expectations: A Theoretical Framework." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62, no. 1 (September 2018): 1761–65. https://doi.org/10.1177/1541931218621398.

Chavaillaz, Alain, Adrian Schwaninger, Stefan Michel, and Juergen Sauer. "Expertise, Automation and Trust in X-Ray Screening of Cabin Baggage." *Frontiers in Psychology* 10 (February 14, 2019): 256. https://doi.org/10.3389/fpsyg.2019.00256.

Ferrario, Andrea, Michele Loi, and Eleonora Viganò. "In AI We Trust Incrementally: A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions." *Philosophy & Technology* 33, no. 3 (September 2020): 523–39. https://doi.org/10.1007/s13347-019-00378-3.

Fischer, Kerstin, Hanna Mareike Weigelin, and Leon Bodenhagen. "Increasing Trust in Human–Robot Medical Interactions: Effects of Transparency and Adaptability." *Paladyn, Journal of Behavioral Robotics* 9, no. 1 (June 1, 2018): 95–109. https://doi.org/10.1515/pjbr-2018-0007.

Forster, Yannick, Frederik Naujoks, and Alexandra Neukum. "Increasing Anthropomorphism and Trust in Automated Driving Functions by Adding Speech Output." In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 365–72. Los Angeles, CA, USA: IEEE, 2017. https://doi.org/10.1109/IVS.2017.7995746.

Gallimore, Darci, Joseph B. Lyons, Thy Vo, Sean Mahoney, and Kevin T. Wynne. "Trusting Robocop: Gender-Based Effects on Trust of an Autonomous Robot." *Frontiers in Psychology* 10 (March 8, 2019): 482. https://doi.org/10.3389/fpsyg.2019.00482.

Geiskkovitch, Denise Y., Raquel Thiessen, James E. Young, and Melanie R. Glenwright. "What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots." In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 48–56. Daegu, Korea (South): IEEE, 2019. https://doi.org/10.1109/HRI.2019.8673024.

Ghazali, Aimi Shazwani, Jaap Ham, Emilia Barakova, and Panos Markopoulos. "Assessing the Effect of Persuasive Robots Interactive Social Cues on Users' Psychological Reactance, Liking, Trusting Beliefs and Compliance." *Advanced Robotics* 33, no. 7–8 (April 18, 2019): 325–37. https://doi.org/10.1080/01691864.2019.1589570.

Gillath, Omri, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. "Attachment and Trust in Artificial Intelligence." *Computers in Human Behavior* 115 (February 2021): 106607. https://doi.org/10.1016/j.chb.2020.106607.

Gold, Christian, Moritz Körber, Christoph Hohenberger, David Lechner, and Klaus Bengler. "Trust in Automation – Before and After the Experience of Take-over Scenarios in a Highly Automated Vehicle." *Procedia Manufacturing* 3 (2015): 3025–32. https://doi.org/10.1016/j.promfg.2015.07.847.

Ha, Taehyun, Sangyeon Kim, Donghak Seo, and Sangwon Lee. "Effects of Explanation Types and Perceived Risk on Trust in Autonomous Vehicles." *Transportation Research Part F: Traffic Psychology and Behaviour* 73 (August 2020): 271–80. https://doi.org/10.1016/j.trf.2020.06.021.

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, no. 5 (October 2011): 517–27. https://doi.org/10.1177/0018720811417254.

Herse, Sarita, Jonathan Vitale, Meg Tonkin, Daniel Ebrahimian, Suman Ojha, Benjamin Johnston, William Judge, and Mary-Anne Williams. "Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System." In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 7–14. Nanjing: IEEE, 2018. https://doi.org/10.1109/ROMAN.2018.8525581.

Jensen, Theodore, Yusuf Albayram, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. "The Apple Does Fall Far from the Tree: User Separation of a System from Its Developers in Human-Automation Trust Repair." In *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1071–82. San Diego CA USA: ACM, 2019. https://doi.org/10.1145/3322276.3322349.

Keller, David, and Stephen Rice. "System-Wide versus Component-Specific Trust Using Multiple Aids." *The Journal of General Psychology* 137, no. 1 (December 21, 2009): 114–28. https://doi.org/10.1080/00221300903266713.

Kraus, Johannes Maria, Yannick Forster, Sebastian Hergeth, and Martin Baumann. "Two Routes to Trust Calibration: Effects of Reliability and Brand Information on Trust in Automation." *International Journal of Mobile Human Computer Interaction* 11, no. 3 (July 2019): 1–17. https://doi.org/10.4018/IJMHCI.2019070101.

Kraus, Matthias, Nicolas Wagner, and Wolfgang Minker. "Effects of Proactive Dialogue Strategies on Human-Computer Trust." In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 107–16. Genoa Italy: ACM, 2020. https://doi.org/10.1145/3340631.3394840.

Kunze, Alexander, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. "Automation Transparency: Implications of Uncertainty Communication for Human-Automation Interaction and Interfaces." *Ergonomics* 62, no. 3 (March 4, 2019): 345–60. https://doi.org/10.1080/00140139.2018.1547842.

Large, David R., and Gary E. Burnett. "The Effect of Different Navigation Voices on Trust and Attention While Using In-Vehicle Navigation Systems." *Journal of Safety Research* 49 (June 2014): 69.e1-75. https://doi.org/10.1016/j.jsr.2014.02.009.

Law, Theresa, Josh de Leeuw, and John H. Long. "How Movements of a Non-Humanoid Robot Affect Emotional Perceptions and Trust." *International Journal of Social Robotics*, October 21, 2020. https://doi.org/10.1007/s12369-020-00711-3.

Law, Theresa, Bertram F. Malle, and Matthias Scheutz. “A Touching Connection: How Observing Robotic Touch Can Affect Human Trust in a Robot.” *International Journal of Social Robotics*, January 5, 2021. https://doi.org/10.1007/s12369-020-00729-7.

Lee, Jae-Gil, Ki Joon Kim, Sangwon Lee, and Dong-Hee Shin. “Can Autonomous Vehicles Be Safe and Trustworthy? Effects of Appearance and Autonomy of Unmanned Driving Systems.” *International Journal of Human-Computer Interaction* 31, no. 10 (October 3, 2015): 682–91. https://doi.org/10.1080/10447318.2015.1070547.

Lee, Jieun, Genya Abe, Kenji Sato, and Makoto Itoh. “Effects of Demographic Characteristics on Trust in Driving Automation.” *Journal of Robotics and Mechatronics* 32, no. 3 (June 20, 2020): 605–12. https://doi.org/10.20965/jrm.2020.p0605.

Lee, Jiin, Naeun Kim, Chaerin Imm, Beomjun Kim, Kyongsu Yi, and Jinwoo Kim. “A Question of Trust: An Ethnographic Study of Automated Cars on Real Roads.” In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 201–8. Ann Arbor MI USA: ACM, 2016. https://doi.org/10.1145/3003715.3005405.

Li, Wenmin, Nailang Yao, Yanwei Shi, Weiran Nie, Yuhai Zhang, Xiangrong Li, Jiawen Liang, Fang Chen, and Zaifeng Gao. “Personality Openness Predicts Driver Trust in Automated Driving.” *Automotive Innovation* 3, no. 1 (March 2020): 3–13. https://doi.org/10.1007/s42154-019-00086-w.

Löcken, Andreas, Anna-Katharina Frison, Vanessa Fahn, Dominik Kreppold, Maximilian Götz, and Andreas Riener. “Increasing User Experience and Trust in Automated Vehicles via an Ambient Light Display.” In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–10. Oldenburg Germany: ACM, 2020. https://doi.org/10.1145/3379503.3403567.

Ma, Rachel H. Y., Andrew Morris, Paul Herriotts, and Stewart Birrell. “Investigating What Level of Visual Information Inspires Trust in a User of a Highly Automated Vehicle.” *Applied Ergonomics* 90 (January 2021): 103272. https://doi.org/10.1016/j.apergo.2020.103272.

Mackay, Ana, Inês Fortes, Catarina Santos, Dário Machado, Patrícia Barbosa, Vera Vilas Boas, João Pedro Ferreira, Nélson Costa, Carlos Silva, and Emanuel Sousa. “The Impact of Autonomous Vehicles’ Active Feedback on Trust.” In *Advances in Safety Management and Human Factors*, edited by Pedro M. Arezes, 969:342–52. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-20497-6_32.

Maris, Anouk van, Hagen Lehmann, Lorenzo Natale, and Beata Grzyb. “The Influence of a Robot’s Embodiment on Trust: A Longitudinal Study.” In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 313–14. Vienna Austria: ACM, 2017. https://doi.org/10.1145/3029798.3038435.

Niu, Dongfang, Jacques Terken, and Berry Eggen. "Anthropomorphizing Information to Enhance Trust in Autonomous Vehicles." *Human Factors and Ergonomics in Manufacturing & Service Industries* 28, no. 6 (November 2018): 352–59. https://doi.org/10.1002/hfm.20745.

O' Connor, Yvonne, Matesuz Kupper, and Ciara Heavin. "Trusting Intentions Towards Robots in Healthcare: A Theoretical Framework." In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021. http://scholarspace.manoa.hawaii.edu/handle/10125/70682.

Ogawa, Rui, Sung Park, and Hiroyuki Umemuro. "How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust." In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 606–7. Daegu, Korea (South): IEEE, 2019. https://doi.org/10.1109/HRI.2019.8673090.

Oliveira, Luis, Christopher Burns, Jacob Luton, Sumeet Iyer, and Stewart Birrell. "The Influence of System Transparency on Trust: Evaluating Interfaces in a Highly Automated Vehicle." *Transportation Research Part F: Traffic Psychology and Behaviour* 72 (July 2020): 280–96. https://doi.org/10.1016/j.trf.2020.06.001.

Pak, Richard, Ericka Rovira, Anne Collins McLaughlin, and Natalee Baldwin. "Does the Domain of Technology Impact User Trust? Investigating Trust in Automation across Different Consumer-Oriented Domains in Young Adults, Military, and Older Adults." *Theoretical Issues in Ergonomics Science* 18, no. 3 (May 4, 2017): 199–220. https://doi.org/10.1080/1463922X.2016.1175523.

Pan, Ye, and Anthony Steed. "A Comparison of Avatar-, Video-, and Robot-Mediated Interaction on Users' Trust in Expertise." *Frontiers in Robotics and AI* 3 (March 29, 2016). https://doi.org/10.3389/frobt.2016.00012.

Park, Eunil, and Jaeryoung Lee. "I Am a Warm Robot: The Effects of Temperature in Physical Human–Robot Interaction." *Robotica* 32, no. 1 (January 2014): 133–42. https://doi.org/10.1017/S026357471300074X.

Pearson, Carl J., Michael Geden, and Christopher B. Mayhorn. "Who's the Real Expert Here? Pedigree's Unique Bias on Trust between Human and Automated Advisers." *Applied Ergonomics* 81 (November 2019): 102907. https://doi.org/10.1016/j.apergo.2019.102907.

Pearson, Carl J., Allaire K. Welk, William A. Boettcher, Roger C. Mayer, Sean Streck, Joseph M. Simons-Rudolph, and Christopher B. Mayhorn. "Differences in Trust between Human and Automated Decision Aids." In *Proceedings of the Symposium and Bootcamp on the Science of Security*, 95–98. Pittsburgh Pennsylvania: ACM, 2016. https://doi.org/10.1145/2898375.2898385.

Roff, Heather M., and David Danks. "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems." *Journal of Military Ethics* 17, no. 1 (January 2, 2018): 2–20. https://doi.org/10.1080/15027570.2018.1481907.

Rovira, Ericka, Richard Pak, and Anne McLaughlin. "Effects of Individual Differences in Working Memory on Performance and Trust with Various Degrees of Automation." *Theoretical Issues in Ergonomics Science* 18, no. 6 (November 2, 2017): 573–91. https://doi.org/10.1080/1463922X.2016.1252806.

Ruijten, Peter, Jacques Terken, and Sanjeev Chandramouli. "Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior." *Multimodal Technologies and Interaction* 2, no. 4 (September 24, 2018): 62. https://doi.org/10.3390/mti2040062.

Schmidt, Philipp, Felix Biessmann, and Timm Teubner. "Transparency and Trust in Artificial Intelligence Systems." *Journal of Decision Systems* 29, no. 4 (October 1, 2020): 260–78. https://doi.org/10.1080/12460125.2020.1819094.

Schneider, Sebastian, and Franz Kummert. "Comparing Robot and Human Guided Personalization: Adaptive Exercise Robots Are Perceived as More Competent and Trustworthy." *International Journal of Social Robotics*, February 8, 2020. https://doi.org/10.1007/s12369-020-00629-w.

Schwarz, Chris, John Gaspar, and Timothy Brown. "The Effect of Reliability on Drivers' Trust and Behavior in Conditional Automation." *Cognition, Technology & Work* 21, no. 1 (February 2019): 41–54. https://doi.org/10.1007/s10111-018-0522-y.

Sebo, Sarah Strohkorb, Priyanka Krishnamurthi, and Brian Scassellati. "'I Don't Believe You': Investigating the Effects of Robot Trust Violation and Repair." In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 57–65. Daegu, Korea (South): IEEE, 2019. https://doi.org/10.1109/HRI.2019.8673169.

Smith, Melissa A., M. Mowafak Allaham, and Eva Wiese. "Trust in Automated Agents Is Modulated by the Combined Influence of Agent and Task Type." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, no. 1 (September 2016): 206–10. https://doi.org/10.1177/1541931213601046.

Song, Yao, and Yan Luximon. "The Face of Trust: The Effect of Robot Face Ratio on Consumer Preference." *Computers in Human Behavior* 116 (March 2021): 106620. https://doi.org/10.1016/j.chb.2020.106620.

Stanton, Christopher, and Catherine J. Stevens. "Robot Pressure: The Impact of Robot Eye Gaze and Lifelike Bodily Movements upon Decision-Making and Trust." In *Social Robotics*, edited by Michael Beetz, Benjamin Johnston, and Mary-Anne Williams, 8755:330–39. Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-11973-1_34.

Steain, Andrew, Christopher John Stanton, and Catherine J. Stevens. "The Black Sheep Effect: The Case of the Deviant Ingroup Robot." Edited by Stefano Triberti. *PLOS ONE* 14, no. 10 (October 16, 2019): e0222975. https://doi.org/10.1371/journal.pone.0222975.

Stokes, Charlene K., Joseph B. Lyons, Kenneth Littlejohn, Joseph Natarian, Ellen Case, and Nicholas Speranza. "Accounting for the Human in Cyberspace: Effects of Mood on Trust in Automation." In *2010 International Symposium on Collaborative Technologies and Systems*, 180–87, 2010. https://doi.org/10.1109/CTS.2010.5478512.

Straten, Caroline L. van, Jochen Peter, Rinaldo Kühne, Chiara de Jong, and Alex Barco. "Technological and Interpersonal Trust in Child-Robot Interaction: An Exploratory Study." In *Proceedings of the 6th International Conference on Human-Agent Interaction*, 253–59. Southampton United Kingdom: ACM, 2018. https://doi.org/10.1145/3284432.3284440.

Strohkorb Sebo, Sarah, Margaret Traeger, Malte Jung, and Brian Scassellati. "The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams." In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178–86. Chicago IL USA: ACM, 2018. https://doi.org/10.1145/3171221.3171275.

Stuck, Rachel E., and Wendy A. Rogers. "Older Adults' Perceptions of Supporting Factors of Trust in a Robot Care Provider." *Journal of Robotics* 2018 (2018): 1–11. https://doi.org/10.1155/2018/6519713.

Sun, Xu, Jingpeng Li, Pinyan Tang, Siyuan Zhou, Xiangjun Peng, Hao Nan Li, and Qingfeng Wang. "Exploring Personalised Autonomous Vehicles to Influence User Trust." *Cognitive Computation* 12, no. 6 (November 2020): 1170–86. https://doi.org/10.1007/s12559-020-09757-x.

Thiebes, Scott, Sebastian Lins, and Ali Sunyaev. "Trustworthy Artificial Intelligence." *Electronic Markets*, October 1, 2020. https://doi.org/10.1007/s12525-020-00441-4.

Ullman, Daniel, and Bertram F. Malle. "Human-Robot Trust: Just a Button Press Away." In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 309–10. Vienna Austria: ACM, 2017. https://doi.org/10.1145/3029798.3038423.

Vattheuer, Christopher, Annalena Nora Baecker, Denise Y. Geiskkovitch, Stela Hanbyeol Seo, Daniel J. Rea, and James E. Young. "Blind Trust: How Making a Device Humanoid Reduces the Impact of Functional Errors on Trust." In *Social Robotics*, edited by Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge, 12483:207–19. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-62056-1_18.

Verberne, Frank M. F., Jaap Ham, and Cees J. H. Midden. "Trusting a Virtual Driver That Looks, Acts, and Thinks Like You." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, no. 5 (August 2015): 895–909. https://doi.org/10.1177/0018720815580749.

Volante, William G., Janine Sosna, Theresa Kessler, Tracy Sanders, and P. A. Hancock. "Social Conformity Effects on Trust in Simulation-Based Human-Robot Interaction." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 61, no. 5 (August 2019): 805–15. https://doi.org/10.1177/0018720818811190.

Waggoner, Philip D., Ryan Kennedy, Hayden Le, and Myriam Shiran. “Big Data and Trust in Public Policy Automation.” *Statistics, Politics and Policy* 10, no. 2 (December 18, 2019): 115–36. https://doi.org/10.1515/spp-2019-0005.

Walker, Francesco, Anika Boelhouwer, Tom Alkim, Willem B. Verwey, and Marieke H. Martens. “Changes in Trust after Driving Level 2 Automated Cars.” *Journal of Advanced Transportation* 2018 (August 5, 2018): 1–9. https://doi.org/10.1155/2018/1045186.

Waytz, Adam, Joy Heafner, and Nicholas Epley. “The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle.” *Journal of Experimental Social Psychology* 52 (May 2014): 113–17. https://doi.org/10.1016/j.jesp.2014.01.005.

Wijnen, Luc, Joost Coenen, and Beata Grzyb. “‘It’s Not My Fault!’: Investigating the Effects of the Deceptive Behaviour of a Humanoid Robot.” In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 321–22. Vienna Austria: ACM, 2017. https://doi.org/10.1145/3029798.3038300.

Yamani, Yusuke, Shelby K. Long, and Makoto Itoh. “Human–Automation Trust to Technologies for Naïve Users Amidst and Following the COVID-19 Pandemic.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 62, no. 7 (November 2020): 1087–94. https://doi.org/10.1177/0018720820948981.

Ye, Sean, Karen Feigh, and Ayanna Howard. “Learning in Motion: Dynamic Interactions for Increased Trust in Human-Robot Interaction Games.” In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1186–89. Naples, Italy: IEEE, 2020. https://doi.org/10.1109/RO-MAN47096.2020.9223437.

Yuksel, Beste F., Penny Collisson, and Mary Czerwinski. “Brains or Beauty: How to Engender Trust in User-Agent Interactions.” *ACM Transactions on Internet Technology* 17, no. 1 (March 6, 2017): 1–20. https://doi.org/10.1145/2998572.

Zhang, Qiaoning, X. Jessie Yang, and Lionel Peter Robert. “Expectations and Trust in Automated Vehicles.” In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9. Honolulu HI USA: ACM, 2020. https://doi.org/10.1145/3334480.3382986.

Zhang, Yunfeng, Q. Vera Liao, and Rachel K. E. Bellamy. “Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372852.

Zhu, Lixiao, and Thomas Williams. “Effects of Proactive Explanations by Robots on Human-Robot Trust.” In *Social Robotics*, edited by Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge, 12483:85–95. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-62056-1_8.