PERCEIVED AGENCY CHANGES TRUST IN ROBOTS

by

Chelsea Frazier
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Arts
Psychology

Committee:

________________________________ Director

____

____

________________________________ Department Chairperson

________________________________ Dean, College of Humanities and Social Sciences

Date: ____________________________ Spring Semester 2022
George Mason University
Fairfax, VA

Perceived Agency Changes Trust in Robots

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts at George Mason University

by

Chelsea Frazier
Master of Arts
Webster University, 2018
Bachelor of Science
Morgan State University, 2012

Director: Tyler Shaw, Professor
Department of Psychology

Spring Semester 2022
George Mason University
Fairfax, VA

DISCLAIMER CLAUSE: The views expressed in this paper are those of the author and do not reflect the official policy of the United States Army, Department of Defense, or the U.S. Government.

ACKNOWLEDGEMENTS

First and foremost, I would want to give all the glory and praise to God for guiding and molding me through yet another challenging season of my life.

I would also like to thank my family, friends and colleagues, whose prayers, support, positive energy, and candid feedback helped make this project possible. Specifically, I would like to thank Dr. Greg Trafton for his mentorship, patience, and guidance throughout my time here at George Mason University. He demonstrated the keen ability to help me through complex concepts and procedures in a manner and pace that aligned with my skillset and research interests. I really enjoyed soaking up your unique perspectives on a weekly basis. A big thanks goes out to my other committee members, Dr. Tyler Shaw, and Dr. Elizabeth 'Beth' Phillips. Thank you for your continued support and exposing me to your specific areas of expertise and interests. It was a pleasure working with you all and I am excited to continue collaboration efforts with you in the future.

A special thank you goes out to the Office of Naval Research and the United States Military Academy at West Point, New York for providing the resources towards this project and entrusting me to carry out this unique assignment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

Analysis of Variance	ANOVA
Alpha, type I error rate	α
Beta, regression coefficient	b
Bivariate correlation coefficient	*r*
Confidence interval	CI
Critical t-value	*t*
Negative Attitudes Toward Robots Scale	NARS
Partial eta squared	η_p^2
Sample Mean	$\bar{X}$(M)
Size of sample	N
Standard deviation	SD
Statistical Probability	*P*
Mean squared error	MSE
Multi-dimensional Measure of Trust	MDMT
Multivariate correlation coefficient	*R*
Human-robot interaction	HRI
Human-robot trust	HRT
Variance ratio	*F*

ABSTRACT

PERCEIVED AGENCY CHANGES TRUST IN ROBOTS

Chelsea Frazier, M.A.

George Mason University, 2022

Thesis Director: Dr. Greg Trafton, Dr. Tyler Shaw

This thesis experimentally investigated the effect of people's perception of a robot's compliance (and resistance) to social norms on their evaluation of a robot's perceived agency and trust. Participants reported higher perceived agency and trust to a norm-conforming robot compared to a norm-violating robot (experiment 1). Furthermore, results from experiment 1 found that perceived agency, regardless of how much a robot followed norms, was positively correlated with trust; therefore, suggesting that as people see a robot as having agency, they trust it more. While specifically examining the effect of negative attitudes on the relationship between social norms, perceived agency and trust, results from experiment 2 replicated experiment 1 and found that participants' negative attitudes towards robots did not impact perceptions of trust in robots. These findings are useful in that they provide understanding for situations when people decide whether to trust a norm-conforming robot with perceived agency, regardless of if they feel negatively about robots.

These results suggest that robot designers should pay careful attention on how to integrate robots with perceived agency cues. This is particularly important for safety critical robots in health care or the military because the more perceived agency a robot has, the more people seem to trust it (i.e., agency cues may not be representative of competency or ethics). Ideally, robot designers should understand the implications of integrating a robot high in perceived agency and low in capability, or conversely, low in perceived agency and high in capability. Future HRI research is necessary with specific emphasis on executable tools for measuring perceived agency for a diversified set of robots.

INTRODUCTION

One of the most important questions within human-robot-interaction is "when do people trust a robot?" One possible reason is that people are more likely to trust a robot when they believe that the robot has more perceived agency. Evidence of this possible link comes from an interesting paper by Parasuraman & Miller (2004). In a study of automation etiquette on user trust, Parasuraman and Miller (2004) found that, "good automation etiquette can compensate for low automation reliability." The behavior in Parasuraman and Miller (2004) was manipulated through communication style, where a computer would display "interruptive" and "impatient," or "non-interruptive" and "patient" language while participants completed a flight simulation task. Specifically, the communication style represented a norm where people held each other accountable for communicating in a manner that was not interruptive or impatient. Participants in this study demonstrated a propensity to trust the robot while associating the robot's behavior with good manners which was more meaningful than the robot's skill.

Highlighting that while Parasuraman and Miller focused on etiquette, I believe that etiquette is a type of social norm such that people or systems who follow rules of etiquette are following established social norms. The current experiment directly tests the hypothesis that when an automated system (a robot in this case) follows social norms, the overall system will be more likely to be trusted. Since automated systems that follow social norms also seem to have an increase in perceived agency (Korman et al., 2019), results should demonstrate that when an automated system follows social norms, it will lead to increased perceived agency which in turn will lead to the system being trusted. The remainder of this introduction discusses previous work on trust and perceived agency, especially as it relates to robotics and human-robot interaction.

Defining and Conceptualizing Trust

Capturing the level of trust a person has in a machine/robot is quite challenging due to the diversified ways in which trust is conceptualized. Some researchers have proposed that people trust machines/robots by how consistent they are (van den Brule, Dotsch, Bijlstra & Wigboldus, 2014; Kidd & Breazeal, 2004; Ross, Szalma, Hancock, Barnett & Taylor, 2008). While others suggested that people place more trust in a robot's ability to discern right from wrong (Banks, 2020; Bringsjord, Arkoudas, & Bello, 2006; Wallach, 2010). Hancock, Kessler, Kaplan, Brill, and Szalma (2021) presented a thorough summary of definitions while evaluating their similarities and differences. Among the 21 definitions of trust highlighted in their research, the top five most used terms consisted of "individual", "vulnerable", "expectation", "party", and "action". The diversity within these terms suggests that trust is a dynamic concept and difficult to define. Therefore, developing appropriate methodologies for creating experiment scenarios where trust is needed requires researchers to emphasize the agents involved and the environmental context in which the interaction takes place.

One of the early definitions of trust that has been widely accepted is "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rosseau, Sitkin, Burt & Camerer, 1998) where the trustor psychologically displays confidence in the trustee prior to any behavioral action. This definition primarily focuses on the trustor's willingness to be vulnerable; thus, trust has the capacity to exist. However, others have focused on the shared contextual environment between the trustor and trustee where trust is "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004). In this definition of trust an environment of both vulnerability and uncertainty exists. Another common way trust has been defined is "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others' " which emphasizes the idea of loss avoidance (Hancock, Billings, Schaefer, Chen, & de Visser, 2011). Furthermore, identifying the appropriate definition for trust and contributing factors for HRT remains a challenge due to difficulty with replication.

Also, interesting research by Kim, Wen, de Visser, Zhu, Williams & Phillips (2021) provided insight for other potential outcomes concerning the perceptions of trustworthy or untrustworthy robots. In a study examining moral advice to deter cheating behaviors, Kim et al., (2021), found that participants were more willing to accept moral advice from a human rather than the social robot (NAO) which was proactively offering moral advice. This finding alluded to situations in which a human might choose not to take sound advice from a robot even if the robot was perceived as morally sound to some degree. Perhaps, people that preferred to take sound advice from the human rather than robot perceived the robot as being programmable; and therefore, lacking both competency and morality. The current experiment aims to explore perceptions of people's trust in robots regardless of whether the robot acts morally correct (e.g., follows social norms) . Differently from Kim et al., (2021), the current experiment specifically suggests the contradiction in human behavior to trust in a robot in some cases alludes to the likelihood that perceived agency influences people's decision to trust (or distrust) a robot.

Defining and Conceptualizing Perceived Agency

As robots become more common and "intelligent", there has been a separate effort at understanding how different behaviors influence perceived agency. Much of HRI research has proposed attributions of perceived agency through a robot's appearance (Zhao, Phillips & Malle, 2019), eye gaze (Moon, Troniak, Gleeson, K.X., Pan & Zhen, 2014), and transparency (de Graaf & Malle, 2017). Many researchers have performed similar studies; and therefore, attempted to adopt inclusive ways of thinking about, defining and measuring perceived agency. To develop a hypothesis for the relationship between perceived agency and trust, the current experiment attempts to define perceived agency similarly to Dennett's (1978) definition: "People perceive agency in another when its actions may be assumed by an outside observer to be driven by its internal cognitive and/or emotional states." For a robot, this means it has not been programmed to behave in a specific way for a specific situation.

Gray, Gray and Wegner's (2007), research for how people perceived the mind in robots (machines and other inanimate beings) led to a principal component analysis (PCA) and factor analysis (varimax rotation) of 18 mental capacities resulting in two main dimensions: agency and experience. In this context agency included seven capacities from self-control and morality to memory and planning. Experience contained 11 capacities such as hunger, rage, consciousness, and joy. The agency and experience dimensions were developed through the comparison of 13 characters on each mental capacity. An interesting result emerged concerning how people ranked the level of agency each character had on a scale between “0” to “1”. For example, in comparing four of the 13 agents- girl, robot, frog, and fetus, people ranked the robot and girl relatively equally between “0” and “1”, and the frog and fetus equally at “0”. Gray et al.’s (2007) research provides a framework in which perceived agency in robots can be better understood.

Almost a decade later, Malle (2019) sought to compare previous research on people’s perception of minds while adopting a different methodology from Gray et al. (2007). Instead of analyzing 11 mental capacities, Malle (2019) used 28 capacities, resulting in a conditional solution including three-to-five dimensional structures. The three-dimensional structure included Affect (positive and negative feelings), Moral and Social Cognition (upholding moral values and setting goals), and Reality Interaction (communicating verbally). The five-dimension structure was a split of the Affect dimension (Social Cognition and Positive Social Affect) and Moral and Social Cognition (Moral Cognition and Social Cognition) respectively. Similarly, to the multi-layered construct of trust, the results from Gray et al. (2007) and Malle (2019) illustrated the complexities with defining and measuring agency.

Where researchers such as Gray et al. (2007) and Malle (2019) emphasized a conceptual approach toward understanding how people might consider robots and their minds, others have explored the features and capabilities that people ascribe to a robot an attempt to answer how much agency it has, and conditions in which perceived agency increases or decreases. For example, Short, Hart, Vu and Scassellati (2010) conducted an experiment where humans played a simple game of *rock-paper-scissors* with robots. In a between-participants' experimental design, each participant interacted with a humanoid robot (Nico) that either played according to the rules, declared itself the winner when it had not won (verbal cheat), or changing its hand symbol after seeing the opponents winning (or tie) hand symbol (action cheat). Interestingly, Short et al. (2010) found that participants engaged more and made greater mentalistic attributions to robots in conditions in which it cheats. In this case, the robot's interesting behavior elicited an observation of the robot's agency cues to the participant; and therefore, suggesting that people perceive robots that demonstrate the capacity to cheat as more agentic than robots that do not.

In comparing Short et al. (2019), people that perceive robots explicitly violating a norm (e.g., cheating), might feel as though the robot has a greater ability to carry out its own intentions and desires. Korman, Harrison, McCurry and Trafton (2019) specifically examined the effect of social norms on perceived agency while manipulating a robot's behavior. In a study of participants watching videos of a DRC-HUBO conducting realistic tasks by way of a norm violation, norm-conforming, and unintentional violation , Korman et al. (2019) sought to understand the relationship between social norms and perceived agency. In contrast to Short et al. (2010), Korman et al. (2019) found that people attributed more agency to the robot that did not violate a social norm compared to the robot that intentionally or unintentionally violates a social norm. The contrast between what Short et al. (2010), and Korman et al. (2019) found presents a unique relationship between social norms and perceived agency, such that people are willing to attribute a robot as having high perceived agency regardless of whether the robot followed a social norm.

Aside from the features and capabilities of a robot, a common theme presented in previous studies is that perceived agency includes a robot's intentional behavior (Levin, Adams, Saylor & Biswas, 2013) and a subjective assessment unique to an observer (Pollini, 2009). Thus, perceived agency is a multifaceted and complex factor in which previous studies reinforce the need for the replication, classification, and comparison to effectively develop a comprehensive measure.

The current set of experiments combines the results from Korman et al.'s (2019) and Short et al.'s (2010) experiments, and research from Levin et al. (2013) and Pollini (2009) to develop a hypothesis for perceived agency. Furthermore, the following experiment experimentally examines whether the relationship between perceived agency and trust helps to explain situations when people attribute a robot as having high perceived agency even when a robot violates a social norm.

Consequence of robot following social norms: First, I predict that when a robot follows social norms, people will attribute it as having higher perceived agency. Evidence of this relationship is demonstrated in research by Korman et al. (2019), and Yasuda, Doheny, Salomons, Strohkorb, and Scassellati (2020) suggesting that if a robot understands human social norms, people perceive a robot following human social norms to have more understanding of the world, more autonomy, and thus more agency.

Also, in an experiment by Ullman, Leite, Phillips, Cohen and Scassellati (2014), they explored how behavior types demonstrated by both a human and a robot affected attributions of trustworthiness, intelligence, and intentionality to human observers. Their results showed that a robot's behavior (e.g., honest, dishonest) significantly impacted attributions of their trustworthiness and intentionality.

Consequence of robot acting in a socially accepted way: Second, I predict that when a robot follows social norms, it will lead to higher trust in robots. Evidence of this relationship is demonstrated in research by Ullman et al. (2014) suggesting that when a robot acts in an expected way, people are more willing to trust it.

Consequence of robot having perceived agency: Third, I predict that higher perceived agency will lead to higher trust in robots. Evidence of this relationship is subsumed in research by Parasuraman and Miller (2004) and Ullman et al.'s (2014) suggesting that a robot that acts more like a person, capable of making its own decisions (not just 'programmed'), will be perceived as more trustworthy.

The following experiment utilizes a combination of measures, to develop a framework for investigating the relationship between a norm-conforming and norm-violating robot with perceived agency and trust, and the relationship between perceived agency and trust. Therefore, I believe that when people observe a robot that follows social norms, they will attribute it as having higher perceived agency. Also, as people observe a robot as having high perceived agency, regardless of whether the robot follows social norms, the more likely they are to trust it.

NORMING STUDY

First, a norm study was conducted to validate the video stimuli from the Korman et al.'s (2019) experiment. The following experiment adopted Korman et al.'s (2019) methodology by setting up a 3 x 3 between subjects' experiment to examine how people interpreted the three norm-behaviors (norm-conforming, violation, and mistake conditions) demonstrated by three norm-types (line, elevator, and trash scenarios). To begin the experiment, participants were instructed to watch one 30-second video, where a DRC-HUBO robot on wheels carried out one of three norm-behaviors through one of three norm-types. For example, the *norm-conforming* condition illustrated the robot performing normal actions while joining the end of a line, entering an elevator at an arms-length distance from an individual, or throwing away garbage in a trashcan. Conversely, the *norm-violation* condition illustrated the robot disregarding social norms in a blatant manner. In the videos, the robot either cut in line, invaded personal space in the elevator, or littered. In the *mistake* condition, the robot performed all actions in an unintentional manner. For example, the robot entered the perceived break in a line (scenario) where a group of people were occluded around a corner. In the *elevator* scenario, the robot accidentally bumped into an individual while entering inside. In the *trash* scenario, the robot dropped garbage just short of the trashcan.

Results of the norm study revealed that people distinctly recognized when a robot demonstrated norm-conforming or violating behaviors in only the *line* and *elevator* scenarios. Additionally, people did not recognize when a robot behaved unintentionally (or behaviors deemed a mistake) regardless of the norm-type. Therefore, the following two experiments will focus on the *line* and *elevator* scenarios with the pure *norm-violating* and *norm-conforming* conditions. I hypothesize that people in the non-violation condition (collapsed across scenario) will attribute more perceived agency and trust to the robot compared to people in the violation condition. Furthermore, I believe that when people attribute a robot high in perceived agency, regardless of condition, are more likely to trust the robot.

EXPERIMENT 1

Methodology

Participants

One hundred forty-one participants (84 males, 55 females, 2 prefer not to answer) participated in this study. Of the 141 participants, 11 (8%) participants were excluded for missing the attention check question (10 participants) or providing incomplete responses (1 participant). The final sample included 130 (77 males, 52 females, 2 prefer not to answer) with ages ranging from 23 to 70 years old (*M* = 36, *SD* = 9.38). Participants were recruited from Amazon's Mechanical Turk online platform and were invited to complete the online survey in exchange for pay. The number of participants needed to detect a medium effect size with 80% power and $\alpha = 0.05$ was $n = 134$. The Power analysis was calculated *a priori* using G*Power Software, $t = 1.98$.

Design

A between-subjects design was used for this study. Each participant was randomly assigned to watch one 30-second video associated with each of the study conditions. After finding no differences in participant scores for the line and elevator scenarios (see Experiment 1 Results section), norm-behaviors were treated as one factor with two levels consisting of the norm-conforming and violation conditions. Table 1 shows the number of participants in each condition prior to data cleaning.

Scenarios	Conditions	
	Norm-Conforming	**Norm-Violation**
Line	37 Participants	34 Participants
Elevator	36 Participants	34 Participants

Table 1: Number of participants in each experimental condition (Experiment 1)

Materials

Stimuli The experimental setup included four sets of the 30-second videos from Korman et al.'s (2019) research replicating situations in which socially acceptable behavior is demonstrated in day-to-day activity (e.g., entering an elevator, standing in a line) video of the DRC-HUBO robot.

Measures Participants answered the three questions where Korman et al. (2019) found perceived agency: "Did the robot perform this behavior intentionally: Was the robot aware of engaging in this behavior: Did the robot want to perform this behavior?" The questions were combined into a single "perceived agency" scale (by averaging) on a 7-point rating scale from 1 (Not at all) to 7 (Very much). Also, trust was assessed using Ullman and Malle's (2019) Multi-Dimensional Measure of Trust (MDMT v2) where participants rated the robot on a scale from 0 (Not at all) to 7 (Very) or 8 (Does Not Fit) to a total of 20 Likert scale items consisting of two sub-scales-performance trust (e.g., reliable, and competent) and moral trust (e.g., ethical, transparent, benevolent). Participant scores were computed as the average of all 20 Likert scale items. Participants who selected "Does Not Fit" were treated as missing values.

Procedure

Experimental Procedure To begin the experiment, participants were instructed to watch one video of the DRC-HUBO robot. After watching the video, participants were prompted to describe the robot's behavior in written form. Next, participants responded to the perceived agency and trust questionnaires in which one attention check question was inserted. After responding to the two questionnaires, participants were prompted with a free-text box for general feedback and debriefed on the experiment.

Experiment 1 Results

A two-sample t-test was conducted to inspect differences in participant scores for perceived agency and trust between the line and elevator scenarios. The t-test did not show significant differences for perceived agency $t(128) = -1.65$, $p=.101$ or trust $t(127) = -0.08$, $p=.937$. As a result, I collapsed scores in the two scenarios. First, I performed two t-tests to examine significant differences in participant means scores between the norm-conforming and norm-violating conditions (independent variable) with perceived agency and trust (dependent variables). A t-test comparing means on perceived agency showed that participants in the norm-conforming condition ($M = 5.83$; $SD = 1.65$), attributed the robot as having more perceived agency compared to participants in the norm-violating condition ($M = 5.14$; $SD = 1.80$, $t(128) = 2.94$, $p=.004$ $d = 0.52$ 95% CI [0.22 – 1.15]), $MSE = 1.75$, *partial eta squared* = 0.06[1].

Next, a t-test examining significant differences in participant means scores between the norm-conforming condition and norm-violating condition with trust showed that people felt a greater sense of trust in a norm-conforming robot ($M = 5.34$; $SD =$

[1] A Wilcoxon test showed similar results where people attributed perceived agency to a norm-conforming robot significantly greater than a norm-violating robot ($w = 2701.5$, $p=.003$).

1.08), compared a norm-violating robot (*M* = 4.05; *SD* = 2.46),), *t*(127) = 4.36, *p*<.001*d* = 0.77 95% CI [0.71 – 1.88]), *MSE* = 2.79, *partial eta squared* = 0.13[2][3].

For the final analysis, a Pearson correlation coefficient was performed to examine the linear relationship between perceived agency and trust scores by condition. As suggested by figure 1, there was a positive correlation between perceived agency and trust scores *r*(127) = 0.27, *p*=.002[4].

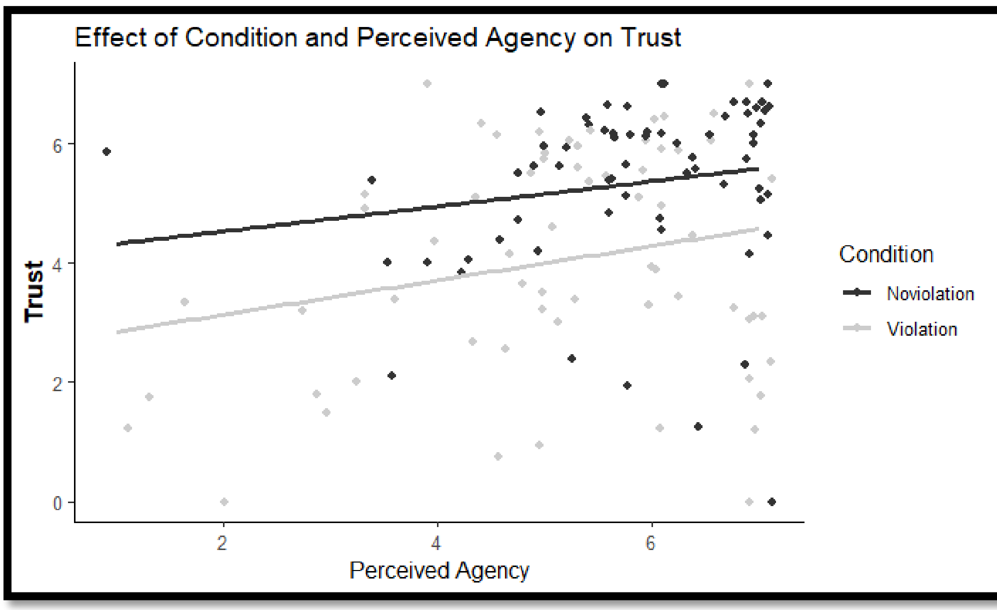

Figure 1: Pearson's correlation examining the relationship between Perceived Agency and Trust scores in norm-conforming and norm-violation conditions (Experiment 1)

[2] A Wilcoxon test showed similar results where people trusted a norm-conforming robot significantly greater than a norm-violating robot (*w* = 2964.5, *p*<.001).
[3] An independent samples t-test showed similar results for performance trust *t*(125) = 4.03, *p*<.001, and moral trust *t*(120) = 4.33, *p*<.001.
[4] Spearman's rank correlation showed similar results *rs*(127) = 0.24, *p*=.006.

Discussion

Experiment 1 examined the effects of people's perception of a robot conforming to and violating social norms on their evaluation of a robot's perceived agency and trust. As hypothesized, the analysis on perceived agency replicated Korman et al. (2019), which found that people rated a norm-conforming robot higher on perceived agency than a norm-violating robot. This result is also consistent with Yasuda et al. (2020), who showed that norm violations did not increase people's perception of perceived agency.

While comparing mean perceptions of trust, the current experiment found that people rated a norm-conforming robot higher than a norm-violating robot. The conclusion made from this analysis is consistent with Ullman et al.'s (2014) research who showed that honest behaviors increased people's perception of perceived trustworthiness compared to dishonest behaviors. Notably, figure 1 suggests that trust, regardless of condition, was positively correlated with perceived agency. Results from experiment 1 illustrated the significant influence of social norms and perceived agency on trust where a robot high in perceived agency was perceived as more trustworthy. This result was consistent with Falcone, Castelfranchi, Cardoso, Jones and Oliveria's (2013), Korman et al.'s (2019), and Ullman et al.'s (2014) research suggesting that people are willing to trust machine/robots that follow social norms or high in perceived agency.

Additionally, researchers Stafford, MacDonald, Jayawardena, Wegner, & Broadbent, (2013), and Tussyadiah, Zach, and Wang (2020) has shown that how people feel about robots is a relevant explanation for understanding how people decide whether to trust a robot. It could be, for example, that people who rated the robots low in trust or low in perceived agency had a predisposed negative attitude towards robots. Thus, I examined people's negative attitudes toward robots while attempting to replicate the primary findings in experiment 2.

EXPERIMENT 2

Most researchers generally agree that psychological dispositions are critical factors for trust development between humans and robots (de Visser et al., 2019; Kohn et al., 2021; Mcknight et al., 2011). Evidently, a negative attitude towards a robot is one factor contributing to a person's holistic perception of the robot. Therefore, it could be that negative attitudes towards robots influences perceptions of trust. For example, Tussyadiah et al.'s (2020) research found that trust is negatively influenced by negative attitudes towards robots and technology in general while investigating perceptions of U.S. travelers' trust in intelligent service robots. In Tussyadiah et al.'s (2020) experiment U.S.-based participants participated in an online scenario where they were instructed to plan a cruise ship featuring robot bartenders. Regardless of robot type (e.g., robot arm, humaniod), Tussyadiah et al. (2020) found that people who had negative attitudes towards robots trusted these robots less.

While experiment 1 demonstrated a significant main effect of social norms and perceived agency on trust, the purpose of experiment 2 is to explore whether the relationship between social norms, perceived agency and trust was dependent upon people's attitudes towards robots. While considering that negative attitudes towards robots is an element of a psychological construct contributing to perceptions of robots, it is expected that the effect of social norms or perceived agency is related to trust above and beyond the effect of negative attitudes on trust. Experiment 2 aims to rule out the notion that there is a relationship between negative attitudes towards robots and trust suggesting that people who rated the robots low in trust or low in perceived agency had a predisposed negative attitude towards robots.

Methodology

Participants

Two hundred sixty-nine participants (160 males, 106 females, 2 prefer not to answer, 1 other) participated in this study. Of the 269 participants, 27 (10%) participants were excluded for missing the attention check question (22 participants) or providing incomplete responses (5 participants). The final sample included 242 (138 males, 101 females, 2 prefer not to answer, 1 other) with ages ranging from 19 to 77 years old (M = 37, SD = 10.81). Participants were recruited from Amazon's Mechanical Turk online platform and were invited to complete the online survey in exchange for pay. The number of participants needed to detect a medium effect size with 80% power and $\alpha = 0.05$ was n = 204. The Power analysis was calculated *a priori* using G*Power Software, t = 1.65.

Design

A between-subjects design was used for this study. Each participant was randomly assigned to watch one 30-second video associated with each of the study conditions. After finding no differences in participant scores for the line and elevator scenarios (see Experiment 2 Results section), norm-behaviors were treated as one factor with two levels consisting of the norm-conforming and violation conditions. Table 2 shows the number of participants in each condition prior to data cleaning.

Scenarios	Conditions	
	Norm-Conforming	**Norm-Violation**
Line	68 Participants	64 Participants
Elevator	70 Participants	67 Participants

Table 2: Number of participants in each experimental condition (Experiment 2)

Materials

Stimuli The stimuli for experiment 2 were identical to experiment 1.

Measures The measures administered in experiment 2 included the perceived agency questionnaire (Korman et al., 2019) and MDMT-v2 (Ullman & Malle, 2019) from experiment 1. Additionally, negative attitudes towards robots were assessed using Normua, Suzuki, Kanda and Kato's (2006) Negative Attitude toward Robots Scale (NARS). Normua et al.'s (2006) scale consisted of 14 Likert items were participants chose from one of five grades for item (1: I strongly disagree, 2: I disagree, 3: Undecided, 4: I agree, 5: I strongly agree). NARS consisted of three sub-scales where "S1:Negative Attitude toward Situations of Interaction with Robots" (e.g., I would feel paranoid talking with a robot) consisted of six items, "S2:Negative Attitude toward Social Influence of Robots" (e.g., Something bad might happen if robots developed into living beings) consisted of five items. The third subscale- "S3:Negative Attitude towards Emotions in Interaction with Robots" (e.g., I would feel relaxed talking with robots) consisted of three reverse items. Participant scores were computed as the average of all 14 Likert scale items.

Procedure

Experimental Procedure To begin the experiment , participants responded to the NARS questionnaire. The following procedures for experiment 2 were identical to experiment 1.

Experiment 2 Results

A two-sample t-test was conducted to inspect differences in participant scores for perceived agency and trust between the line and elevator scenarios. The t-test did not show significant differences for perceived agency $t(240) = -1.74$, $p=.083$ or trust $t(239) = 1.14$, $p=.254$. As a result, I collapsed scores in the two scenarios. Next, I examined significant differences in participant means scores on the norm-conforming and norm-violating conditions with perceived agency and trust to replicate the results of experiment 1. A t-test comparing means on perceived agency revealed participants in the norm-conforming condition ($M = 5.66$; $SD = 1.09$), attributed the robot as having more perceived agency compared to participants in the violation condition ($M = 5.12$; $SD = 1.41$), $t(240) = 3.35$, $p=.001$ $d = 0.43$ 95% CI [0.17 – 0.69]), $MSE = 1.57$, *partial eta squared* = 0.05. Similarly, a t-test revealed that people felt a greater sense of trust in a norm-conforming robot ($M = 5.33$; $SD = 1.18$), compared to a norm-violating robot ($M = 4.53$; $SD = 1.57$), $t(239) = 4.49$, $p<.001$, $d = 0.58$ 95% CI [0.32 – 0.84]), $MSE = 2.00$, *partial eta squared* = 0.08.

Next, experiment 2 examined the linear relationship between perceived agency and trust scores by condition. Identical to experiment 1, a Pearson correlation coefficient revealed that there was a strong positive correlation between perceived agency and trust scores $r(239) = 0.32$, $p<.001$ (figure 2)[5].

For the final analysis, Pearson correlation coefficients were performed to examine whether there was a linear relationship between negative attitudes (dependent variable) and trust scores (independent variable), or negative attitudes (dependent variable) and perceived agency (independent variable) scores. As suggested by figure 3, there was generally no relationship between negative attitudes and trust scores $r(238) = 0.05$, $p=.478$[6], or between negative attitudes and perceived agency scores $r(239) = 0.02$, $p=.785$[7]. Furthermore, a regression was performed to examine whether perceived agency and negative attitudes (independent variables) impacted trust scores (dependent variable). The regression revealed that perceived agency and negative attitudes did not statistically significantly interact, $b=0.14$, $t(236) = 1.51$, $p = .132$[8].

---

[5] Spearman's rank correlation showed similar results $rs(239) = 0.34$, $p<.001$.
[6] Spearman's rank correlation showed similar results $rs(239) = 0.10$, $p=.111$.
[7] Spearman's rank correlation showed similar results $rs(239) = -0.02$, $p=.748$.
[8] A regression revealed similar results for perceived agency, negative attitudes and condition, $b=0.23$, $t(232) = 1.20$, $p = .232$.

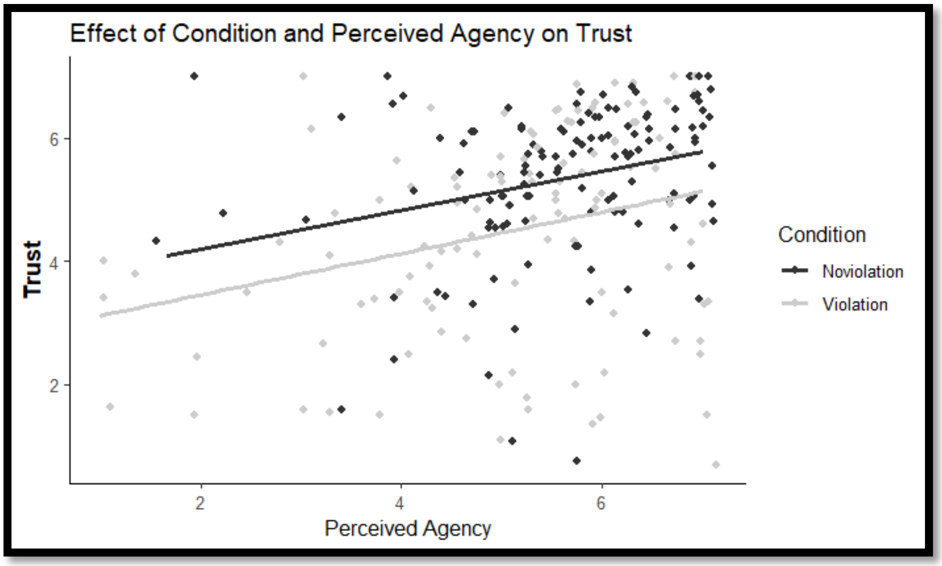

Figure 2: Pearson's correlation examining the relationship between Perceived Agency and Trust scores in norm-conforming and norm-violation conditions (Experiment 2)

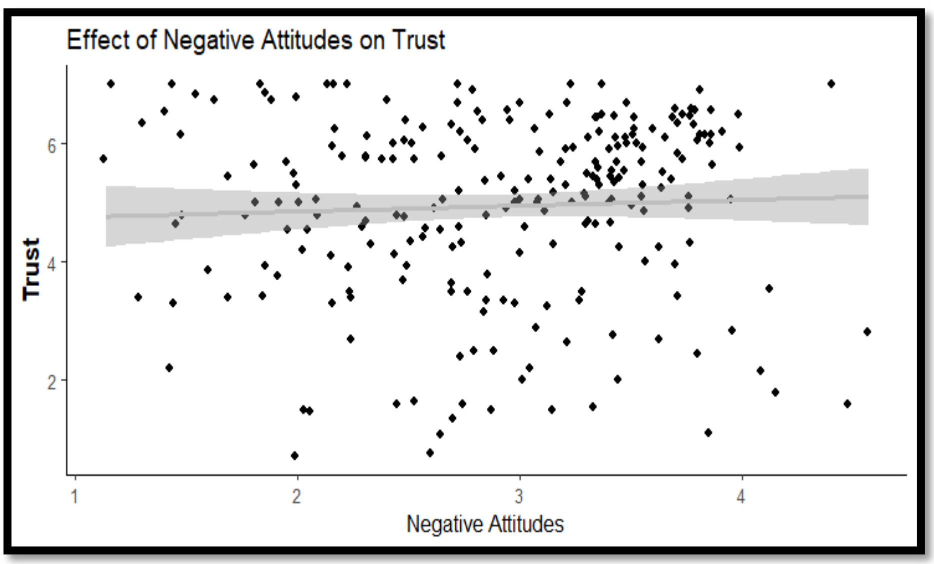

Figure 3: Pearson's correlation examining the relationship between Negative Attitudes and Trust scores (Experiment 2)

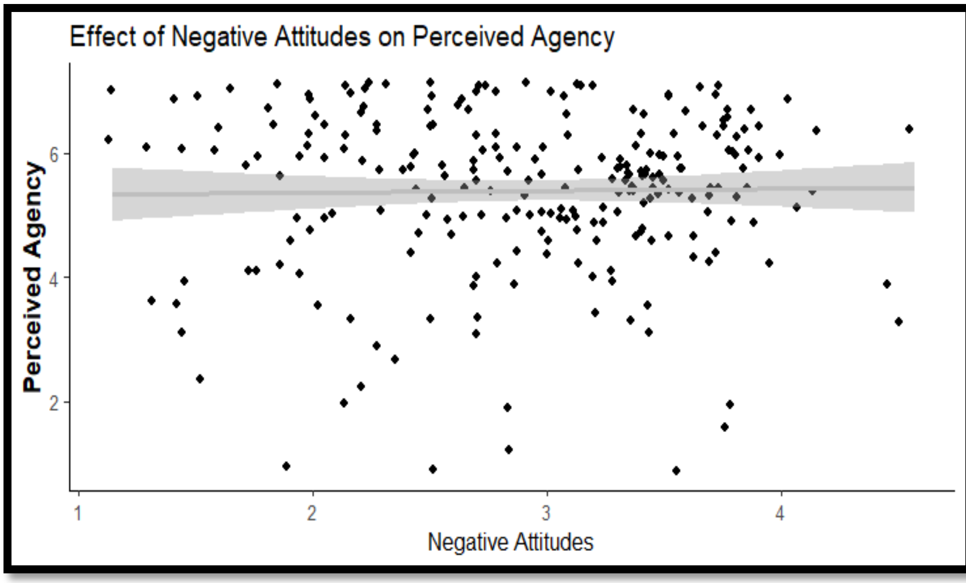

Figure 4: Pearson's correlation examining the relationship between Negative Attitudes and Perceived Agency scores (Experiment 2)

Experiment 2 Discussion

While analyzing whether negative attitudes correlated with trust scores or perceived agency scores, experiment 2 found that negative attitudes was generally not related to evaluations of trust in the robot or how much perceived agency people attributed to the robot. Furthermore, while analyzing the influence of perceived agency and negative attitudes on trust scores results from experiment 2 found no interaction between perceived agency and negative attitudes on trust. That is, people's evaluations of how much perceived agency a robot has and their respective levels of trust in robots was not impacted by whether they felt negatively about robots in general. Furthermore, results from experiment 2 demonstrated a conflicting pattern concerning perceived agency and negative attitudes towards robots on trust while comparing what Tussyadiah et al. (2020) found. Tussyadiah et al.'s (2020) experiment found evidence for situations where negative attitudes alone influence trust in robots. Also, experiment 2 found evidence for situations where negative attitudes does not influence trust in robots.

GENERAL DISCUSSION

To examine the relationship between trust and perceived agency, experiment 1 examined the effects of people's perception of a robot's compliance to social norms on their evaluation of a robot's perceived agency and trust. Consistent with the hypotheses, experiment 1 showed that when a robot followed social norms, it was perceived as having more agency (Korman et al., 2019), and participants were more willing to trust it compared to a robot that violated social norms. This result was also comparable to Ullman et al. (2014) who showed that a humanoid (NAO) robot's honest behavior during a game of Battleship increased people's perception of its perceived intentionality and perceived trustworthiness compared to dishonest behaviors.

Experiment 2 was conducted to assess whether the relationship found in experiment 1 was dependent upon people' s negative attitudes towards robots while replicating results from experiment 1. As expected, experiment 2 replicated the results found in experiment 1. Furthermore, results from experiment 2 showed when people trusted the robot based on the robot's compliance (or resistance) to social norms and the amount of perceived agency it had was not influenced by whether people felt negatively about robots. This result was different from Miller et al. (2021) who found that the relationship between negative attitudes towards robots and dynamic learned trust is mediated by initial learned trust of a humanoid domestic robot (TIAGo).

Most significantly, results from experiments 1 and 2 support the hypothesis concerning the relationship between perceived agency and trust such that when perceptions of how much perceived agency a robot has increases, people are willing to trust it. The strong positive relationship between perceived agency and trust exceeds beyond a robot's compliance to social norms, and negative attitudes towards robots. This result was also consistent with Brink and Wellman's (2020) research suggesting that children trusted information more from a robot that was perceived as having psychological agency, regardless of whether the robot was accurate. Brink and Wellman's (2020) research supports the hypotheses made in this thesis, while alluding to the consequences of designing a robot tutor (or equivalent) low in perceived agency such that when a child perceives a robot low in perceived agency, they are less likely to trust it.

Limitations and Future Directions

The previous experiments illuminate crucial insight for designers who should place more emphasis on deploying robots and automated systems, specifically in high-stake environments, perceived as having high agency (e.g., health care, military) to establish relationships built on trust. This is particularly interesting because these findings suggest that enhancing trust between humans and robots is possible even when they are less capable or violate social norms. Perhaps this phenomenon occurs because, a robot with a high degree of certain perceived agency cues (e.g., perceives the environment, navigates in the environment) takes higher precedence than a robot's actual capabilities or morality.

However, other environments also reliant upon trust (e.g., healthcare) may not benefit from an incompetent robot high in perceived agency. Ideally, designs for healthcare robots should ensure that a robot has relatively low perceived agency so that the robot's actual performance takes priority in daily interactions with human counterparts. Of note, social robots are not limited to the design utilized in the previous experiments or healthcare robots. Results from experiments 1 and 2 strengthen previous theories concerning the significant influence of perceived agency in interactions with robots and found a strong positive relationship between trust and perceived agency. Furthermore, results from experiment 1 and 2 suggest that the significant influence of perceived agency on trust evaluations is particularly evident in simplistic yet regularly occurring interactions where robots are sharing an elevator or joining a line with people. Future HRI research is necessary with specific emphasis on executable tools for measuring perceived agency for a diversified set of robots. Therefore, future work should manipulate various perceived agency cues and experimentally test its effects on trust and related constructs of interest.

APPENDIX A - MULTI-DIMENSIONAL MEASURE OF TRUST (MDMT V2)

Please rate the robot using the scale from 0 (Not at all) to 7 (Very). If a particular item does not seem to fit the robot in the situation, please select the option that says "Does Not Fit."

	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit
reliable									
competent									
ethical									
transparent									
benevolent									
predictable									
skilled									
principled									
genuine									
kind									
	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit
dependable									
capable									
moral									
sincere									
considerate									
consistent									
meticulous									
has integrity									
candid									
has goodwill									
	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit

APPENDIX B - NEGATIVE ATTITUDE TOWARD ROBOTS SCALE (NARS)

Table 1: The English Version of Negative Attitude toward Robots Scale and the Subscales that Each Item Is Included

Item No.	Questionnaire Items	Subcale
1	I would feel uneasy if robots really had emotions.	**S2**
2	Something bad might happen if robots developed into living beings.	**S2**
3	I would feel relaxed talking with robots.*	**S3**
4	I would feel uneasy if I was given a job where I had to use robots.	**S1**
5	If robots had emotions, I would be able to make friends with them.*	**S3**
6	I feel comforted being with robots that have emotions.*	**S3**
7	The word "robot" means nothing to me.	**S1**
8	I would feel nervous operating a robot in front of other people.	**S1**
9	I would hate the idea that robots or artificial intelligences were making judgments about things.	**S1**
10	I would feel very nervous just standing in front of a robot.	**S1**
11	I feel that if I depend on robots too much, something bad might happen.	**S2**
12	I would feel paranoid talking with a robot.	**S1**
13	I am concerned that robots would be a bad influence on children.	**S2**
14	I feel that in the future society will be dominated by robots.	**S2**

(*Reverse Item)

REFERENCES

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, *90*, 363–371. https://doi.org/10.1016/j.chb.2018.08.028

Banks, J. (2021). Good Robots, Bad Robots: Morally Valanced Behavior Effects on Perceived Mind, Morality, and Trust. *International Journal of Social Robotics*, *13*(8), 2021–2038. https://doi.org/10.1007/s12369-020-00692-3

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a General Logicist Methodology for Engineering Ethically Correct Robots. *IEEE Intelligent Systems*, *21*(4), 38–44. https://doi.org/10.1109/MIS.2006.82

Brink, K. A., & Wellman, H. M. (2020). Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology*, *56*(7), 1268–1277. https://doi.org/10.1037/dev0000884

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can You Trust Your Trust Measure? *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 92–100. https://doi.org/10.1145/3434073.3444677

De Graaf, M., & Malle, B. (2017). *How People Explain Action (and Autonomous Intelligent Systems Should Too)* (AAAI Fall Symposium Series). https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/vie

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Dennett, D. C. (1978). Current Issues in the Philosophy of Mind. *American Philosophical Quarterly*, *15*(4), 249–261.
16009/15283

Falcone, R., Castelfranchi, C., Cardoso, H. L., Jones, A., & Oliveira, E. (2013). Norms and Trust. In S. Ossowski (Ed.), *Agreement Technologies* (pp. 221–231). Springer Netherlands. https://doi.org/10.1007/978-94-007-5583-3_15

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *63*(7), 1196–1229. https://doi.org/10.1177/0018720820922080

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, *4*, 3559–3564. https://doi.org/10.1109/IROS.2004.1389967

Kim, B., Haring, K., Schellin, H., Oberley, T., Patterson, K., Phillips, E., de Visser, E., & Tossell, C. (2020). *How Early Task Success Affects Attitudes Toward Social Robots* (p. 289). https://doi.org/10.1145/3371382.3378241

Korman, J., Harrison, A., McCurry, M., & Trafton, G. (2019). Beyond Programming: Can Robots' Norm-Violating Actions Elicit Mental State Attributions? *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 530–531. https://doi.org/10.1109/HRI.2019.8673293

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Levin, D. T., Adams, J. A., Saylor, M. M., & Biswas, G. (2013). A transition model for cognitions about agency. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 373–380. https://doi.org/10.1109/HRI.2013.6483612

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), 12:1-12:25. https://doi.org/10.1145/1985347.1985353

Miller, L., Kraus, J., Babel, F., & Baumann, M. (2021). More Than a Feeling—Interrelation of Trust Layers in Human-Robot Interaction and the Role of User Dispositions and State Anxiety. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.592711

Moon, Aj., Troniak, D. M., Gleeson, B., Pan, M. K. X. J., Zheng, M., Blumer, B. A., MacLean, K., & Croft, E. A. (2014). Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 334–341. https://doi.org/10.1145/2559636.2559656

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 35–40. https://doi.org/10.1109/ROMAN.2004.1374726

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437–454. https://doi.org/10.1075/is.7.3.14nom

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. https://doi.org/10.1145/975817.975844

Phillips, E., Ullman, D., de Graaf, M. M. A., & Malle, B. F. (2017). What Does A Robot Look Like?: A Multi-Site Examination of User Expectations About Robot Appearance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1215–1219. https://doi.org/10.1177/1541931213601786

Pollini, A. (2009). A theoretical perspective on social agency. *AI & SOCIETY*, *24*(2), 165–171. https://doi.org/10.1007/s00146-009-0189-2

Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., & Taylor, G. (2008). The Effect of Automation Reliability on User Automation Trust and Reliance in a Search-and-Rescue Scenario. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1340–1344. https://doi.org/10.1177/154193120805201908

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!!: An interaction with a cheating robot. *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction - HRI '10*, 219. https://doi.org/10.1145/1734454.1734546

Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., & Broadbent, E. (2014). Does the Robot Have a Mind? Mind Perception and Attitudes Towards Robots Predict Use of an Eldercare Robot. *International Journal of Social Robotics*, *6*(1), 17–32. https://doi.org/10.1007/s12369-013-0186-y

Tussyadiah, I., Zach, F., & Wang, J. (2020). Do travelers trust intelligent service robots? *Annals of Tourism Research*, *81*, 102886. https://doi.org/10.1016/j.annals.2020.102886

Ullman, D., Leite, I., Phillips, J., Kim-Cohen, J., & Scassellati, B. (n.d.). *Smart Human, Smarter Robot: How Cheating Affects Perceptions of Social Agency*. 7.

Ullman, D., & Malle, B. F. (2019). Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 618–619. https://doi.org/10.1109/HRI.2019.8673154

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100. https://doi.org/10.1016/j.cognition.2010.04.003

van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, P. (2014). Do Robot Performance and Behavioral Style affect Human Trust?: A Multi-Method Approach. *International Journal of Social Robotics*, *6*(4), 519–531. https://doi.org/10.1007/s12369-014-0231-5

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, *12*(3), 243–250. https://doi.org/10.1007/s10676-010-9232-8

Yasuda, S., Doheny, D., Salomons, N., Sebo, S. S., & Scassellati, B. (2020). Perceived Agency of a Social Norm Violating Robot. *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://par.nsf.gov/biblio/10284325-perceived-agency-social-norm-violating-robot

Zhao, X., Phillips, E., & Malle, B., (2019). Beyond Anthropomorphism: Differentiated Inferences About Robot Mind From Appearance. *Na – Advances in Consumer Research Volume 47, eds.*, 352-358. http://www.acrwebsite.org/volumes/2551817/volumes/v47/NA-47

BIOGRAPHY

Chelsea Frazier is an active-duty Officer in the U.S. Army. She was born in Yokosuka, Japan and raised primarily in Gaithersburg, Maryland. Chelsea earned her Bachelor of Science in Business Administration degree from Morgan State University in Baltimore, MD in 2012 in which she also received her commissioning through the Reserve Officers' Training Corps (ROTC). Chelsea has served in a multitude of positions to include an Executive Officer, Company Commander, and Division Network Officer throughout her ten years of service. In 2018, Chelsea earned a Master's in Information Technology Management from Webster University in St. Louis, MO.