Perceived Agency Changes Performance and Moral Trust in Robots

Chelsea R. Frazier
(chelsea.r.frazier3.mil@army.mil)
United States Military Academy
West Point, NY

J. Malcolm McCurry
(malcolmmccurry@gmail.com)
Arcfield
Centreville, VA

Kevin Zish
(zish.kevin@gmail.com)
Global Systems Technologies
Arlington, VA

J. Gregory Trafton
(greg.trafton@nrl.navy.mil)
Naval Research Laboratory
Washington, DC

Abstract

What is the relationship between trust and perceived agency? The present study 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, performance trust, and moral trust. Participants reported a norm-conforming robot to have higher perceived agency and a greater sense of trust than a robot that violated social norms. We also found that perceived agency, regardless of how much a robot followed norms, was correlated with trust. We interpret this finding as evidence that as people see a robot as having agency, they trust it more.

Keywords: human-robot interaction; moral trust; perceived agency; performance trust; social norms; trust

Introduction

One of the most important questions within human-robot-interaction is "why do people trust a robot?" In this paper, we suggest 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.

We highlight that while Parasuraman and Miller focused on etiquette, we consider etiquette a form of social norm following: people or systems who follow rules of etiquette are following established social norms. We seek in this report to directly test the hypothesis that when an automated system (a robot in our 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), we expect to find that when an automated systems follows social norms, it will lead to increased perceived agency which in turn will lead to the system being trusted.

In the remainder of this introduction, we discuss 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).

Researchers Hancock, Kessler, Kaplan, Brill, and Szalma (2020) 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 have 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

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

"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 human-robot trust (HRT) remains a challenge due to difficulty with replication.

Also, interesting research by Kim, Wen, de Visser, Zhu, Williams, and 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. Therefore, to develop a reliable hypothesis for perceived agency and trust, next we summarize human-robot interaction (HRI) literature with the specific emphasis on how researchers conceptualized perceived agency.

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) experiment demonstrated a replicable and succinct approach for future contributions.

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 perceived 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 in attempt to answer how much agency it has, and conditions in which 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. 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 social norms 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 perceived 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 alludes to an additional variable contributing to how people perceive, think about, and interact with norm compliant (or non-compliant) robots.

In general, it appears that robots that follow social norms are more likely to have a higher degree of perceived agency than robots that do not follow social norms (Korman et al., 2019; Yasuda, Doheny, Salomons, Strohkorb & Scassellati, 2020). Given other research that suggests that perceived agency and trust are also related (Kim et al., 2019; Parasuraman & Miller, 2004), we will explicitly test this link in the following experiment.

Stimuli Validation Experiment

Norm Study

We first conducted a norm study to validate the video stimuli from the Korman et al. (2019) experiment. We adopted their 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 through 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 *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 the line where a group of people were occluded around a corner. In the *elevator*, 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. Thus, our investigation will focus on the *line* and *elevator* scenarios with the pure *norm-violating* and *norm-conforming* conditions.

Methodology

Participants

The sample included 160 men and women between 23 to 70 years old from various ethnicities and generally holding a bachelor's degree. Participants were recruited from the 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 = 90. The Power analysis was calculated *a priori* using G*Power Software, $f = 0.30$.

Procedure

We set up a between subjects' experiment where we treated norm-behaviors as one factor with two levels consisting of the norm-conforming and violation conditions.

To begin the experiment, participants were instructed to watch the video of the DRC-HUBO robot. Upon the completion of each video, participants utilized a free-text box to respond to an open-ended question regarding the description of the robot's behavior. Following the initial question, participants responded to three questions capturing the essence of perceived agency, and the 20-item Multi-Dimensional Measure of Trust (MDMT-v2) (Ullman & Malle, 2019). After responding to the two questionnaires, participants were prompted with a free-text box for general feedback and debriefed on the study. One foil question was inserted to verify the accuracy of participant responses.

Measures

Following the video, participants responded to six questions capturing the essence of perceived agency (Korman et al., 2019). The six Likert-style questions on a 7-point rating scale (from 1 = "not at all aware" to 7 = "very much aware") and question six rated from 1 = "definitely not just a tool" to 7 = "definitely just a tool" included whether the robot performed the behavior *intentionally, aware* of engaging in the behavior, *chose* to, *wanted* to, could have *chosen not* to, and lastly, "rate if the robot is just a tool".

Ullman and Malle's (2019) Multi-Dimensional Measure of Trust (MDMT v2) demonstrated a reliable link to measure people's evaluations of trust in the context of social norms.

Therefore, Performance Trust and Moral Trust was assessed using MDMT v2. Ullman and Malle's (2019) trust dimensions were grouped by two main factors consisting of performance trust (e.g., reliable and competent) and moral trust (e.g., ethical, transparent, benevolent) where participants rated the robot on a scale from 0 (Not at all) to 7 (Very) to a total of 20 Likert scale items.

Perceived Agency Results

Of the 160 participants who completed the initial screening survey, 38 (24%) participants were excluded for missing the attention check question or providing incomplete responses. The remaining 122 participants from the norm-conforming (n

= 64) and norm-violating (n = 58) conditions contributed to this study.

To assess the relationship between perceived agency, performance trust and moral trust among two conditions (norm-conforming and violation), we first analyzed perceived agency, performance trust, and moral trust independently. Consistent with Korman et al. (2019), we performed three separate ANOVAs comparing means on the rating scales of interest for the *intentionality*, *awareness,* and *want* questions (Figure 1)[1].

There was a significant difference for the *intentionality* question where participants in the norm-conforming condition (*M* = 6.14; *SD* = 1.19), attributed the robot as having more intentionality compared to participants in the violation condition (*M* = 5.41; *SD* = 1.76), *F* (1, 120) = 7.26, *p* <.01, *d* = 0.49 95% CI [0.12 – 0.85]).

Also, there was a significant difference for the *awareness* question where participants in the norm-conforming condition (*M* = 5.83; *SD* = 1.23), attributed the robot as having greater awareness compared to participants in the violation condition (*M* = 4.88; *SD* = 1.96), *F* (1, 120) = 10.49, *p* <.01, *d* = 0.59 95% CI [0.22 – 0.95]).

However, there was not a significant difference between participants in the norm-conforming (*M* = 5.41; *SD* = 1.46) and the violation conditions (*M* = 5.03; *SD* = 1.71) for the *want* question *F* (1, 120) = 1.69, *p* <.01, *d* = 0.24 95% CI [-0.12 – 0.60]).

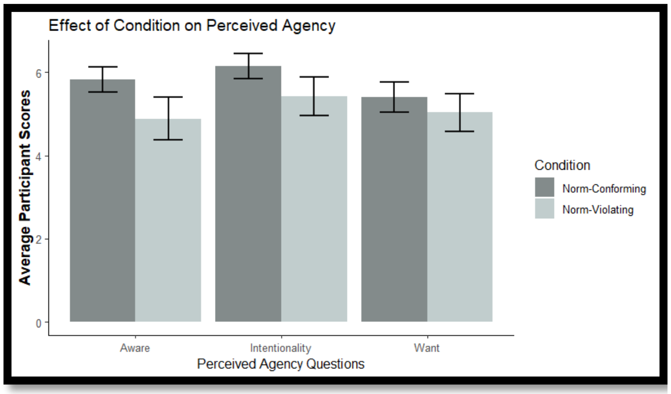

Figure 1: Perceived Agency Questionnaire by Condition

Perceived Agency Discussion

Our analyses on perceived agency replicated Korman et al. (2019), which we found that people rated a robot that followed social norms higher on perceived agency (intentionality and awareness) than a robot that violated social norms. This result is also consistent with Yasuda et al. (2020), who showed that norm violations did not increase people's perception of perceived agency. These results suggest a strong positive relationship between social norms and perceived agency.

Trust Results

Effect of Condition on Performance Trust and Moral Trust

Next, we performed two separate ANOVAs and found that people felt a greater sense of performance trust when the robot did not violate a norm (*M* = 5.71; *SD* = 1.14), compared to situations when the robot violated a norm (*M* = 4.61; *SD* = 1.62), *F* (1, 118) = 18.59, *MSE* = 1.90, *p <.01, partial eta squared* = 0.14 (Figure 2).

Also, our results revealed a similar pattern for moral trust where people felt a greater sense of moral trust when the robot did not violate a norm (*M* = 5.07; *SD* = 1.73), compared to situations when the robot violated a norm (*M* = 3.25; *SD* = 2.30), *F* (1, 113) = 23.26, *MSE* = 4.02 *p <.01, partial eta squared* = 0.17 (Figure 3).

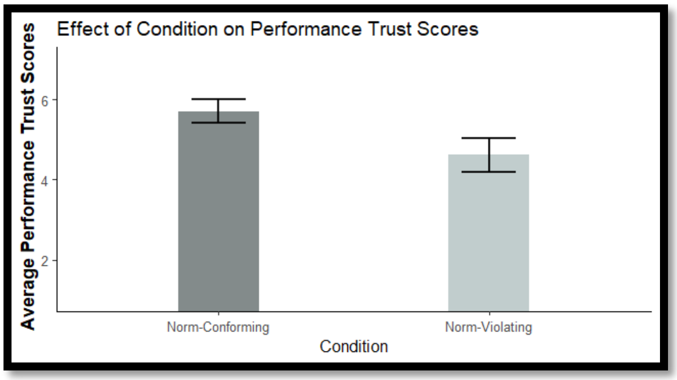

Figure 2: Mean perceptions of performance trust. (Error bars show 95% confidence intervals)

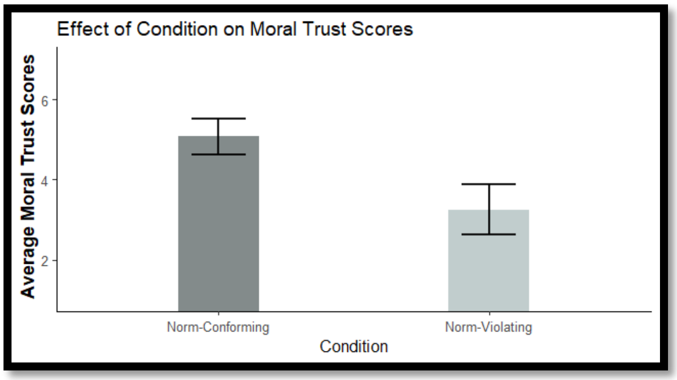

Figure 3: Mean perceptions of moral trust. (Error bars show 95% confidence intervals)

Trust Discussion

While comparing mean perceptions of performance trust and moral trust, we found that people rated a robot that followed social norms higher than a robot that violated social norms. Our result is consistent with Ullman et al. (2014) who showed that honest behaviors increased people's perception of perceived trustworthiness compared to dishonest behaviors.

[1] Combing all three questions showed similar results.

These results also suggest a strong positive relationship between social norms and both performance trust and moral trust.

Perceived Agency and Trust Results

Effect of Perceived Agency on Performance and Moral Trust

For our final analyses, we examined the relationship between perceived agency and trust while collapsing across condition. As suggested by Figure 4, there is a positive relationship between the amount of perceived agency that a robot is perceived to have and the amount that people trusted its performance, r = 0.34, p < .01 Interestingly, there is also a positive relationship between the amount of perceived agency a robot has and the amount of moral trust, r=0.24, p < .01 (Figure 5).

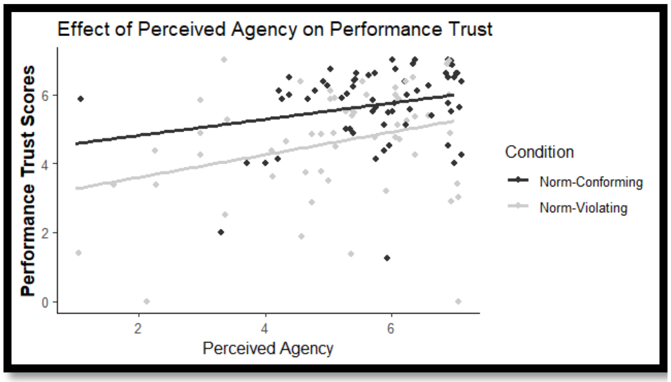

Figure 4: Pearson's correlation examining the relationship between Perceived Agency and Performance Trust

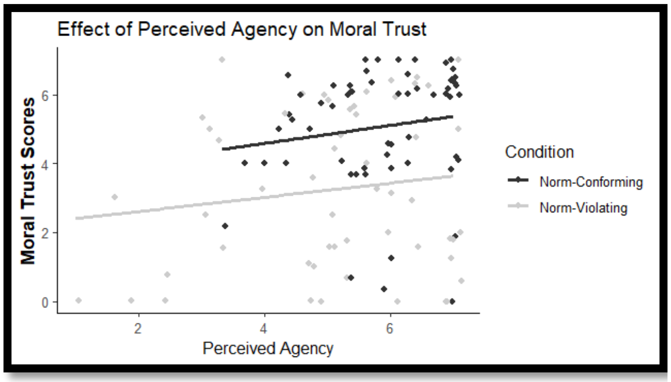

Figure 5: Pearson's correlation examining the relationship between Perceived Agency and Moral Trust

Perceived Agency and Trust Discussion

While performing a correlation with perceived agency, performance trust and moral trust, we found that both performance trust and moral trust are positively correlated with perceived agency. This result is consistent with our hypothesis specifically concerning the relationship between perceived agency and trust: that people will trust a robot more (at least partially) depending on how much perceived agency the robot has.

General Discussion

This experiment has examined the effects of people's perception of a robot's compliance (and resistance) to social norms on their evaluation of a robot's perceived agency, performance trust, and moral trust. Consistent with our hypotheses, participants in the norm-conforming condition continually attributed higher evaluations of perceived agency, performance trust and moral trust compared to participants in the violation condition.

Therefore, we found evidence for a strong positive relationship between social norms and perceived agency. Our results are consistent with Uttich and Lombrozo's (2010) research suggesting that an entity engaging in social ways was expected to respect social norms. When this was met, people felt as though the machine/robot was more intentional. Also, our results illustrated a strong positive relationship between social norms and trust. This result was consistent with Falcone, Castelfranchi, Cardoso, Jones and Oliveria's (2013) research suggesting that people are willing to trust a machine/robot that fulfills norm expectations.

Most significantly, our findings support our hypothesis concerning the relationship between perceived agency and trust whereas perceived agency positively correlated with performance trust and moral trust. Therefore, our results imply that when perceptions of how much perceived agency a robot has increases, people are willing to trust it. This is a crucial insight for designers who should place more emphasis on deploying robots and automated systems, specifically in high-stake environments that are perceived as having high agency (e.g., health care, military) to establish relationships built on trust. Furthermore, this research illuminates the necessary research concerning the measurement, replication, and wide publication of perceived agency among a broad range of robots and machines.

Of note, we believe an additional factor contributes to the discrepancy between social norms, perceived agency and trust which might explain why social norms effect people's evaluation of perceived agency and trust independently, but not conjointly. Perhaps, people's preconceived notions and interactions (or the lack thereof) with robots influences their perceptions of a robot's perceived agency to begin with (Stafford, MacDonald, Jayawardena, Wegner, & Broadbent, 2013). As a result, people with positive feelings towards robots might be more likely to perceive robots with higher perceived agency compared to people who do not. Further research is necessary with the specific intention of analyzing the effects of people's attitudes towards robots on perceived agency and trust through other realistic settings. As robots continue to coexist in familiar environments alongside humans, the more precise our interpretation of human behavior must be for engaging robots in social ways.

Continued research helps to investigate future implications for social, cultural, and political contexts.

Acknowledgments

This work was funded by the Office of the Naval Research to JGT. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies of the U.S. Armed Forces.

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 Valenced 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

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

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

Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308. https://doi.org/10.1109/HRI.2013.6483603

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

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

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

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

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

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