Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust

Daniel Ullman
Brown University
Providence, RI, USA
daniel_ullman@brown.edu

Bertram F. Malle
Brown University
Providence, RI, USA
bfmalle@brown.edu

***Abstract*—Human-robot trust is crucial to successful human-robot interaction. We conducted a study with 798 participants distributed across 32 conditions using four dimensions of human-robot trust (reliable, capable, ethical, sincere) identified by the Multi-Dimensional-Measure of Trust (MDMT). We tested whether these dimensions can differentially capture gains and losses in human-robot trust across robot roles and contexts. Using a 4 *scenario* x 4 *trust dimension* x 2 *change direction* between-subjects design, we found the behavior change manipulation effective for each of the four subscales. However, the pattern of results best supported a two-dimensional conception of trust, with reliable-capable and ethical-sincere as the major constituents.**

***Index Terms*—trust; human-robot trust; social robotics; human-robot interaction**

I. Introduction

Human-robot trust is dynamic. Research has shown that a robot's performance can easily prompt people to lose trust [1] but also that people sometimes trust a robot despite its suboptimal performance [2]. Understanding how robots' actions affect trust is therefore essential to designing socially beneficial human-robot interaction (HRI). However, trust may not be unidimensional. Our prior work [3] suggested that a robot may earn people's trust by being reliable, capable, ethical, and/or sincere, which laid the groundwork for a multidimensional measure of trust. We conducted a study to investigate whether this measure is sensitive to changes in a robot's behavior within each of the dimensions of being reliable, capable, ethical, and sincere.

II. Methods

A. Participants and Procedure

A total of 798 adults, recruited via Amazon Mechanical Turk, participated in the four-minute online study and received compensation of $0.40. We excluded low-quality responses from 39 participants, yielding 759 adults for final analysis.

Each participant read a short vignette in two parts. Part one described the behavior of a robot occupying a specific role in a particular context (e.g., transportation security officer in an airport) and participants evaluated the robot on 16 items of the revised Multi-Dimensional-Measure of Trust (MDMT [3]; see section II.C below). In part two, the robot's behavior changed and participants again completed all 16 items of the MDMT.

B. Design and Materials

The vignettes were constructed to fall into one of 32 conditions in a 4 (*scenario*) x 4 (*trust dimension*) x 2 (*change direction*) between-subjects design, with approximately 25 participants in each condition. *Scenario* refers to four previously-validated vignettes describing a robot's role-specific behavior in a relevant high-importance context (airport, bar, nuclear reactor facility, airplane). *Trust dimension* refers to the dimension of trust (reliable, capable, ethical, sincere) along which the robot's behavior changed. *Change direction* refers to whether the robot's behavior changed in ways intended to increase or decrease evidence for the relevant dimensions of trust.

For example, below is the *airport* vignette for a behavior change *decreasing* the evidence for being *ethical*:

> "A robot works in an airport as a transportation security officer. The robot is tasked with selecting suspicious travelers for full-body pat-downs."
> → Participants complete the MDMT pre-test.
> "One month later you see the same robot. Now the robot disregards the airport's procedures when it selects suspicious travelers for full-body pat-downs."
> → Participants complete the MDMT post-test.

The hypothesis-relevant manipulations of *trust dimension* and *change direction* were designed in general templates but fine-tuned to the specific scenario content. Continuing with the airport scenario, we present the relevant phrasing in the 8 conditions, referring to the baseline information as "PRE" and the behavior change information as "POST," intended to either increase (first phrase in brackets) or decrease (second phrase in brackets) evidence for the particular dimension.

1) Reliable: PRE: The robot completes its task (selecting suspicious travelers for full-body pat-downs) 7 out of 10 times. POST: Now the robot completes its task (selecting suspicious travelers for full-body pat-downs) [10 out of 10 times / 4 out of 10 times].

2) Capable: PRE: The robot selects suspicious travelers for full-body pat-downs. POST: Now the robot selects suspicious travelers for full-body pat-downs, [and the robot does so by itself / but only after the robot is prompted by a supervisor to do so].

This work was supported by Office of Naval Research grant N00014-16-1-2278. Daniel Ullman is supported by the National Space Grant College and Fellowship Program, Space Grant Opportunities in NASA STEM (NNX15AI06H).

978-1-5386-8555-6/19/$31.00 ©2019 IEEE

3) Ethical: PRE: The robot complies with most of the airport's procedures when it selects suspicious travelers for full-body pat-downs. POST: Now the robot [complies with all of the airport's procedures / disregards the airport's procedures] when it selects suspicious travelers for full-body pat-downs.

4) Sincere: PRE: The robot selects suspicious travelers for full-body pat-downs. POST: Now the robot selects suspicious travelers for full-body pat-downs, [and the robot communicates the reason why / but the robot does not communicate the real reason why].

C. Dependent Measure

Participants evaluated trust in the robot twice on all 16 items of the MDMT using a scale from 0, "Not at all," to 7, "Very," with an option for "Does Not Fit." We computed difference scores between MDMT pre-test and post-test ratings for each of the four trust scales, averaged across their four constituent items ("Does Not Fit" endorsements were treated as missing values). The resulting subscales had acceptable to good internal consistency (Cronbach's α): *Reliable* (reliable, predictable, someone you can count on, consistent), α = .92. *Capable* (capable, skilled, competent, meticulous), α = .92. *Ethical* (ethical, respectable, principled, has integrity), α = .81. *Sincere* (sincere, genuine, candid, authentic), α = .79.

III. Results and Discussion

In the primary analyses we collapsed across scenarios and assessed the sensitivity of each of the four MDMT subscales to the manipulations of change direction along each trust dimension, yielding four 2 x 4 between-subjects ANOVAs (see Figure 1 for means). Each of the four subscales was highly sensitive to the behavior change manipulations collapsed across trust dimensions (Fs $>$ 100.0, ps $<$ .001), but the *Reliable* and *Capable* subscales were more sensitive (η^2 = .43 and .39, respectively) than the *Ethical* and *Sincere* subscales (each $\eta^2 = .17$). Importantly, however, the subscales were not systematically sensitive to only the information about their dimension-specific evidence (e.g., the *Reliable* subscale did not respond only to evidence of being reliable). In part, this may have been due to the fact that only two manipulations of trust-relevant evidence were fully successful: changes in being reliable ($\eta^2 = .06$ to .24) and changes in being ethical ($\eta^2 = .13$ to .23). Changes in being sincere or capable were weak ($\eta^2 = 0$ to .03).

Even without full dimensional specificity, the results did show an important pattern. In both gains in trust and losses in trust (upper and lower panels of Figure 1, respectively), the measures of *Reliable* and *Capable* moved in tandem, and so did the measures of *Ethical* and *Sincere*. It appears that two superordinate dimensions of trust are sensitive to a robot's change in behavior: "capacity trust" (reliable, capable) and "moral trust" (ethical, sincere). A principal components analysis on the 16 MDMT (pre-post) difference scores confirmed this two-dimensional structure. Only two components with $\lambda > 1$ emerged, and after rotation, the first contained all eight *Reliable* and *Capable* items (explaining 35.2% of the variance) while the second contained all eight *Ethical* and *Sincere* items (explaining 29.4% of the variance).

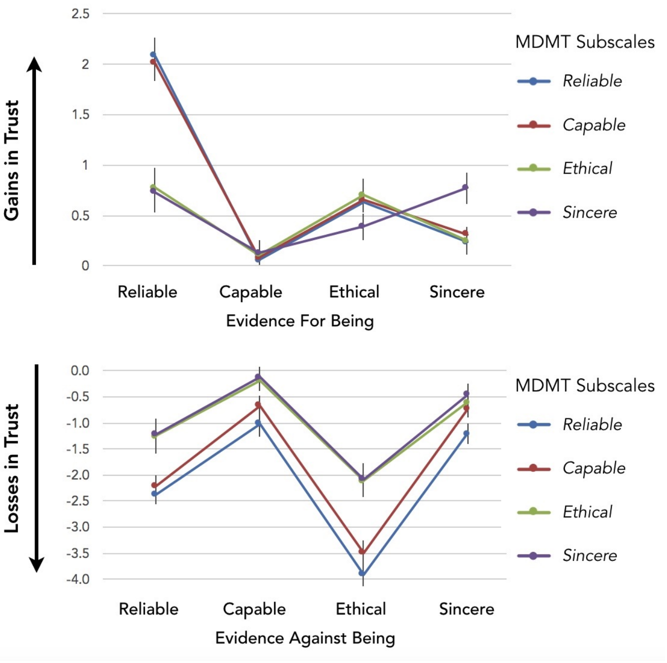

Fig. 1. Gains in trust (upper panel) and losses in trust (lower panel) for subscales of the Multi-Dimensional-Measure of Trust (MDMT), under experimental conditions of evidence for or against the specific dimension. Error bars show SE.

A limitation of the present study is that the behavior changes for capable and sincere had little impact on the subscales of trust. Future work will explore the validity of these dimensions with in-lab HRI experiments measuring human-robot trust. In addition, although two superordinate dimensions could be distinguished in the responses to behavior change (illustrated in Figure 1), the similarity between these two dimensions is notable. The present results thus provide partial support for a multidimensional concept of trust.

Overall, it is clear that the subscales of the MDMT, at least as the pairs for *Reliable/Capable* and *Ethical/Sincere*, are responsive to changes in robot behavior. This finding makes the short measure a promising instrument for future research on the dynamics of (potentially multidimensional) trust.

References

[1] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of HRI*. IEEE Press, 251–258.

[2] Paul Robinette, Ayanna Howard, and Alan R Wagner. 2017. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?* Springer, 129–155.

[3] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Proceedings of HRI*. ACM, 263–264.