What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust

Daniel Ullman
Brown University
Providence, Rhode Island
daniel_ullman@brown.edu

Bertram F. Malle
Brown University
Providence, Rhode Island
bfmalle@brown.edu

ABSTRACT

Research on trust in human-human interaction has typically focused on notions of vulnerability, integrity, and exploitation whereas research on trust in human-machine interaction has typically focused on competence and reliability. In this initial study, we explore whether these different aspects of trust can be considered parts of a multidimensional conception and measure of trust. We gathered 62 words from dictionaries and trust literatures and asked participants to evaluate the words as belonging to a "personal" meaning or a "capacity" meaning. Through an iterative process using Principal Components Analysis (PCA) and item analysis, we derived four components that capture the multidimensional space occupied by the concept of trust. The resulting four components yield four subscales of trust with five items each and α reliabilities as follows: *Capable* = .88, *Ethical* = .87, *Sincere* = .84, and *Reliable* = .72.

KEYWORDS

trust, human-robot trust, social robotics, human-robot interaction

ACM Reference Format:
Daniel Ullman and Bertram F. Malle. 2018. What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3173386.3176991

1 INTRODUCTION

Trust is a cornerstone of sustainable relationships between human agents, from child and caregiver to members of a search and rescue team. Robots, too, increasingly occupy roles that require trust, such as companions for older adults [1] or educational tutors [6]. We therefore must understand under what conditions people do or do not trust robots, and how calibrated such trust is. To that end trust must be measurable; but the measurement of trust is complicated by a conceptual division across different literatures.

The human-human interaction literature often studies trust as an agent's acceptance of vulnerability in an interaction or relationship, believing that the other will not exploit the vulnerability (e.g., [7]). By contrast, the human-automation literature often focuses on trust that is grounded in a machine's performance and reliability (e.g., [4, 5]). We may label the former aspect of trust "relational" trust and the second aspect "capacity" trust. The two aspects are neither redundant nor incompatible; but to study their precise relationship we must be able to measure them in distinct ways. To take a first step in such distinct measurement is the goal of this project.

Most measures of trust in human-robot interaction focus on the person's belief that the robot is capable of completing a given task [8]; most measures of trust in human-human interaction focus on integrity, loyalty, and other social-moral constructs [3]. We broadly culled the vocabulary used to describe and query trust in human-human, human-robot, and human-automation literature. Then we reached further, using dictionaries and thesauruses, to populate the semantic space of trust and its related constructs, ending up with 62 candidate words. With this broad vocabulary in hand we began the task to look for potential dimensions of trust—bundles of words that capture, and make measurable, distinct features of trust.

2 METHOD

A total of 45 adults, recruited via Amazon Mechanical Turk, participated in the five-minute online study for a compensation of $0.50. We excluded data from five participants who failed a comprehension check of the concepts grounding the dependent variable. We then examined inter-rater agreement. Five participants were outliers, as their ratings correlated with the whole group of raters at $r < .30$, and they were excluded from the analysis (in analogy to excluding items from scales that have item-total correlations of less than .30; see [2]). The final analysis included 35 adults, and their ratings correlated with the whole group on average at $r = .67$.

Participants were informed that they would read a series of 62 words and be asked to relate them to one of two ideas about trust, explained this way: "One idea is about trusting that an agent is capable of completing a task (referred to as "capacity trust"); the other idea is about trusting that an agent will not place you at risk (referred to as "personal trust")." After a comprehension check on the two ideas about trust, participants were asked to rate where they thought each word falls on a slider scale from "more similar to capacity trust" to "more similar to personal trust," with the middle indicating "equally similar to capacity trust and personal trust" (scale ranging from -50 for capacity trust to +50 for personal trust). For each word, participants had the option of selecting "Don't know word." Finally, participants were instructed that "Many of the words describe the presence of trust, while some describe the absence of trust; please place the words with the kind of trust they are most similar to, regardless of whether it is the presence or absence of this kind of trust."

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3176991

Item	Capable	Ethical	Sincere	Reliable
Capable	0.89			
Diligent	0.78			
Rigorous	0.77			
Accurate	0.77			
Meticulous	0.76			
Honest		0.83		
Principled		0.75	0.38	
Reputable		0.72	0.32	
Respectable		0.70	0.37	
Scrupulous		0.67	0.37	
Sincere			0.90	
Genuine			0.81	
Truthful			0.75	
Benevolent		0.30	0.70	
Authentic			0.63	0.37
Count on				0.82
Depend on				0.81
Reliable				0.57
Faith in				0.56
Confide in	-0.34			0.55
Explained variance (rotated)	19.2%	19.0%	16.9%	13.5%
5-item Cronbach's α	0.88	0.87	0.84	0.72

Figure 1: Final PCA with 4 components and 5 items each; loadings ≥ .30 displayed.

3 RESULTS

We conducted an initial Principal Components Analysis (PCA) of all 62 items, but an extraction criterion of $\lambda > 1$ allowed too many components, of which many had few and weakly loading items. After examining the intercorrelation matrix, we concluded that the item pool was too heterogeneous. We therefore took a step back and grouped items into semantically similar subsets. We arrived at four interpretable sets of items, roughly corresponding to general ethical qualities, relational qualities, competence, and reliability, as well as four unclassifiable items. We conducted item analyses on each of the four subsets, moving items in and out of sets to improve Cronbach's α reliabilities and item-total correlations. We noticed that, despite explicit instructions, negative words often confused participants (e.g., evaluating the word "inaccurate" as related to competence would be correct but is counterintuitive). Five negative items had to be removed, along with eight other items that did not fit any of the four sets.

Arriving at a reduced pool of 49 items, we performed a second PCA. An extraction criterion of $\lambda > 1$ still produced several components with little explained variance, but the scree plot and comparison on rotated solutions pointed to 4 strong, interpretable components. We selected items that loaded at least .50 on one component and less than .40 on any of the other components. We arrived at 32 items distributed over the four components. The interpretation of these scales became sharper—well represented by the labels *Capable* (8 items), *Ethical* (11 items), *Sincere* (6 items), and *Reliable* (7 items).

Next we resumed item analysis of the four subsets, removing items that did not enhance α reliability and aiming for subscales of 5 items each. We then conducted a final PCA of the resulting 20 items. A 4-factor solution was clearly supported and explained 68.6% of the total item variance. The loading matrix after Varimax rotation is shown in Figure 1. The corresponding subscales (with five items each) had α reliabilities as follows: *Capable* = .88, *Ethical* = .87, *Sincere* = .84, and *Reliable* = .72. Intercorrelations among the subscales demonstrated that *Sincere* was related to *Ethical* (r = .46, p = .01) and less so to *Capable* (r = -.30, p = .09) and *Reliable* (r = .22, p = .21). All remaining intercorrelations ranged from -.07 to .10 (ps > .50). (Details of the entire item selection procedure and exact formulations of all items can be found here: http://research.clps.brown.edu/SocCogSci/Measures)

4 CONCLUSION

Our analyses of the semantic space of trust aspects suggest that trust has a multidimensional structure. We found four distinct dimensions (*Capable*, *Ethical*, *Sincere*, and *Reliable*) and created internally consistent 5-item sets for each.

There are obvious limitations to this initial study, considering the small sample and entirely exploratory analyses. However, we believe to have succeeded in considering a large number of candidate items and systematically narrowing them to four distinct sets. Our next goals are to replicate this structure using different dependent variables (e.g., sorting task, prediction task) and to create a parsimonious, intuitive instrument that makes the multidimensional nature of trust measurable. For example, we can test the differential responsiveness of each subscale to variations in previously created social and nonsocial scenarios of human-robot trust [9].

ACKNOWLEDGMENTS

The authors thank Elizabeth Phillips and Salomi Aladia for their contributions to this project. This work is supported by Office of Naval Research grant #N00014-14-1-0144. Daniel Ullman is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

[1] Elizabeth Broadbent, Rebecca Stafford, and Bruce MacDonald. 2009. Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics* 1, 4 (2009), 319–330.
[2] David De Vaus. 2002. *Analyzing social science data: 50 key problems in data analysis.* Sage.
[3] Anthony M Evans and William Revelle. 2008. Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality* 42, 6 (2008), 1585–1593.
[4] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.
[5] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
[6] Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene Stokes, Susan Rivers, and Brian Scassellati. 2017. Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Frontiers in Robotics and AI* 4 (2017), 29.
[7] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of Personality* 35, 4 (1967), 651–665.
[8] Kristin Schaefer. 2013. *The perception and measurement of human-robot trust.* STARS, University of Central Florida.
[9] Daniel Ullman and Bertram F Malle. 2017. Human-robot trust: Just a button press away. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, 309–310.