Developing a Multi-Dimensional Model and Measure of Human-Robot Trust

By

Daniel Ullman

B.A., Yale University, 2015

Sc.M., Brown University, 2017

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the Department of Cognitive, Linguistic, and Psychological Sciences at

Brown University

Providence, Rhode Island

October 2021

This dissertation by Daniel Ullman is accepted in its present form

by the Department of Cognitive, Linguistic, and Psychological Sciences as satisfying the

dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Dr. Bertram F. Malle, Advisor

Recommended to the Graduate Council

Date _____ _____

Dr. Oriel FeldmanHall, Reader

Date _____ _____

Dr. Stefanie Tellex, Reader

Approved by the Graduate Council

Date _____ _____

Dr. Andrew Campbell, Dean of the Graduate School

Daniel Ullman

EDUCATION & EXPERIENCE

2021 | **Brown University**
Ph.D. in Cognitive Science—2021
Sc.M. in Cognitive Science—2017

Brown Social Cognitive Science Research Center, 2015–2021
- Advisor: Dr. Bertram F. Malle
- Conducted research within the Humanity-Centered Robotics Initiative (HCRI)
- Created a Multi-Dimensional Measure of Trust (MDMT) to model and measure trust in human-robot interaction (HRI)

2020 | **Facebook AR/VR**
User Experience Researcher Intern

Hardware UX Research & Human Factors
- Conducted research on augmented reality (AR) technology

2018 | **Naval Air Warfare Center Training Systems Division (NAWCTSD)**
Naval Research Enterprise Internship Program (NREIP)

STRIKE Lab, 2018
(Simulation & Training Research to Improve Knowledge & Effectiveness Lab)
- Advisor: Dr. Cheryl I. Johnson
- Conducted research on virtual training environments

2015 | **Yale University**
B.A. in Cognitive Science, Cum Laude & Distinction in the Major

Yale Social Robotics Lab, 2011–2015
- Advisor: Dr. Brian Scassellati
- Conducted research within the NSF Expedition on Socially Assistive Robotics

Yale Comparative Cognition Lab, 2012
- Advisor: Dr. Laurie R. Santos
- Studied a free-ranging rhesus macaque monkey population via non-invasive behavioral research on the island of Cayo Santiago located off the coast of Puerto Rico

BOOK CHAPTERS

2021 | Malle, B. F., & **Ullman, D.** (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction*. Elsevier.

2021 | Haring, K. S., Phillips, E., Lazzara, E. H., **Ullman, D.**, Baker, A. L., & Keebler, J. R. (2021). Applying the swift trust model to human-robot teaming. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction*. Elsevier.

JOURNAL PUBLICATIONS

2018 | Baker, A. L., Phillips, E. K., **Ullman, D.**, & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 30.

2017 | Leite, I., McCoy, M., Lohani, M., **Ullman, D.**, Salomons, N., Stokes, C., Rivers, S., & Scassellati, B. (2017). Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Frontiers in Robotics and AI*, *4*.

CONFERENCE PUBLICATIONS

2021 | **Ullman, D.***, Phillips, E.*, Aladia, S., Haas, P., Fowler, H. S., Iqbal, I. S., Mi, K. L., Riches, I. W., Omori, M., & Malle, B. F. (2021). Evaluating psychosocial support provided by an augmented reality device for children with type 1 diabetes. *Proceedings of the Human Factors and Ergonomics Society Health Care Symposium*. Virtual Conference.
***Denotes equal authorship**

2021 | **Ullman, D.**, Aladia, S., & Malle, B. F. (2021). Challenges and opportunities for replication science in HRI: A case study in human-robot trust. *Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction*. Virtual Conference.
^Best Theory and Methods Paper Award

2020 | Rosen, E.*, Whitney, D.*, Fishman, M.*, **Ullman, D.**, & Tellex, S. (2020). Mixed reality as a bidirectional communication interface for human-robot interaction. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Virtual Conference.
^Finalist, Best Paper Award on Cognitive Robotics
***Denotes equal authorship**

2020 | Rosen, E., Kumar, N., Gopalan, N., **Ullman, D.**, Konidaris, G., & Tellex, S. (2020). Building plannable representations with mixed reality. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Virtual Conference.

2019 | Whitmer, D. E.*, **Ullman, D.***, & Johnson, C. I. (2019). Virtual reality training improves real-world performance on a speeded task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Seattle, Washington.
***Denotes equal authorship**

2019 | Huang, B., Bayazit, D., **Ullman, D.**, Gopalan, N., & Tellex, S. (2019). Flight, camera, action! Using natural language and mixed reality to control a drone. *Proceedings of the IEEE International Conference on Robotics and Automation*. Montreal, Canada.

2019 | **Ullman, D.**, & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. Daegu, South Korea.

2018 | Whitney, D., Rosen, E., **Ullman, D.**, Phillips, E., & Tellex, S. (2018). ROS Reality: A virtual reality framework using consumer-grade hardware for ROS-enabled robots. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Madrid, Spain.

2018 | Phillips, E., Zhao, X., **Ullman, D.**, & Malle, B. F. (2018). What is human-like?: Decomposing robot human-like appearance using the Anthropomorphic roBOT (ABOT) Database. *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*. Chicago, Illinois.
^Best Paper Nominee

2018 | **Ullman, D.**, & Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*. Chicago, Illinois.

2017 | Phillips, E., **Ullman, D.**, de Graaf, M., & Malle, B. F. (2017). What does a robot look like?: A multi-site examination of user expectations about robot form. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Austin, Texas.

2017 | **Ullman, D.**, & Malle, B. F. (2017). Human-robot trust: Just a button press away. *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction*. Vienna, Austria.

2016 | **Ullman, D.**, & Malle, B. (2016). The effect of perceived involvement on trust in human-robot interaction. *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*. Christchurch, New Zealand.

2015 | Leite, I., McCoy, M., Lohani, M., **Ullman, D.**, Salomons, N., Stokes, C., Rivers, S., & Scassellati, B. (2015). Emotional storytelling in the classroom: Individual versus group interaction between children and robots. *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon.

2015 | Leite, I., McCoy, M., **Ullman, D.**, Salomons, N., & Scassellati, B. (2015). Comparing models of disengagement in individual and group interactions. *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon.

2015 | Litoiu, A., **Ullman, D.**, Kim, J., & Scassellati, B. (2015). Evidence that robots trigger a cheating detector in humans. *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon.

2014 | Hayes, B., **Ullman, D.**, Alexander, E., Bank, C., & Scassellati, B. (2014). People help robots who help others, not robots who help themselves. *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. Edinburgh, Scotland.
^Finalist, RSJ/KROS Distinguished Interdisciplinary Research Award

2014 | **Ullman, D.**, Leite, I., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Quebec City, Canada.

2013 | Admoni, H., Hayes, B., Feil-Seifer, D., **Ullman, D.**, & Scassellati, B. (2013). Dancing with myself: The effect of majority group size on perceptions of majority and minority robot group members. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Berlin, Germany.

2013 | Admoni, H., Hayes, B., Feil-Seifer, D., **Ullman, D.**, & Scassellati, B. (2013). Are you looking at me? Perception of robot attention is mediated by gaze type and group size. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. Tokyo, Japan.

WORKSHOP PUBLICATIONS

2018 | Rosen, E., Whitney, D., Phillips, E., **Ullman, D.**, & Tellex, S. (2018). Testing robot teleoperation using a virtual reality interface with ROS Reality. *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

TALKS

2020 | *Human-Robot Trust: A Multi-Dimensional Approach*, Social Cognitive Science Seminar Series, Brown University—October 2, 2020

2019 | *Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust*, NASA RI Space Grant Symposium, Roger Williams University—April 13, 2019

2015 | *Social Robotics: Investigating the Role of Agency in Human-Robot Interaction*, Yale Bulldog Days STEM Symposium, Yale College—April 20, 2015

2015 | *Social Robotics: Investigating the Role of Agency in Human-Robot Interaction*, Yale Engineering and Science Weekend, Yale College—February 15, 2015

2014 | *Humans, Robots, and Monkeys: How Science Resonates with Me*, Yale Resonance, TEDxYale—November 2, 2014

2014 | *Human-Robot Interaction: Work at the Yale Social Robotics Lab*, Social Cognitive Science Research Center, Brown University—August 26, 2014

2014 | *Social Robotics: Understanding and Improving Human-Robot Interaction*, Yale Bulldog Days STEM Symposium, Yale College—April 22, 2014

2014 | *Social Robotics: Understanding and Improving Human-Robot Interaction*, Yale Engineering and Science Weekend, Yale College—February 16, 2014

2013 | *Towards Collaboration with the Yale Social Robotics Laboratory,* McPartland Lab, Yale School of Medicine—May 8, 2013

ACADEMIC AWARDS

2018 | Brown-Hyundai Visionary Challenge—*Winner, Team "Improving Man-Machine Partnership using Mixed Reality Social Feedback from Robots"*
2018 | NASA RI Space Grant Graduate Fellowship—*Full graduate funding for 2018–2019*
2018 | Naval Research Enterprise Internship Program (NREIP)—*Graduate funding for Summer 2018*
2015 | National Defense Science & Engineering Graduate Fellowship (NDSEG)—*Full graduate funding for 2015–2018*
2015 | National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP)—*Honorable Mention*
2012 | Yale College Dean's Fellowship in the Sciences—*Undergraduate funding for Summer 2012*
2011 | National Merit Finalist and Scholarship Winner
2011 | National AP Scholar
2011 | Blind Brook High School Academic Renaissance Award
2010 | University of Pennsylvania Book Award

ACADEMIC SERVICE

Assorted Conferences and Journals—*Reviewer* 2016–present
Brown University Social Cognitive Science Seminar Series—*Coordinator* 2017–2018
Brown University CLPS Department—*Graduate Representative* 2016–2017
Psi Chi, International Honor Society in Psychology (Lifetime Member)—*Executive Board* 2014–2015

OTHER AWARDS

2020 | St. Vincent's Medical Center EMS Values Recognition Award—*Awarded by hospital*
2020 | Trumbull EMS Certificate of Recognition for Service during COVID-19
2020 | Trumbull EMS Certificate of Excellence for Outstanding Management of a Mass Casualty Event
2015 | Yale-Jefferson Public Service Award—*Awarded by Yale University in partnership with the Jefferson Awards Foundation*
2015 | Yale Herbert & Jean Cahoon Award—*Awarded by Yale University*
2011 | The President's Volunteer Service Award—*Awarded by President Barack Obama*
2011 | NY State Senate Legislative Resolution Commendation—*Adopted on March 8, 2011*
2011 | Distinguished Finalist Prudential Spirit of Community Award
2010 | United Nations Outstanding Youth Achievers' Recognition

---

OTHER SERVICE

Trumbull EMS—*Emergency Medical Technician (EMT)* 2014–present
RIde Safe: Bicycle Safety Outreach Program—*Program Creator & Director* 2016–2021
Bristol Fire Department—*Emergency Medical Technician (EMT)* 2015–2021
Downtown Evening Soup Kitchen—*Board of Directors* 2012–2015
Yale Hunger and Homelessness Action Project—*Board of Directors* 2011–2015
HEARTH For All—*Founder & Executive Director* 2009–2013

Preface and Acknowledgements

I am so very grateful to all of the people who have helped me get to this point in my life and made this dissertation possible.

Mom: You have always been my biggest supporter, from preschool to graduate school—thank you for inspiring a love of knowledge and psychology, thank you for always helping me find and pursue my passions, and thank you for your unconditional love and support. I love you to the sky and beyond.

Bertram: I am so immensely grateful for all of the support, guidance, and mentorship you have shared with me these past six years. I have learned an incredible amount from working with you, from multivariate statistics and rigorous experimental design to some of the finer things in life.

My dissertation committee—Bertram, Oriel, and Stefanie: Thank you for your guidance on my dissertation and throughout graduate school, sharing your invaluable expertise and insights with me.

The Malle Lab—especially our lab managers Vincent, Salomi, and Stuti, and our then post-docs, now faculty, Beth, Maartje, and Malik: Thank you for all of your help and feedback over the years, as well as for being a wonderful community of scholars and friends.

Scaz and ScazLab: Thank you for all of your mentorship since I first started in the research world as an undergrad—and for your continued support and friendship.

Beth and Cheryl: Thank you for your research mentorship and friendship, always reminding me that "We have to remember what's important in life: friends, waffles, and work."

My various communities: Thank you to my research collaborators, my CLPS Office family, my research internship teams at the Navy and Facebook Reality Labs, and my EMS families from Trumbull, Bristol, and LifePACT (especially as we worked through the pandemic).

My teachers and mentors, from preschool to graduate school: Thank you for educating and inspiring me, as well as for serving as extraordinary role models.

Finally, thank you to all of my friends and family who have always been there for me. I am so very grateful to all of the amazing people in my life.

This work was supported by Office of Naval Research grants N00014-14-1-0144 & N00014-16-1-2278 and National Science Foundation grant 1717701. This work was also supported by the Brown University Humanity-Centered Robotics Initiative (HCRI). Daniel Ullman was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, and by the National Space Grant College and Fellowship Program, Space Grant Opportunities in NASA STEM (NNX15AI06H).

My graduate training, including all of the research and work that has gone into my PhD, has been the product of the time and effort of numerous people. None of the research I detail in this dissertation was done in isolation, rather everything was completed in partnership with my advisor, my lab, and my fellow researchers. I discuss papers where much of this work has been published, and thus typically use "we" to transparently denote this work and "I" when sharing additional personal anecdotes or insight into this work.

Table of Contents

Chapter 1: Introduction **1**
1.1 Defining Trust 2
1.2 Motivating the Need for Appropriate Trust in Robots 3
1.2.1 Human Expectations for Robots 3
1.2.2 Trust in a Robot as a Subjective State 5
1.3 Conceptualizing Trust 8
1.3.1 Trust in Human-Automation Interaction 8
1.3.2 Trust in Human-Human Interaction 10
1.3.3 Trust in Human-Robot Interaction 16
1.4 Measuring Trust in Human-Robot Interaction 21
1.5 Psychometric Scale Design 24
1.6 Trust Terminology 25

Chapter 2: Trust Items Rating Study **27**
2.1 Method 27
2.1.1 Procedure 28
2.1.2 Participants 29
2.2 Results 30
2.3 Discussion 32

Chapter 3: Trust Items Sorting Study **36**
3.1 Method 36
3.1.1 Procedure 36
3.1.2 Participants 39
3.2 Results 39
3.3 Discussion 40

Chapter 4: Trust Change Study **43**
4.1 Method 44
4.1.1 Procedure 44
4.1.2 Design 45
4.1.3 Participants 46
4.2 Results 46
4.3 Discussion 53

Chapter 5: Trust Items Sorting Expansion Study **55**
5.1 Method 55
5.1.1 Procedure 56
5.1.2 Participants 59
5.2 Results 59
5.3 Discussion 67

Chapter 6: Trust Change Expansion Study **73**
6.1 Method 74
6.1.1 Procedure 75
6.1.2 Design 76
6.1.3 Participants 77
6.2 Results 78
6.3 Discussion 93

Chapter 7: Applications of the MDMT **98**
7.1 "Challenges and Opportunities for Replication Science in HRI: A Case Study in Human-Robot Trust" (Ullman et al., 2021) 98
7.2 "Mixed Reality as a Bidirectional Communication Interface for Human-Robot Interaction" (Rosen et al., 2020) 103
7.3 "Can You Trust Your Trust Measure?" (Chita-Tegmark et al., 2021) 106

Chapter 8: General Discussion **110**
8.1 Trust Factors, Dimensions, and Items 110
8.1.1 Trust Factors 112
8.1.2 Trust Dimensions 112
8.1.3 Trust Items 113
8.2 MDMT Psychometric Design 115
8.3 Conclusion 118

Appendix A: MDMT v1 **120**

Appendix B: MDMT v2 **123**

Appendix C: Trust Change Study Vignettes **127**

Appendix D: Trust Change Expansion Baseline **130**
D.1 Method 130
D.1.1 Procedure 131
D.1.2 Design 131
D.1.3 Participants 133
D.2 Results 133
D.3 Discussion 135

Appendix E: Trust Change Expansion Baseline Vignettes **137**

Appendix F: Trust Change Expansion Study Vignettes **139**

References **143**

List of Figures

Figure 1-A. Literature Review: Dimensions of trust expectations (characteristics of trustworthiness) in selected models. Figure from our book chapter (Malle & Ullman, 2021).

Figure 2-A. Trust Items Rating Study: Example of an excerpted set of items a participant might see in the task.

Figure 2-B. Trust Items Rating Study: Final PCA with 4 components and 5 items each; loadings ≥ .30 displayed.

Figure 3-A. Trust Items Sorting Study: Example of the task participants completed, which consisted of an item bank on the left (items presented in random order to each participant) and the labeled sorting boxes on the right with an "Other" box.

Figure 3-B. Trust Items Sorting Study: Task consensus scores (0%–100%) as the percentages of participants who classified a given item into a category, with items clustered by category/dimension. Filler items in red (creative, easy going, humorous, intelligent, intentional) were assumed to be unrelated to trust and expected to be sorted into the "Other" box. Degree of consensus is shaded from dark gray (0%) to green (100%). Figure from our book chapter (Malle & Ullman, 2021).

Figure 3-C. MDMT v1: Model structure of the MDMT v1, including trust factors, trust dimensions, and trust items.

Figure 4-A. Trust Change Study: Each graph shows MDMT pre-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

Figure 4-B. Trust Change Study: Each graph shows MDMT post-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

Figure 4-C. Trust Change Study: Each graph shows MDMT difference subscale scores (legend) on appropriate scaling (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

Figure 5-A. Trust Items Sorting Expansion Study: Example of the task participants completed, consisting of an item bank on the left (items presented in random order to each participant) and the unlabeled sorting boxes on the right with an "Other" box.

Figure 5-B. Trust Items Sorting Expansion Study: Example of the free response boxes in the task where participants assigned labels to the groupings of trust items they had created.

Figure 5-C. Trust Items Sorting Expansion Study: Frequencies of participants who chose a particular number of trust sorting categories (2, 3, 4, 5, or 6).

Figure 5-D. Trust Items Sorting Expansion Study: Frequencies of the most common trust category labels that participants used to characterize the categories they formed. Columns arranged by five dimensions using experimenter-assigned labels (Reliable, Competent, Ethical, Transparent, Benevolent) and the single-item candidate for the capture of trust (*trustworthy*). Rows arranged by the set of first five most frequent labels, then second five, and then third five.

Figure 5-E. Trust Items Sorting Expansion Study: The proportion of time a specific item was placed in the same bin as another item; the diagonal shows the proportion of time each item was placed into any one of the trust bins (i.e., less than 1.00 on the diagonal indicates that an item was sometimes classified into the "Other" bin). Items organized alphabetically on both axes.

Figure 5-F. Trust Items Sorting Expansion Study: A heat map showing clusters of items based on high co-occurrence proportions. Co-occurrences of items with themselves on the diagonal are grayed out. Items are arranged on both axes based on the co-occurrence clusters. Rate of co-occurrence is shaded from white (0%) to red (100%).

Figure 5-G. Trust Items Sorting Expansion Study: k-means cluster analysis optimized to the AIC value with elbow plot on the left yielding 5 clusters (displayed in Figure 5-H) and k-means cluster analysis optimized to the BIC value with elbow plot on the right yielding 4 clusters.

Figure 5-H. Trust Items Sorting Expansion Study: Visualization from the k-means cluster analysis. The five extracted clusters correspond to the updated five hypothesized dimensions of trust.

Figure 5-I. Trust Items Sorting Expansion Study: Multidimensional scaling using an alternating least-squares algorithm performed on the data.

Figure 5-J. Trust Items Sorting Expansion Study: Visualization of the final selected set of 20 items split across five dimensions for the MDMT v2 (with the parsimonious single-item *trustworthy* also displayed) excerpted from the k-means cluster analysis.

Figure 5-K. MDMT v2: Model structure of the MDMT v2, including trust factors, trust dimensions, and trust items.

Figure 6-A. Trust Change Expansion Study: Each graph shows MDMT pre-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

Figure 6-B. Trust Change Expansion Study: Each graph shows MDMT post-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

Figure 6-C. Trust Change Expansion Study: Each graph shows MDMT difference subscale scores (legend) on appropriate scaling (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right

graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

Figure 6-D. Trust Change Expansion Study: *t* values from Helmert contrasts on the difference scores shown in the diagonal cells, comparing the focal subscale score to the other four subscale scores for the corresponding focal *Targeted Trust Dimension*. Each column shows the subscale scores that resulted from evidence targeting the dimension specified at the top of the column. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

Figure 6-E. Trust Change Expansion Study: Cohen's *d* values from Helmert contrasts on the difference scores shown in the diagonal cells, comparing the focal subscale score to the other four subscale scores for the corresponding focal *Targeted Trust Dimension*. Each column shows the subscale scores that resulted from evidence targeting the dimension specified at the top of the column. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

Figure 6-F. Trust Change Expansion Study: *F* tests using Helmert contrasts on the difference scores comparing the magnitude of difference for the focal subscale score for the corresponding focal *Targeted Trust Dimension* against the scores for the same subscale when other dimensions were the *Targeted Trust Dimension*. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

Figure 6-G. Trust Change Expansion Study: Average number of items (out of four per dimension) that received a "Does Not Fit" selection on the MDMT pre-test and the MDMT post-test for the human and robot agents.

Figure 6-H. Trust Change Expansion Study: Average number of items (out of four per dimension) that received a "Does Not Fit" selection on the MDMT pre-test and the MDMT post-test for the **Performance Trust** factor and the **Moral Trust** factor for the human agent on top and the robot agent on the bottom.

Figure 6-I. Trust Change Expansion Study: Number of participants out of 985 total who selected "Does Not Fit" for all items in a dimension (left columns) or all items across dimensions for a factor (right columns) on the MDMT pre-test and the MDMT post-test for the human and robot agents.

Figure 6-J. Trust Change Expansion Study: MDMT pre-test factor scores for the **Performance Trust** and **Moral Trust** factors (legend) on a 0-7 scale (y-axis) for the human agent on the left and the robot agent on the right (x-axis). Error bars show *SE*.

Figure 6-K. Trust Change Expansion Study: Absolute value difference scores (y-axis) for the human agent on the left and the robot agent on the right (x-axis) for evidence expected to prompt trust gain or trust loss (legend). Error bars show *SE*.

Figure 7-A. MDMT Application One: Trust in imagined future robots for the original study (Thymio), Study 1 (Create), Study 2 (Thymio), and online Study 3 (Both Robots). Figure demonstrates

social-nonsocial trust gap for robots; this finding was assessed using a different two-item trust measure, not the MDMT. Error bars show *SE*.

Figure 7-B. MDMT Application One: Trust in the Create robot in the first conceptual replication study, assessed on the MDMT subscales, the MDMT overall score, and the original two-item trust measure. Error bars show *SE*.

Figure 7-C. MDMT Application Two: MDMT **Capacity Trust** score for the three between-subjects conditions of no feedback control (CTRL), physical feedback (PHYS), and MR feedback (MR). Error bars show *SE*.

Figure 7-D. MDMT Application Three: Figure used with permission (Chita-Tegmark et al., 2021). Original figure caption: "Rating percentages by item: "N/A to robots in general" (red), "N/A to this robot" (yellow)."

Figure D-A. Trust Change Expansion Baseline: Heat map of subscale scores (on a 0-7 scale) for each *Scenario* x *Agent* x *Targeted Trust Dimension* condition (gradient from minimum scores in green, to 50th percentile in yellow, to maximum scores in red).

Abstract of "Developing a Multi-Dimensional Model and Measure of Human-Robot Trust"
by Daniel Ullman, Ph.D., Brown University, October 2021

Abstract

Trust, and specifically appropriate trust, is essential to beneficial interaction between agents. Robots are becoming increasingly present in everyday life and offer a multitude of potential benefits to people, from physically assistive robotics technology for people who have experienced a stroke to socially assistive robots designed for children with autism. Just as trust is essential to people interacting with other people, so too is trust essential to people interacting with robots. The purpose of this dissertation work is threefold: (1) To accurately conceptualize and model trust and its constituent components; (2) To design a measurement tool to capture the nuance of trust in an agent, especially for human-robot trust; and (3) To demonstrate the validity of this measurement tool for human-robot interaction. The empirical findings from this work support a two-factor superordinate conception of trust, as well as reveal a more nuanced five-dimensional structure of trust: the **Performance Trust** factor consists of Reliable and Competent dimensions, and the **Moral Trust** factor consists of Ethical, Transparent, and Benevolent dimensions. This research resulted in the creation of the Multi-Dimensional Measure of Trust (MDMT), which is a model and measurement tool that captures the identified differentiable dimensions of trust in an agent; the MDMT is publicly available for researcher use. The use of the MDMT was tested in studies that measured changes in trust in an agent resulting from changes in salient evidence about the agent along the theorized dimensions of trust; the MDMT was validated for both robot agents and human agents. This dissertation documents the process underlying this research effort, integrating papers published as part of this effort together with additional detail and new work.

Chapter 1: Introduction

When friends ask what my dissertation is about, I usually offer a concise version of the following anecdote; we used a detailed version of this anecdote to introduce our recent book chapter on human-robot trust (Malle & Ullman, 2021).

You are standing in front of a robot, with your back to it. Your task is to let yourself fall backward, and the robot's task is to catch you. Would you trust the robot?

Commonly referred to as the "trust fall" exercise, this situation serves as a quintessential example of trust between two agents and is often used by researchers as a theoretical example to ground discussions and explorations of trust (Hieronymi, 2008; Holton, 1994; Miller et al., 2016; Wagner, 2009). The exercise hinges on the expectation of the falling agent (the trustor) that the catching agent (the trustee) will not let them hit the ground. This expectation has two features: First, the falling agent needs to believe that the catching agent is capable of catching them before they hit the ground. Second, the falling agent needs to believe that the catching agent has the moral integrity and commitment to catch them and is sincere when promising to do so. The first idea is what we henceforth call **Performance Trust**, while the second idea is what we call **Moral Trust**. **Performance Trust** refers to the trustor's confidence that the trustee is capable of completing a given task, while **Moral Trust** refers to the trustor's confidence that the trustee will choose the morally right action and not exploit the trustor's vulnerability. These two beliefs can diverge: We may be confident that another agent is capable of doing what benefits us but not be morally motivated to do so; or we may be confident of the other's moral commitment but not of their capacity to fulfill it. Thus, trust seems to have at least two conceptually independent factors.

Robots are beginning to offer benefits to humans in a range of settings, from schools to workplaces, and across a variety of application domains, from medical care to support in one's home. Robots have moved out of their cages of isolation and safety precautions, where their primary contribution is to reliably complete repetitive physical tasks. Increasingly, robots are used in the world of social interaction, where their expected contribution is to assist and support people. But for robots to successfully fulfill roles in this social context of human-robot interaction (HRI), people must be willing to

interact with these robots, entrust them with tasks that have socially beneficial results, and be confident that the robots are both capable of and committed to bringing about those results. Because trust is integral to social relationships, people will likely be inclined to trust robots with whom they interact. If robots are to succeed in these social interactions, we must first design robots that are worthy of human trust.

The trust fall scenario described above nicely captures the high-level conception of trust in robots that this dissertation identifies and substantiates across a range of empirical studies. This two-factor model of human-robot trust has been strongly and consistently supported across all of the studies we have run in our lab to date. A more nuanced conception of human-robot trust consists of these two superordinate factors of trust (**Performance Trust**, **Moral Trust**) broken down into five dimensions (Reliable and Competent within **Performance Trust**; Ethical, Transparent, and Benevolent within **Moral Trust**). We have accumulated empirical evidence for this multi-dimensional conception of trust, and have continuously worked to validate and refine this conception and the resulting measurement tool we call the Multi-Dimensional Measure of Trust (MDMT). This dissertation documents how this conception and measurement tool were developed and tested over the past several years.

1.1 Defining Trust

In our book chapter we offer our working definition of trust (Malle & Ullman, 2021). I first present this definition here to help anchor the reader to our present conception of trust, before jumping back to the literature review and associated considerations that drove the formation of this definition. Our trust literature review (documented below) indicated reasonable consensus that trust can be defined as follows (slightly updated definition below):

> **Trust** = a dyadic relation in which one agent accepts vulnerability because they expect that the other agent's future action will have certain characteristics; these characteristics include some mix of performance features (reliability, competence) and/or morality features (ethical integrity, transparency, benevolence).

We take this definition as our starting point to introduce a Multi-Dimensional Measure of Trust (MDMT). We believe that the subjective state of trusting (accepting vulnerability) cannot easily be divorced from the trust *expectations* regarding the other's capability, reliability, and morality; instead, the

subjective state is typically directed at one or more of these characteristics. That is, when people say, "X trusts Y," they (implicitly or explicitly) refer to X trusting that Y is capable of and/or reliable in performing a certain action and/or sincere when uttering a statement (e.g., promise, information offer) and/or has the moral integrity and/or benevolence not to exploit or otherwise harm X. Measuring these focal characteristics of trust expectations lies at the heart of our proposed instrument.

I first explore the motivation behind the need for appropriate trust in robots—a focus of and strong belief underlying this work—before diving into how we derived this definition of trust from the component literatures and identifying how it fits with robots.

1.2 Motivating the Need for Appropriate Trust in Robots

Appropriate trust in robots is essential to safe and beneficial human-robot interaction. Robots have the potential to help people, from providing physical assistance to older adults in stroke rehabilitation (Norouzi-Gheidari et al., 2012) to providing socially assistive support to children with autism (Scassellati et al., 2012). Robots range vastly in type, capabilities, and application domains—and variations in robot design can impact human expectations for and interactions with robots. Understanding people's evaluations of robots—including robot trustworthiness—is essential to designing appropriate, socially beneficial human-robot interaction (HRI). How a robot looks (appearance) and how a robot acts (behavior) greatly influences human evaluations of robots as agents.

1.2.1 Human Expectations for Robots

The axiom "form follows function" is commonly considered an important principle of design and links to the idea that people often generate expectations for what a thing does based on how it looks. Consider the low-tech example of a door that is not clearly marked "push" or "pull" and does not give clear cues to which way the door works. Most, if not all, people can empathize with the experience of pushing a pull door or vice versa—both of which are not particularly fun experiences. Likewise, even though robots are much more complex instances of technology, people generate expectations for robots based in part on appearance (Duffy, 2003; Goetz et al., 2003). And just like the low-tech example of a door, robot appearance does not always transparently convey what a robot does.

Appearance is one example of a feature that influences human expectations for a robot and is often the first thing that people observe when interacting with a robot. The question "What is a robot?" can prompt vastly different responses from people, let alone associated mental models of robots based on personal experience. In one line of research I worked on together with collaborators, we showed how people generate different ideas of what a robot looks like when they were given slightly different category information (Phillips et al., 2017), and later identified the component features people think makes a robot look human-like (Phillips et al., 2018). Appearance can influence people's expectations for a robot, including that people make inferences about a robot's mental capacities based on its human-likeness (Zhao et al., 2019). In particular, robot form can influence initial perceptions of robot trustworthiness (Schaefer et al., 2012).

Although robot appearance can influence initial expectations for a robot, robot behavior typically drives continued use or disuse of the technology. Understanding how people respond to robot behavior is thus essential to understanding human-robot interaction—with the focus here on behaviors that potentially prompt changes in trust in a robot. Research has shown the potential dangers of overtrust in robots. In one such study involving a simulated emergency evacuation scenario, researchers found that all 26 participants followed a robot even though half of those participants had observed the robot perform poorly in a navigation guidance task just minutes prior (Robinette et al., 2016). Understanding a robot's actual ability or decision making is essential to conveying accurate information that can match appropriately to **Performance Trust**. This aligns with the push from the artificial intelligence (AI) community to develop explainable artificial intelligence (XAI), which strives to transparently communicate decision making to users so that they understand and can appropriately trust and interact with artificial intelligence systems (Gunning, 2017; Gunning et al., 2019). System behaviors and performance can foster user understanding of those systems; researchers and designers must work to ensure that technology ability is sufficiently communicated so that users have the necessary information to evaluate those systems. Inappropriate trust poses a real danger to human-robot interaction, especially given the various applications of robots with potentially vulnerable populations—such as children (Stower et al., 2021).

The vast variation in types of robots prompts numerous questions about human expectations for social robots, ranging from how people think about existing robots based on their specific features to what this means for design of future robots (Fong et al., 2003). Human-robot interaction research is often conducted with a single particular robot, such that questions emerge about whether findings are confined to that specific type of robot or generalize to robots in general. For example, we conducted a replication effort of a possible human-in-the-loop involvement effect on trust in robots (Ullman & Malle, 2017), thinking at one point (based on research findings) that the effect was possibly driven by variation in the robot—only to find the effect itself did not replicate (Ullman et al., 2021). My aim with this dissertation research was thus to develop a model and measure of trust in robots that can be used across the wide range of robot types and application domains. This goal informed the design of research studies we developed in order to ensure the applicability of the model and measure to robots in general. Rather than anchor to any specific robot(s) or agent(s), we therefore designed initial conceptual research studies to be agnostic to type of agent and context during development of the trust model (discussed further in study chapters). We then designed later evaluative studies to test the trust model and associated measure using vignettes that describe an assortment of robots performing in a range of roles and contexts—ensuring generalizability across robots. These evaluative studies used variations in robot behavior to test whether the developed measurement tool can in fact capture corresponding changes in a person's trust in a robot; this aligns with an important conceptual point about how we consider trust in this work, discussed next.

1.2.2 Trust in a Robot as a Subjective State

In this research, we work to understand trust as the "subjective state" of a person (the trustor) toward another agent (the trustee)—i.e., a person's experience of trust in another agent as a point in time capture based on the features of the other agent and the contextualizing scenario. This idea of subjective trust integrates across contributing features and reflects a person's experience of the "trustworthiness" of another agent at a point in time. It is helpful to distinguish this holistic conception of trust from considerations that may influence the subjective state of trust as possible contributing factors.

First, a contributing factor to a person's subjective state of trust as a trustor may be a more stable, dispositional, trait-like individual "propensity to trust" that varies across individuals (Frazier et al., 2013). That is, some people may generally be more trusting than other people and thus have higher rates of trust. While this may contribute to a person's subjective trust across situations, it is not our sole focus. Our goal is not to assess stable, dispositional individual differences in trust tendencies across trustors, but to capture a trustor's one-time subjective trust state in another agent for a particular situation.

Second, the conception we are focused on hinges on the trustor's subjective belief state, rather than assume there are constant, underlying, knowable traits in the trustee that can be objectively quantified and measured. For instance, a person may believe a robot is morally trustworthy without a robot demonstrating any "actual" properties of moral competence (see Section 1.3.3 for a more detailed discussion of robots as moral agents). The focus is thus on the trustor's trust beliefs, rather than an attempt to measure any presumed stable traits of trustworthiness inherent to the trustee. A wide variety of features can influence trust in robots, including robot appearance and behavior, as well as the particular context involving the potential for trust between a person and a robot. The trust that one places in another is not a constant, but rather changes in response to new information. In addition to any consistent features demonstrated by the trustee that might influence perceived trustworthiness (such as successful repeated performance on tasks, like always catching a person in the trust fall), one-time instances related to trust (such as almost dropping a person but catching them at the last minute) may contribute to a trustor's trust in the agent. While a particular trustee may have and exhibit features generally perceived to be more or less trustworthy than another agent, the focus of this trust measurement is on a person's subjective experience of trust in an agent at a point in time.

Finally, an important distinction must be made here between trust as a subjective mental state and trust as behavioral reliance on an agent—where behavioral reliance is centered on how a person interacts with another agent. For example, if the robot lets the person fall in the first instance of the trust fall exercise, then the person would likely trust the robot less (subjective state) but for a variety of reasons might still participate in a second trust fall (behavioral reliance). However, should the person choose to

either fall or not fall a second time, exhibiting the presence or absence of behavioral reliance, we would be left with a critical question—why? Everybody makes mistakes—but does the person believe that the event was the result of a robot error (i.e., a drop in the perceived **Performance Trust** of the robot), or the result of ill will (i.e., a drop in the perceived **Moral Trust** of the robot)? This is precisely what we aim to assess: the subjective state of trust of a person in another agent as reflected in superordinate factors and dimensions of trust. In this dissertation, I work to understand a person's subjective state of trust in another agent, conceptualizing and measuring the contributing components of that trust—or the lack thereof.

Behavioral reliance is undoubtedly important to understanding human-robot interaction—and is addressed by a variety of research in HRI—but ultimately it shows the results of a more complex process that includes trust as one piece of the puzzle...albeit an important one. Just because a person participates in a trust fall exercise with a robot for a second time—or follows a robot in a simulated emergency evacuation scenario after observing it perform poorly in a previous task—does not mean they have the same subjective state of trust in the robot as before. Accurately measuring and understanding a person's subjective state of trust in a robot—and thus a person's evaluation of the perceived trustworthiness of a robot—is a critical piece of this puzzle. People should ideally only rely on robots to the extent that they consistently and competently perform tasks; being able to understand and measure trust in robots is an important aspect of designing robots for ethical, appropriate human-robot interaction.

There are many areas of research in HRI that deserve attention. My research interests have gravitated toward trust in robots, and specifically toward measuring trust in order to facilitate appropriate trust in robots. Robots have the potential to offer a vast array of benefits to people—but only if implementation is reasonable and appropriate. Extensive literature review at the beginning of graduate school indicated a gap in the consistent conceptualization and measurement of trust in agents, both within and across various domains (including human-automation interaction, human-human interaction, and human-robot interaction). For my dissertation, I aimed to help close this gap; I began by first working to understand existing conceptions of trust across domains, discussed next.

1.3 Conceptualizing Trust

I conducted an extensive literature review on the topic of trust spanning the domains of human-automation interaction, human-human interaction, and human-robot interaction in the first few years of my graduate studies, which I then collated and used to write a theoretical review in a paper for my preliminary examination at Brown University. The paper, entitled "A Two-Factor Theory of Human-Robot Trust: Capacity Trust and Moral Trust" detailed research on conceptions of trust spanning these domains and described my working two-factor theory of trust (with the term **Capacity Trust** since updated to **Performance Trust**). This paper served as the foundation for the previously mentioned book chapter that my advisor and I wrote (Malle & Ullman, 2021). I include the literature review and resulting theoretical framework described in the chapter here, as well as in the next section on measuring trust in human-robot interaction (with updates from additional research reviewed since).

There is a spectrum of agency that we can consider from the endpoint of a simple machine (such as a toaster) to a human. And then we have the corresponding interactions between humans and the agents on this spectrum. On the one hand we have human interaction with automation, and on the other we have human interaction with another human. Human-robot interaction (HRI) is often hypothesized to sit somewhere in the middle, featuring aspects of both human-automation interaction and human-human interaction. We therefore first examine the human-automation literature and the human-human literature, before turning to the intersection that is the human-robot literature.

1.3.1 Trust in Human-Automation Interaction

Work in the domain of automation emphasizes the performance of automated systems. Automation serves its purpose when a system can safely and efficiently perform a task that reduces a burden on human users. Many of these task domains have historically consisted of nonsocial tasks, often combined with system oversight by a human operator but little collaboration between humans and the system. A useful heuristic in thinking about whether or not to trust a system or an agent comes in the form of a basic question: "What do I worry about that would prevent me from interacting with this agent?" In the automation literature, such worry is tied to the performance of the system (i.e., **Performance Trust**);

this worry can be alleviated, and allows for trust, if the system is capable of performing its task and does so consistently (Schaefer et al., 2016).

Sheridan and Parasuraman (2005) reviewed research in human-automation interaction and offered two sets of features of trust. One set comes from the system's lower-level reliability of performance, while the other comes from the system's higher-level ability. Ideally, trust in a system is grounded in an accurate conception of the system's ability and reliability; however, trust is not always appropriately calibrated. Parasuraman and Riley (1997) discussed multiple miscalibrations: misuse (overreliance on automation), disuse (underutilization of automation), and abuse (inappropriate application of automation). Based on previous research, Lee and See (2004) proposed that systems must be designed to help match users' expectations to the system's actual performance capabilities. Calibrated trust is grounded in an accurate assessment of what a system can and cannot do—as well as why the system fails when it does. System transparency is a key issue for trust in human-automation interaction and refers to the degree to which a person understands the actions of a system; systems that are opaque are often frustrating to users, and the inability to understand a system can lead to misuse or disuse (Ososky et al., 2014).

Because errors reveal a system's performance quality, they have been a particular focus in human-automation work. For example, Madhavan and colleagues (2006) showed that trust in an automated decision aid declined when the user observed system errors. However, not all errors were treated the same way: when a system made errors on easy trials rather than difficult trials, users mistrusted the system far more. People potentially use task difficulty as a diagnostic indicator: failing to complete easy trials shows low ability.

In the absence of trust in a system, users will opt to not use it. Lee and Moray (1992) investigated the determinants that influence whether human operators will rely on a system performing a task, or opt for manual control in the task. They used an experimental task where human operators would choose between manual control or automatic control for operating a simulated semiautomatic pasteurization plant. The researchers found a tradeoff between human operators' self-confidence and trust in the system, such that operators opted for automatic control when trust in the system exceeded self-confidence and

opted for manual control when self-confidence exceeded trust in the system. Trusting a system involves accepting some kind of risk and believing that the system is able to limit that risk.

A recent review of 127 empirical studies on human trust in automation (Hoff & Bashir, 2015) identified numerous factors that affect trust, including culture, personality, task characteristics, workload, self-confidence, and more; but the object of the trust itself—the automation system—was described only in terms of its performance: its reliability, predictability, and error-proneness. In sum, in the literature on trust in automation, the primary focus is on appropriately matching a person's expectations for a system with information about the performance of the system. Trust here is the expectation that a system will perform a task as intended and expected, and the only worries that arise concern the system's reliability and ability.

In sum, the human-automation literature review showed a primary focus on features that fall under the broad heading of what we term **Performance Trust**.

1.3.2 Trust in Human-Human Interaction

As we move from the domain of human-automation trust into the domain of human-human trust, we can pose the same heuristic question: "What do I worry about that would prevent me from interacting with this agent?" Unlike the primary focus on **Performance Trust** in human-automation interaction, here the focus widens to include expectations of whether a human agent will act morally—what we refer to as **Moral Trust**. In situations of human-human trust, the question becomes whether a human agent will exploit another human's potential vulnerability—either unintentionally because of a lack of ability, or intentionally because of a lack of moral integrity.

It is instructive to imagine what a social community without trust would look like. People would not expect that another person would have their best interest in mind; instead, they would expect that others do what benefits them even when such actions exploit or harm others. Lack of trust is thus a reflection of a society without prosocial norms, without moral commitments. By contrast, in societies that have such commitments, trust is possible and has the power to enable and sustain cooperative behavior (e.g., Gambetta, 1988; Jones & George, 1998). Trust acts as a glue that enables people to live and work

together without constant worry and threat of being exploited. In fact, such societies uphold a norm to trust other people (Dunning et al., 2014), which places demands on those people to justify the trust they are granted—and makes violations of trust all the more salient.

Philosophical analyses typically treat trust as a three-part relation among a trustor, a trustee, and something that the trustor expects the trustee to do (Hardin, 2002) or to care for (Baier, 1986). The oft-cited conceptualization of trust by Mayer and colleagues (1995) has at its core a three-part relationship as well, but in addition to the trustor and a trustee, the authors focus on the role of risk. Integrating these elements, we can say that the trustor typically expects the trustee to act so as to avoid or reduce the trustor's risk in the situation. The object of this expectation (what the trustee is expected to do or be) corresponds to the notion of trustworthiness—a trustee's characteristics that either inspire or justify the trusting expectation: their ability, reliability, ethical integrity, and so on. The exact characteristics of trustworthiness, and thus the objects of trust expectations, are debated. Therefore, we review several of these proposals, ranging from few to many characteristics, and identify the common denominators. With those in hand, we can then examine the place of human-robot trust in the broader psychological landscape of trust.

Rotter (1967) conceptualized trust as "an expectancy…that the word, promise, verbal or written statement of another…can be relied upon." Rotter's proposal thus highlighted the trust expectation of sincerity (truthful words and standing by promises), which is confirmed by a closer look at the items in his interpersonal trust scale. Chun and Campbell (1974) conducted cluster and factor analyses on this scale, and in their results, we see the primary interpersonal facet of sincerity (honest, truthful). Selfish exploitation formed another factor, perhaps the opposite pole of what other authors call benevolence. Many of the other items in Rotter's scale referred to institutional trust and specifically to concerns about institutions being sincere and ethical (e.g., unbiased, not cheating). However, there is no mention of competence or reliability/predictability. The worries people have about others are cast here entirely in terms of morality.

Several authors who examined human-human trust, from sociology to management, highlighted competence and integrity as the major expectations (Kim et al., 2009; Parsons, 1951). In Cook and Wall's (1980) conception, trust was organized into faith in the intentions of others (often labeled benevolence), and confidence in the ability and the reliability of others. However, when the authors measured trust in an organizational setting, these three aspects did not come apart. Barber (1983), in his analysis of the broader societal role of trust, distinguished between three trust expectations: persistence (predictability), competence, and moral duties (to have others' interests in mind, which most authors call benevolence). Slovic and colleagues (1993) asked ordinary citizens to express their trust or distrust in power plant management as a function of various behaviors by the management; the authors' choice of such behaviors revealed their conception of trust as expectations about competence (e.g., being prepared for accidents) and about two moral dispositions that are often labeled benevolence (good motives) and sincerity/transparency (being truthful, providing access). However, no attempt was made to distinguish these expectations in measurement. Focusing on people's attitudes toward institutions, Carnevale's (1995) definition of trust included reliability and competence as well as three moral aspects: nonthreatening (benevolent), fair, and ethical. Caldwell and Clapham's (2003) proposal, tailored to organizations, included seven aspects of trustworthiness that can be divided broadly into competence (knowledge, ability), the responsibility to inform (transparency), and various moral or normative aspects (quality assurance, respect, legal compliance).

Gabarro (1978) conducted interviews with 4 company presidents and 33 subordinates over the span of 3 years. He extracted six objects of trust. While maintaining aspects of competence (skills and good judgments) and consistency (reliability, predictability), he differentiated the moral dimension into integrity (encompassing honesty and moral character), motives (benevolent motives, commitment), openness (defined as honesty, being straight, not hiding—aspects most other authors label sincerity), and discreetness (not violating confidence). Butler and Cantrell (1984), as well as Schindler and Thomas (1993), experimentally tested five of these characteristics (omitting discreetness, perhaps because it could

be grouped under integrity) as determinants of overall trust judgments. Integrity and competence showed the most considerable impact, while consistency was weak, and openness had little to no impact.

Butler (1991) also interviewed managers and content-analyzed characteristics that the managers mentioned in describing trusted and mistrusted people. In subsequent scale development and iterated factor analyses, Butler postulated nine characteristics, but only six seem to directly capture trust: competence and consistency as the familiar performance aspects, as well as the moral characteristics of integrity, fairness/loyalty, discreetness, and promise fulfillment (the others were availability, receptivity, and openness). However, when examined in people's judgments, Butler's moral characteristics were very highly correlated (*r*s between .65 and .76), suggesting one large moral cluster. Competence was somewhat differentiated from these moral components (correlating with them in the .40s and .50s), while consistency more clearly set itself apart (correlating with all the other components in the .30s and .40s).

Despite differences among the various authors' conceptions of human-human trust, we see that almost all of them support a multi-dimensional concept. Most authors assigned a prominent status to competence and reliability, mirroring the human-automation literature, but many added a moral dimension, with anywhere from one to four moral facets. Mayer and colleagues (1995) tried to integrate these variations and to consolidate them into the major characteristics of trustworthiness that sway a trustor. Against the background of 23 previous proposals, they derived three such characteristics: ability (including knowledge, expertise, competence), benevolence (a positive orientation toward the trustor), and integrity (adhering to moral principles shared with the trustor). Two omissions are noteworthy here. First, the authors excluded predictability from the conceptual space, mainly because they argued that reliability was not sufficient for trust. However, none of the individual characteristics are sufficient for trust, so reliability should not be discarded. Second, sincerity (common among many other models) was absent, not because of direct empirical evidence but because of the authors' conceptual decisions in selecting and compiling characteristics of trust. Interestingly, McKnight and colleagues (1998) cited Mayer as the basis for their conception of trust expectations but worked with four dimensions, including

competence (ability), predictability (added back in), honesty (rather than integrity, thus bringing sincerity back into the picture), and benevolence.

Nonetheless, a meta-analysis showed that Mayer and colleagues' three-dimensional conception is successful in predicting trust states (overall willingness to accept vulnerability) from trust expectations (Colquitt et al., 2007). The predictive correlations were in the .60s for each dimension, but the three dimensions were also correlated with each other between $r = .62$ and $r = .68$. Because reliability and sincerity were not included in the meta-analysis we do not know what their role would be in affecting subjective trust states.

A decade later, after yet more and varied proposals of conceptualizing trust, Burke and colleagues (2007) tabulated 27 such conceptualizations. We performed a frequency analysis of the most-used content words in these definitions and found considerable common ground in what trust is: a dyadic relation in which one person accepts vulnerability because they expect that the other person's future action will be governed by certain characteristics. And though the specific characteristics (of trustworthiness) are rarely mentioned in the definitions, those that are mentioned include the by now familiar notions of ability, reliability, and a bundle of moral characteristics such as benevolence, honesty, and integrity.

One of the challenges in trust work across domains is the difficulty in representing ecologically valid risk in human subjects research. Experiments in the lab (and online) must ultimately be safe, making the inclusion of true vulnerability inherently difficult due to the near-intractable problem of meaningfulness of the stakes (Camerer, 2011; Kee & Knox, 1970). Real-life situations with trust must have risk and often involve something personal at stake. Behavioral economics games, such as the aptly named Trust Game (Berg et al., 1995), can simulate risk through means of money or points. Behavioral games offer the opportunity to test a range of circumstances using a comprehensive and systematic investigation of factors (Camerer, 2011). Many of these studies can be run with relative ease in the lab or online, and the assortment of behavioral games allows for testing of a variety of relationships between agents using an array of parameters. However, money does not capture all of the types or nuances of risk in human interaction, including the potential for physical harm (such as in the trust fall exercise) or

emotional harm. The challenge thus becomes not just how to best simulate risk in a situation, but how to simulate the most appropriate kind of risk for the types of interactions and situations being studied. In addition, Camerer notes substantial cross-cultural variation in the results of behavioral games, including the Trust Game; this variation is interesting in its own right, but for this discussion is even more interesting as we start to consider how people interact with technology—if expectations surrounding human-human trust vary by culture, so likely will expectations for human-robot trust.

Our review of dimensions of trust expectations in the literature from our book chapter (Malle & Ullman, 2021) is summarized in Figure 1-A. Though not all assignments are clear-cut, the overall picture suggests that trust has two broad superordinate factors: a **Performance Trust** factor, with dimensions of competence and reliability, and a **Moral Trust** factor, with dimensions of integrity (being ethical), sincerity, and benevolence. The importance of this moral factor is what appears to differentiate trust in humans from trust in automation, and this difference is best explained by the significance of social interaction and risks involved in human-human relationships. The question arises whether human-robot trust demonstrates aspects of a **Moral Trust** factor.

	Performance		Moral		
	Competence	**Reliability**	**Integrity**	**Sincerity**	**Benevolence**
Rotter (1967)				✓	
Chun and Campbell (1974)			✓	✓	✓
Gabarro (1978)	✓	✓	✓	✓[a]	✓[b]
Parsons (1969)	✓		✓		
Cook and Wall (1980)	✓	✓			✓
Barber (1983)	✓	✓			✓[c]
Butler (1991)	✓	✓	✓		✓[d]
Mayer et al. (1995)	✓		✓		✓
Slovic et al. (1993)	✓			✓	✓
Carnevale (1995)	✓	✓	✓		✓
McKnight et al. (1998)	✓	✓	✓[e]	✓[e]	✓
Caldwell and Clapham (2003)	✓		✓		✓
Analysis of definitions in Burke et al. (2007)	✓	✓	✓		✓
Kim et al. (2009)	✓		✓		

[a] *Calls it "openness" but explicates as honest, straight, and not hiding things—arguably elements of sincerity.*
[b] *Calls it "motives" and arguably refers to benevolent motives.*
[c] *Calls it "moral duties" (with a focus on having others' interests in mind).*
[d] *Calls it "fairness/loyalty."*
[e] *Call it "honesty," which, according to premier dictionaries has both meanings of being sincere, truthful and having integrity, moral principles (though with a stronger emphasis on shades of sincerity).*

Figure 1-A. Literature Review: Dimensions of trust expectations (characteristics of trustworthiness) in selected models. Figure from our book chapter (Malle & Ullman, 2021).

1.3.3 Trust in Human-Robot Interaction

Much of the extant work on human-robot trust focuses on the actual reliability and ability of robotic systems, so trust in intelligent robots is typically considered along the same factors seen in the

broader human-automation literature (Yanco et al., 2016). For example, Hancock and colleagues (2011) conducted a meta-analysis of factors that prior research has identified as influencing trust in human-robot interaction. The authors collated 21 studies and found that properties of the robot—specifically, robot performance (such as reliability)—had the strongest association with trust. Human-related factors (e.g., attitudes and comfort with robots) and environmental factors (e.g., culture and physical environment) contributed relatively less.

In a recent review, Lewis and colleagues (2018) argued for a distinction between performance-based interactions between humans and robots and social-based interactions. This social focus points to some of the distinguishing features that set robots apart from automation. In particular, robots are being introduced into more and more social settings, projected to be social companions for older adults, tutors for children in schools, or assistants to people with health needs (Broadbent et al., 2009). Such contexts often involve an element of risk. Risk may arise from a robot moving around an older adult's house and accidentally knocking them to the ground, which is a traditional safety concern; but risk may also arise from a robot refusing to increase a person's pain medication because the decision authority cannot be reached, or from a robot unknowingly presenting incorrect information to a child. The latter situations involve social and moral norms and thus raise the question of moral trust in a robot. But do people actually treat robots as moral agents?

The question of whether people treat a robot as a moral agent involves two parts: do people treat the robot as an agent, and do they evaluate it as an agent with moral capacities. This question is important to the discussion of human-robot trust; ultimately, empirical evidence establishes that people do in fact treat robots as moral agents.

First we consider the idea of robot agency. Research has shown that when people explain robot behavior, they use the same concepts and language of intentional agency and mind as they do when explaining human behavior (De Graaf & Malle, 2019). Broadly, people seem to think about robots as agents in similar (though distinct) ways as they do human agents. Research using gaze following as a measure of perceiving agency showed that even human infants respond to a social robot (i.e., a robot

exhibiting sensical social interactive behaviors) as a psychological agent, compared to a robot without the same behaviors (Meltzoff et al., 2010). Perceptions of robot agency can vary based on contexts and features of the interaction—as can perceptions of agency in other agents in general; focusing on the actions of a robot in-the-moment can prompt greater ascriptions of agency to the robot (Takayama, 2012).

We next turn to the idea of attributing moral capacity to robot agents. Empirical evidence shows that people do in fact treat robots as moral agents—such that **Moral Trust** is therefore likely to be attributed in varying degrees to robot agents in relevant human-robot interaction contexts. When asked to infer the mental capacities of robots, respondents are inclined to grant robots some capacity for moral decision making (Malle, 2019; Weisman et al., 2017), and significantly more so the more human-like the robot looks (Malle, Zhao, et al., 2019). By contrast, people do not welcome moral decisions by cars and other machines if they lack human-like mental capacities (Bigman & Gray, 2018). An increasing number of studies also show that people apply moral norms to AI and robotic agents and blame those agents when they violate the pertinent norms (Malle et al., 2015; Malle, Magar, et al., 2019; Shank & DeSanti, 2018). In a human-robot interaction study too (Kahn et al., 2012), a majority of people—approximately two-thirds—thought of the robot as morally accountable for a specific transgressive behavior. While not all people evaluate all robots as moral agents across various contexts (Banks, 2019), it appears that the majority of people do. The rate at which people do not treat a robot as a moral agent—i.e., a "moral disqualification" rate—is an important metric in understanding human perceptions of robots in particular situations. A series of research studies using moral dilemmas showed similar moral disqualification rates of about one-third, such that the majority of people (two-thirds) blamed a robot for its decision (Malle et al., 2016; Scheutz & Malle, 2021). Additional cross-cultural research (comparing Japanese and U.S. participants) on moral dilemmas involving robots showed the majority of both Japanese and U.S. participants treated robots as moral agents, and that Japanese participants showed even lower moral disqualification rates (Komatsu et al., 2021). If people treat robots as capable of moral decision making and blame them if they make the wrong moral decisions, then there is room for developing trust in robots if they are sincere, ethical, and benevolent—and to lose trust if they are not.

A series of studies I worked on that were published in 2014 and 2015 investigated whether people recognize a robot's attempt to cheat in a game of rock-paper-scissors by changing its hand gesture after seeing its partner's choice (Litoiu et al., 2015; Short et al., 2010) or in a game of Battleship by lying about a ship's position (Ullman et al., 2014)—and people did recognize those violations. Such recognition, rather than the belief that the robot was simply malfunctioning, may reveal that people saw the robot as being unethical or insincere. Indeed, Wijnen and colleagues (2017) showed that a lying robot (diverting blame to a human for a negative act the robot committed) was trusted less in a behavioral trust game than an honest robot.

For people to gain or lose moral trust in a robot, the robot does not have to be considered a "genuine" moral agent. Even when the designer is really the one who is insincere and untrustworthy, the human interacting with the robot may well direct their distrust and disappointment at the robot. And what holds for sincerity could in principle hold for benevolence and integrity as well, though these two characteristics may demand more behavioral evidence.

Evidence for ethical integrity would be seen in following moral norms and principles: for example, being fair, nondiscriminatory, cooperative, and respectful (Kuipers, 2018; Malle & Scheutz, 2019). Especially when robots are taking on a broader range of roles (e.g., in domains of education or health care; Broadbent et al., 2009), the norms associated with these roles must be made explicit and the robots must be equipped to comply with them. To build robots that guide their behavior by adhering to social and moral norms is still a significant challenge given the current state of the art in robotics (Malle, Bello, et al., 2019); but even just seeing a robot try to follow norms is likely to instill a degree of trust.

Benevolence requires putting one's own interests behind others' interests, rather than selfishly benefitting while others incur costs. Even when robots are not designed to be selfish, they will sometimes pursue goals that run counter to people's interests—such as when a robot hinders or inconveniences someone or is unable to meet a request. In some cases, a robot's built-in goals may purposefully circumvent people's interests or values—such as when a robot tries to coax people into buying something

or revealing private information (Calo, 2011; Hartzog, 2015). Such behaviors are often disguised and therefore insincere as well, suggesting an overall lack of ethical standards.

When a person attributes blame in a moral situation involving a robot, is the person blaming the robot, the programmer, or someone/something else? The question of the "real target" of responsibility or blame in moral scenarios involving robots is interesting and important. For example, research on expectations for autonomous vehicles shows that people evaluate responsibility differently when traffic accidents involve self-driving vehicles compared to vehicles operated by human drivers—allocating greater responsibility to the manufacturer and the government for the self-driving vehicle than the vehicle with a human driver (Li et al., 2016). There are many possible "real targets" of blame here, including the vehicle itself, the manufacturer, the programmer(s), government regulatory agencies, the car owner, etc. And there are numerous potential ramifications tied to determining/assigning responsibility and blame, including issues of import for legal systems and regulatory agencies (Gless et al., 2016). These are important questions concerning robot agents that warrant further investigation and study in evaluations of human-robot interaction in society. Few studies give people an explicit choice of whom to blame; the aforementioned moral dilemma studies (Komatsu et al., 2021; Malle et al., 2016; Scheutz & Malle, 2021) that gave people a chance to opt out (morally disqualify the robot) found that around one-third of participants morally disqualified the robot, and then no more than half of those participants mentioned the programmer, designer, etc. An analogous parallel in the human-human interaction domain might be evaluations of moral actions with a child: When a young child does something evaluated to be morally wrong, people may place varying degrees of responsibility or blame on the child, the child's parents, society, etc. In any case, whether the robot agent or child agent is deserving of responsibility or blame (or whether the blame is placed elsewhere), the agent may nonetheless be evaluated and treated as a trustworthy or untrustworthy agent.

Even if current robots have only minimal moral capacities, their ever-widening roles in society, their routine spoken communication, and their often humanoid looks will (and perhaps already do) prompt people to treat them as social and moral agents (Coeckelbergh, 2012). Robots are machines, evaluated for

their performance, but increasingly they are also social agents, evaluated for their potential to cause harm. In a context of vulnerability, people are likely to consider a robot's moral characteristics of trustworthiness, such as sincerity and integrity. Because research shows that most people do treat robots as moral agents, it is critical that robots are designed to be morally competent to facilitate appropriate trust in robots in contexts involving moral aspects. We must, therefore, measure these moral characteristics along with the familiar characteristics of ability and reliability. I next explore the measurement tools currently available to assess these multi-dimensional layers of human-robot trust.

1.4 Measuring Trust in Human-Robot Interaction

Having conducted an extensive, though not necessarily exhaustive, literature review of trust in relevant domains, we set out to explore existing measurement tools intended to capture trust in human-robot interaction. There is no single trust measurement tool that is universally used in HRI studies, but instead researchers use a variety of trust measures and often adapt or change measures for specific use/applications (Chita-Tegmark et al., 2021). We examine existing measures, with particular emphasis on the extent to which measures target either or both of the broader identified trust factors (**Performance Trust**, **Moral Trust**).

Hancock and colleagues (2011) noted that there is a scarcity of validated measures with which to evaluate human-robot trust consistently across experimental designs, which makes the phenomenon difficult to study and compare across studies, labs, and domains. This echoes the call by Steinfeld and colleagues (2006) for a singular toolkit of human-robot interaction metrics that includes trust.

A discussion of measuring human-robot trust must first address the distinction between trust as a subjective state and behavioral reliance (Kee & Knox, 1970); this important differentiation was explored in Section 1.2.2. Although trust often becomes apparent to the outside observer only when an agent makes a choice that embraces some risk, such a choice is ultimately the product of an internal state of trust that is translated into action. Mayer and colleagues (1995) similarly distinguished between the willingness to accept one's vulnerability and the action of taking a risk. These two aspects, the subjective-internal and the public-behavioral, can be measured separately and are predictably related (Colquitt et al., 2007).

However, a risk-taking choice is not always grounded (solely) in subjective trust. Imagine a forced choice to rely on one of two people—relying on Agent A is not necessarily indicative of trust in Agent A, rather it could be due to a much greater benefit one stands to gain from choosing Agent A or a fear of Agent B. Thus, even though the act of risk-taking is often an important consequence of trust, trust itself is more closely captured by an internal state (Lewis et al., 2018). Internal states cannot be measured directly but are themselves operationalized by verbal reports, nonverbal expressions, and the like.

In the human-human domain, this internal state of trust consists of an acceptance of one's vulnerability and of expectations that the other's characteristics warrant this acceptance. These expectations, we have seen, are multi-dimensional and include at least performance characteristics (ability, reliability) and one or more moral characteristics (e.g., sincerity, integrity, benevolence). The measurement tools currently available for human-robot trust vary in how much they consider **Moral Trust** and **Performance Trust** factors, but they are heavily skewed toward the latter.

Schaefer (2013) offered one of the more comprehensive developments of a trust measure. Across 6 experiments, Schaefer developed a 40-item trust scale to assess the 3 antecedents of human trust in robots considered by Hancock and colleagues (2011): features of the robot, the human, and the environment. In addition, a 14-item scale focused specifically on trust expectations: characteristics that make the robot trustworthy. Of these 14 items, 11 reflect dimensions of **Performance Trust** (e.g., function successfully, act consistently) while 3 relate to social aspects (provide appropriate information, communicate with people). No item in the short form directly invokes the **Moral Trust** dimensions, but a few items in the longer 40-item scale refer to open and truthful communication (i.e., sincerity), and one item mentions protecting people. Similarly, Yagoda and Gillan (2012) proposed a scale for human-robot interaction in teams that identifies various system features (e.g., sensor data, effectors, interface), all evaluated with the words consistent, dependable, and reliable (and a few with the words understandable, accessible), but without reference to **Moral Trust** aspects.

Madsen and Gregor (2000) developed a trust measure for human-computer interaction that contains 25 items covering five content domains: perceived reliability, perceived technical competence,

perceived understandability, faith (in a system's advice and solutions), and personal attachment (to a system). The items in these domains juxtapose subjective experiences (being attached to and understanding the system) with familiar trust expectations regarding ability and reliability (e.g., "The system makes use of all the knowledge and information available to it to produce its solution to the problem"; "The system responds the same way under the same conditions at different times"). The authors did not numerically compare the fit of various factor structures but favored a two-factor solution in which understandability dominated the second factor and virtually all ability and reliability items (along with attachment) hung together in the first factor. No aspects of **Moral Trust** were included.

Not all scales are limited to **Performance Trust** measurement. Jian and colleagues (2000) used a bottom-up approach to examine the semantic field of trust-related expressions. After examining a considerable number of such expressions, the authors arrived at a 12-item scale. Three items capture expectations of reliability (confident in, dependable, reliable), two items capture competence (provides security, harmful outcomes), and four items capture **Moral Trust** characteristics—one representing trustworthiness (integrity) and three representing its absence (deceptive, underhanded, suspicious of intent). However, a subsequent confirmatory factor analysis did not uncover separation between any of these dimensions, only between distrust-related and trust-related items (Spain et al., 2008).

This analysis by no means covered all extant measures of human-robot trust. But it is clear that the large majority does not consider **Moral Trust** aspects, and no specific trust measure appears to exist that reliably captures one or more moral aspects in human-robot interaction. Given the growing similarities between human-robot and human-human relations, there is a need to address the moral aspects of people's trust in robots above and beyond **Performance Trust** aspects. We should not assume that these aspects hold for all robots, but they may hold for some of them; and for those robots, a proper measurement tool must be available. Moreover, as ethical requirements for robot behavior increase (Arkin et al., 2011; Malle & Scheutz, 2019, 2014; Wallach & Allen, 2008), measuring the moral factor of trust is key to evaluating the success of designing such ethical robots. In this dissertation, I work to show the development of a model and measure of trust that captures both the hypothesized **Performance Trust** and

Moral Trust components—for use especially in human-robot interaction, but ideally also in other human-agent interaction.

1.5 Psychometric Scale Design

In this dissertation, I describe the creation of a trust measurement tool called the Multi-Dimensional Measure of Trust (MDMT). The MDMT is intended to help capture a person's evaluation of another agent's trustworthiness based on initial impressions, as well as to capture trust at time points after observing or interacting with the agent. As discussed, trust is complex, rooted in a variety of factors, and can change over time as people learn more from an agent's behavior. Ideally, the MDMT reflects the underlying structure of trust and the measurement tool is robust to variations in the exact method of measurement. I do not make strong claims about the features of the precise rating scale design (e.g., number of response points) that we utilize, with one notable exception discussed here.

Psychometric scale design is a complex and involved process. There is no single, universally accepted method of scale design; rather, there is a vast range of research and opinions on the topic (e.g., Armstrong, 1987; Churchill & Peter, 1984; Garland, 1991; Krosnick & Fabrigar, 1997; Weijters et al., 2010). Although this dissertation does not seek to address all of the nuances of scale design, I report the specific design features of the rating scale that we use for items in the MDMT. We opted to use an 8-point rating scale from 0-7, included a zero endpoint, included endpoint labels of "Not at all" for 0 and "Very" for 7, did not include a midpoint, and included a "Does Not Fit" option. The one feature I will specifically explore is the inclusion of the "Does Not Fit" option.

In all of our research studies, participants always have the option to choose to not answer any particular question (and they are apprised of such in study instructions). Beginning with the very first study described in Chapter 2, we have always included a specific "opt-out" option in various forms in the trust studies we run. I describe the inclusion of this option and the particular implementation in each study throughout the dissertation, highlighting the important role it plays. In the MDMT, we include the option "Does Not Fit" so that participants can mark items they do not believe apply to an agent in the given context, rather than be forced to choose between offering a rating or leaving the question blank. In the

absence of an opt-out option, people may omit responses and such blanks are ambiguous: Did the participant not think that the item fit the agent or context? Did the participant not know what the item meant? Did the participant exercise their right to not answer any particular question in the study for another reason? Did the participant leave the item blank by accident? Although a single opt-out option does not offer full specificity to a multitude of possible motivating reasons, it is arguably better than no option at all—a claim I will support with empirical evidence later on.

1.6 Trust Terminology

Throughout this dissertation I use specific terminology to discuss trust. I reference items related to trust (individual words or phrases related to the semantic space of trust), dimensions of trust (clusters of items that cohere together describing differentiable components of trust), and superordinate factors of trust (clusters of dimensions that cohere together and describe broad conceptual types of trust). To clearly denote when I reference a trust item, dimension, or factor, I italicize "*items*," capitalize "Dimensions," and both capitalize and bold "**Factors**."

There have been two major iterations of the Multi-Dimensional Measure of Trust (MDMT) and the associated trust items, dimensions, and factors; I present a concise description with the associated formatting to familiarize the reader with the terms that will appear in this dissertation. The first iteration of the model and measure of trust, the MDMT v1 (Appendix A; Ullman & Malle, 2019a), included 16 trust items (such as the items *reliable* and *respectable*), four dimensions of trust (Reliable, Capable, Ethical, Sincere), and two superordinate factors of trust (**Capacity Trust**, **Moral Trust**). The factor **Capacity Trust** consisted of the two dimensions Reliable and Capable, as well as all of the associated trust items; the factor **Moral Trust** consisted of the two dimensions Ethical and Sincere, as well as all of the associated trust items.

The second and current iteration of the model and measure of trust, the MDMT v2 (Appendix B; Ullman & Malle, 2020), includes 20 trust items (such as the items *reliable* and *moral*), five dimensions of trust (Reliable, Competent, Ethical, Transparent, Benevolent), and two superordinate factors of trust (**Performance Trust**, **Moral Trust**). The factor **Performance Trust** consists of the two dimensions

Reliable and Competent, as well as all of the associated trust items; the factor **Moral Trust** consists of the three dimensions Ethical, Transparent, and Benevolent, as well as all of the associated trust items.

I offer working definitions of each of the trust factors and trust dimensions below (labels intended to reflect the constituent items and ideas) to help anchor the reader:

- **Performance Trust** = the trustor's confidence that the trustee is capable of completing a task.
    - Reliable = capable of completing a task consistently over time.
    - Competent = capable of completing a task well.

- **Moral Trust** = the trustor's confidence that the trustee will choose the morally right action and not exploit the trustor's vulnerability.
    - Ethical = chooses the morally right action.
    - Transparent = acts openly without deception.
    - Benevolent = acts with others in mind.

Chapter 2: Trust Items Rating Study

Our first study aimed to explore the two-factor conception of trust rooted in our extensive literature review and hypothesized model of human-robot trust. In the spirit of starting simple and trying to capture a big picture possible representation of trust, we designed a study to look at these two broad, superordinate ideas of trust (**Performance Trust**, **Moral Trust**). We theorized that if these are in reality two distinct types of trust, we should be able to identify separate semantic spaces using language that people associate with them. We thus took a semantic inventory of words related to trust in general that we asked people to evaluate on a continuum of these two potentially separate ideas. This study was published in a paper in the HRI conference proceedings (Ullman & Malle, 2018).

2.1 Method

We broadly culled the vocabulary used to describe and query trust in human-automation, human-human, and human-robot 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 items. As part of our selection process, we did not include archaic or uncommon words that are not frequently used in everyday language (e.g., incorruptible); ultimately our goal was to understand how people think about ideas of trust, and thus we wanted to ensure the set of items reflected everyday language. Whereas many previous measures consist of longer sentences related to trust, we opted for simplicity in this task and used single words or very short phrases. With this broad vocabulary in hand, we began the task of looking for potential components of trust (bundles of words that capture, and make measurable, distinct features of trust).

The set of 62 items (arranged alphabetically) that was tested consisted of: *ability*, *able*, *accurate*, *appropriate*, *authentic*, *believable*, *believe in*, *benevolent*, *capable*, *competent*, *complex*, *confide in*, *confident in*, *consistent*, *corrupt*, *count on*, *credible*, *deceptive*, *depend on*, *dependable*, *devoted*, *diligent*, *doubtful*, *dutiful*, *erratic*, *error prone*, *ethical*, *experienced*, *expertise*, *faith in*, *faulty*, *flawed*, *frank*, *function well*, *genuine*, *honest*, *honorable*, *inaccurate*, *incompetent*, *integrity*, *loving*, *loyal*, *meticulous*, *predictable*, *principled*, *reliable*, *rely on*, *reputable*, *respectable*, *responsible*, *rigorous*, *scrupulous*,

sincere, *steadfast*, *steady*, *suspicious of*, *trustworthy*, *truthful*, *unfailing*, *unresponsive*, *untrustworthy*, *virtuous*.

2.1.1 Procedure

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")." Our working labels at that time included **Relational Trust** before we adopted **Moral Trust** (term change addressed in the Discussion below); for the purpose of the study, we did not consider the term "relational trust" to be as clear or commonplace as the alternative "personal trust," which we believed captured the same concept—prompting its use in the study task. Participants completed a comprehension check that consisted of asking participants to match the terms "capacity trust" and "personal trust" to the explanations we shared with them in the initial instructions. Participants were then told they would be 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). We showed participants an example of the slider scale that they would use for the task. All items were presented to each participant in a random order, with an example of an excerpted set of items shown in Figure 2-A. For each word, participants had the option of selecting "Don't know word." 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."

Finally, participants were instructed "If there are any words related to the ideas of either capacity trust or personal trust that you did not see in the previous list of words, please add them below:" and could share additional items for both "capacity trust" and "personal trust."

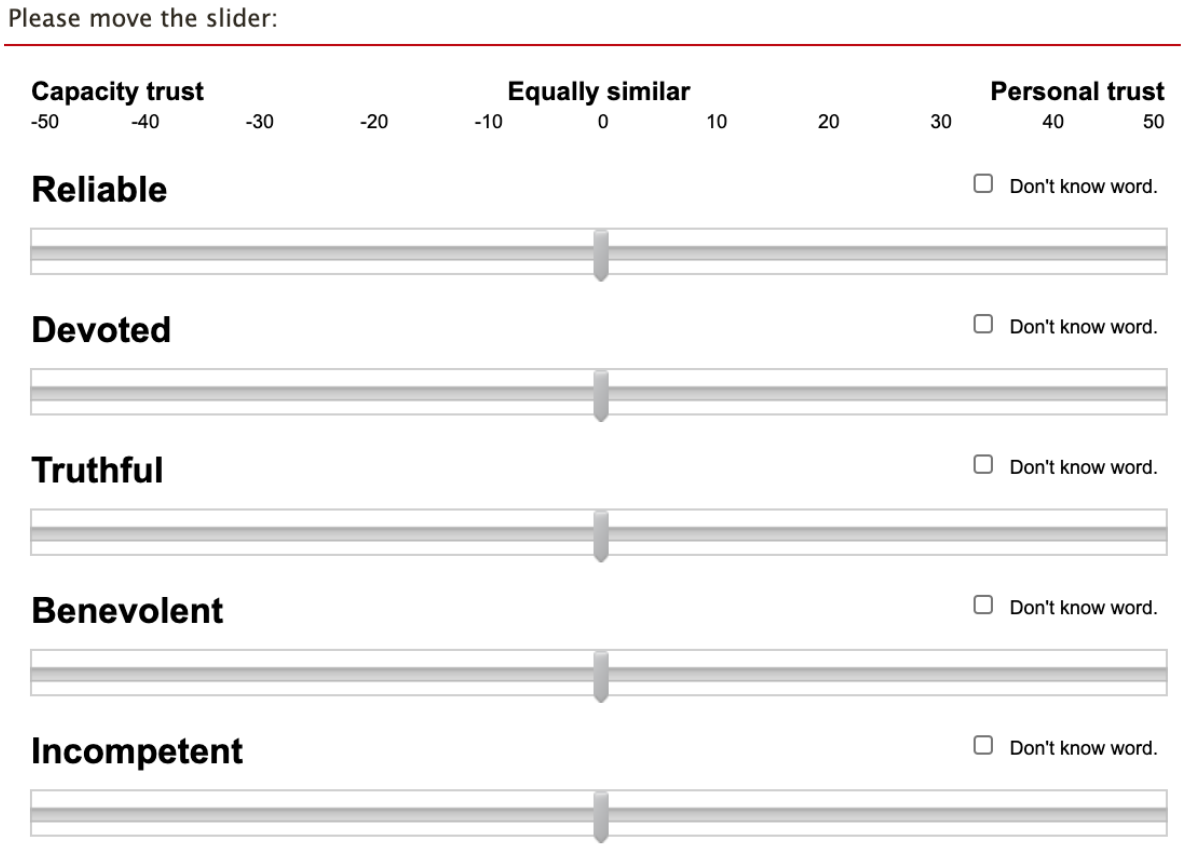


Figure 2-A. Trust Items Rating Study: Example of an excerpted set of items a participant might see in the task.

2.1.2 Participants

A total of 45 adults, recruited via Amazon Mechanical Turk, participated in the approximately 5-minute online study and received compensation of $0.50. We excluded data from five participants who failed a comprehension check of the concepts grounding the dependent variable, for a final sample of n = 40. 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 De Vaus, 2002). The final analysis included 35 adults, and their ratings correlated with the whole group on average at $r = .67$. Participants were not asked any demographic questions.

2.2 Results

When we started this study, we hypothesized we would find two types of trust; however, despite the dependent variable intended to elicit two theorized types of trust, Principal Components Analysis (PCA) from the very start pointed to more than just two components. Our goal was to manage the heterogeneity of the data and identify the best-fitting model of the conceptual space in order to refine our conception of the number and types of trust components. We therefore used an iterative process of PCA and item analysis to analyze the data.

We first 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 reliability, competence, general ethical qualities, and relational qualities, 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 four 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 Reliable (7 items), Capable (8 items), Ethical (11 items), and Sincere (6 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 2-B. 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).

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 2-B. Trust Items Rating Study: Final PCA with 4 components and 5 items each; loadings ≥ .30 displayed.

2.3 Discussion

Our analyses of the semantic space of trust aspects suggested that trust has a multi-dimensional structure. We found four distinct dimensions (Capable, Ethical, Sincere, and Reliable) and created

internally consistent 5-item sets for each. We were somewhat surprised by the fact that a single rating scale with two poles had yielded an emergent four-dimensional structure of items.

There were obvious limitations to this initial study, considering the small sample and entirely exploratory analyses. However, we succeeded in considering a large number of candidate items and systematically narrowing them to four distinct components of trust representations. Our next goals were to replicate and test this structure using different methods in other tasks and to create a parsimonious, intuitive instrument that makes the multi-dimensional nature of trust measurable. This process continued with a trust items sorting task in the subsequent study, where we attempted to replicate the potential dimensions of trust uncovered by this rating study. In this next task we also reintroduced a couple of candidate items, as well as included new items shared by participants in the free response questions that solicited any additional items. We worked to review and integrate items into the candidate set in future studies to refine and validate the robustness of item clusters. Before continuing to the next study, I first make note of a few considerations that shed additional light on some of our thinking and decisions, which evolved with subsequent studies.

Given that we had asked people to rate words only on the single capacity-personal variable, the conceptual structure that emerged was surprising, as these components were strikingly similar to the trust expectations identified by our review of human-automation trust models (Reliable and Capable) and human-human trust models (Ethical and Sincere).

The most conspicuous feature of the experimental design of this first study is likely the absence of any mention of robots—an intentional design characteristic on our part. In the spirit of starting by looking at the big picture of trust, we sought here to explore people's impressions of trust as a concept in general before tying it to particular agents or situations. The absence of robots from experimental design is also apparent in the subsequent trust items sorting task and trust items sorting expansion task, where we attempt to replicate and validate the features of this conceptual framework of trust.

When we were initially designing this study we had arrived at two broad ideas about trust from our extensive literature review, as previously discussed. At the time, we thought about these ideas using

the broad umbrella labels of **Capacity Trust** and **Relational Trust**. Over time we changed these labels to the current superordinate factor labels of **Performance Trust** and **Moral Trust**. We ultimately changed the term from **Capacity** to **Performance** in part because this initial study showed Reliable and Capable as two potentially distinct ideas, such that **Capacity** and its proximity to Capable did not seem to encompass the broader category containing both ideas—**Performance** seemed to fit this criterion. And we changed the term from **Relational** to **Moral** for two reasons: First, we realized that the process of attributing trust to an agent is by nature a relational process across both superordinate factors; second, the word **Moral** seemed to be a more representative and intuitive term for the broader concept, as well as more descriptive than "personal trust." Subsequent studies that were similarly focused on probing the underlying structure of trust did not present participants with superordinate factor labels as part of the tasks; rather, the trust items sorting task (Chapter 3) used the more specific dimension labels, whereas the trust items sorting expansion task (Chapter 5) was free-form in nature and did not use any experimenter-assigned labels. Studies that then examined whether people were responsive to changes along the identified dimensions only used trust items and did not use any factor or dimension labels—as in the trust change task (Chapter 4) and the trust change expansion task (Chapter 6).

The final set of items from the trust items rating task results did not include any "negative" items (e.g., the set included *competent* but not *incompetent*). Even with explicit instructions intended to help avoid or mitigate any confusion, the negative items did not fare well—resulting in the iterative process of item selection quickly eliminating these words. Participants thus demonstrated a clear preference for the positive items. In keeping with the principles of simplicity and parsimony, we have continuously worked to select an intuitive, yet representative, set of candidate trust items; the evolution of this set of items is documented throughout this dissertation. We recognize that variations in terms, including from negation (i.e., *competent* to *incompetent*), could result in different semantic construal of the term; additional research could explore this kind of variation.

One final note comes from the scale presentation used in the study. For each of the items, we gave participants the option to select "Don't know word." The inclusion of this option potentially reduced the

possibility that a participant would have otherwise responded with a 0 rating indicating “equally similar to capacity trust and personal trust” because they did not know the word (or otherwise offered a random rating). This inclusion of an “opt-out” option marked the first instance of a consistent psychometric scale design feature in our trust work, including in the eventual design of the MDMT. As I will show in later studies, consistent inclusion of this type of option offers important insight into participant responses when assessing trust attributions.

Chapter 3: Trust Items Sorting Study

Our second study aimed to test whether the four components we identified in the initial trust items rating study (Chapter 2) would replicate when people were asked to sort candidate items into groups labeled with terms that captured the overarching idea of each category (i.e., Reliable, Capable, Ethical, Sincere). We posited that if the identified components offer an accurate capture of trust then people should agree on grouping specific sets of items along each of the four separate dimensions. This new task was therefore intended as a first replication of the trust item structure identified in the previous study. We also refined the item set based on additional items found in the literature in the interim, as well as included items that appeared in participant free responses in the previous study. This trust items sorting study was published in our book chapter (Malle & Ullman, 2021).

3.1 Method

In addition to replicating the dimensional structure from the previous study, we sought to build in an additional feature that our trust review was increasingly bringing into the spotlight—the idea of attributing features to an agent. Our immersion in the trust literature was making it clear that understanding trust means understanding how one agent thinks about, or evaluates, another agent (conceptualizing trust as a subjective state of the trustor as discussed in Section 1.2.2). We thus took a step toward integrating an agent into our trust tasks by asking participants to consider evaluating person types in the task instructions.

The set of 27 items (arranged alphabetically) that was tested consisted of: *accurate*, *authentic*, *benevolent*, *capable*, *competent*, *consistent*, *diligent*, *ethical*, *genuine*, *honest*, *meticulous*, *principled*, *reliable*, *reputable*, *respectable*, *responsible*, *rigorous*, *scrupulous*, *sincere*, *someone you can confide in*, *someone you can count on*, *someone you can depend on*, *someone you can have faith in*, *steadfast*, *transparent*, *trustworthy*, *truthful*.

3.1.1 Procedure

Participants completed a guided sorting task. Participants were asked to consider 32 items, which included 5 filler items assumed to be unrelated to trust (creative, easy going, humorous, intelligent,

intentional). We expected participants to sort the filler items into an "Other" box as a form of task validation.

Participants were given the following instructions:

> "In this study, you will see a series of words and indicate how well you think they describe different types of people.
>
> Specifically, we would like you to imagine four types of people. One type of person embodies the character trait of being CAPABLE. A second type of person embodies the character trait of being ETHICAL. A third embodies the character trait of being SINCERE. And the fourth embodies the character trait of being RELIABLE.
>
> On the next page, you will drag and drop words from the left side into boxes on the right. Each box represents one of the person types we want you to consider: CAPABLE, ETHICAL, SINCERE, RELIABLE. Your goal is to assign each word to a person type: the type that this word describes the best. There is also one box (OTHER) for words that do not fit any person type. Some of these words will be harder to assign than others; please just follow your intuition for each word."

Participants were told that they could rearrange items as they saw fit if they changed their mind, which afforded participants the opportunity to re-evaluate items as they filled out categories with constituent items. As noted in the instructions, and in alignment with the discussion from the previous study, we included an opt-out option in the form of an "Other" box where participants could place items that did not fit with a particular label. All items were presented to each participant in a random order, with an example of what a participant might have seen displayed in Figure 3-A.

Finally, participants were asked to "Please try to define each of these four ideas (CAPABLE, ETHICAL, SINCERE, RELIABLE) in a few words:" in order to query how participants were thinking about the experimenter-assigned labels we had created based on the initial trust items rating task. We explored participant responses to validate our working understanding of each dimension.

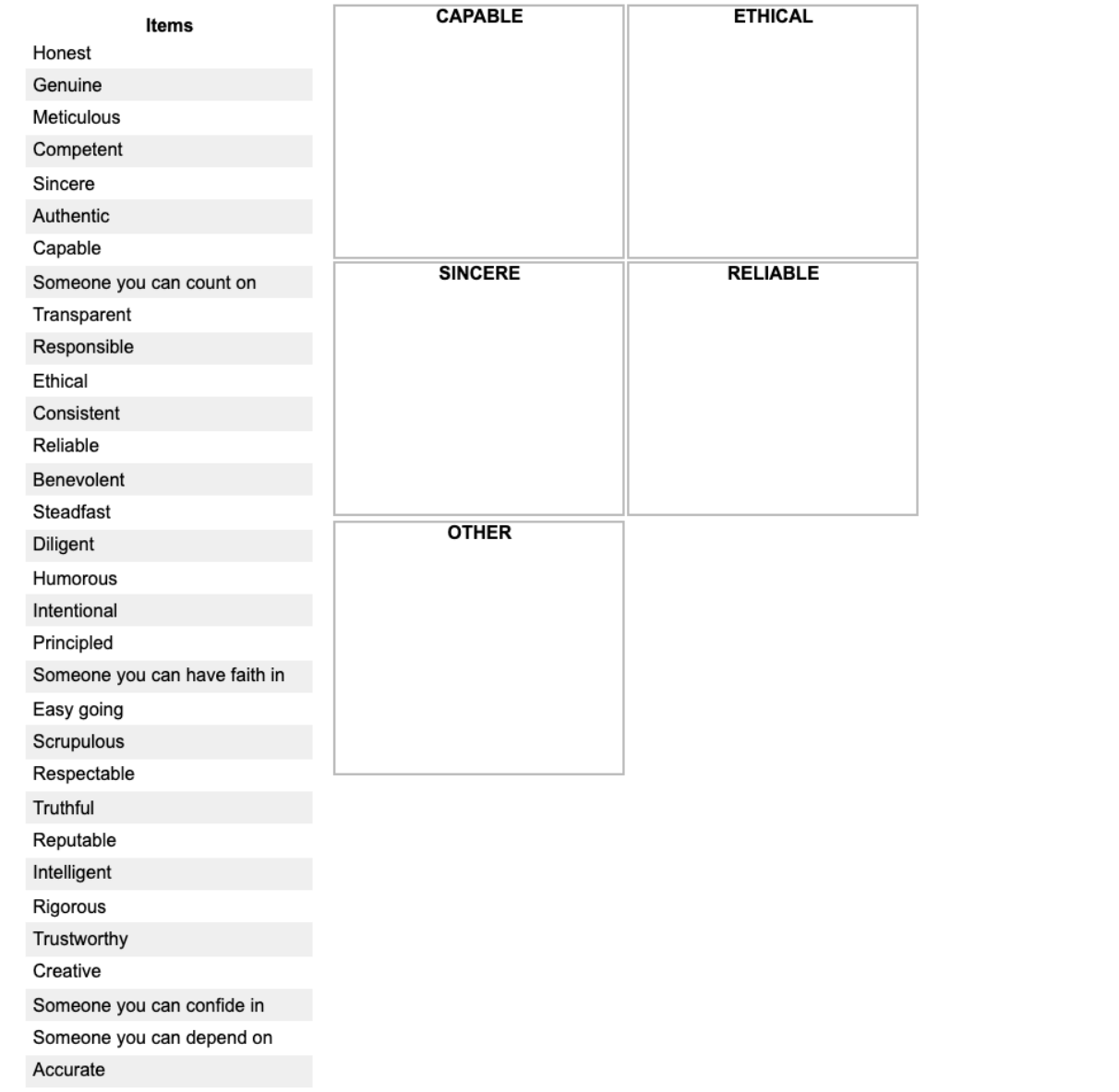


Figure 3-A. Trust Items Sorting Study: Example of the task participants completed, which consisted of an item bank on the left (items presented in random order to each participant) and the labeled sorting boxes on the right with an "Other" box.

3.1.2 Participants

A total of 60 adults, recruited via Amazon Mechanical Turk, participated in the approximately 4-minute online study and received compensation of $0.40. No participants were excluded from the data set, for a final sample of $n = 60$. Participants were not asked any demographic questions.

3.2 Results

We set out to analyze whether participants naturally sorted items into the same groupings that emerged from the PCA in the initial trust items rating study (Chapter 2). We computed sorting consensus scores as the percentages of participants who classified a given item into a category and then grouped items by sorting consensus. Figure 3-B shows the resulting consensus scores, with degree of consensus shaded from dark gray (0%) to green (100%). As is quickly apparent from the figure, the replication succeeded: Each hypothesized dimension of trust expectations contained largely the same items as in the trust items rating study (Chapter 2). These results not only showed that the item clusters that emerged in the initial study appeared to capture consistent constructs, but also that the dimension labels we had assigned (Reliable, Capable, Ethical, Sincere) were intuitive enough that participants did not seem to have difficulty sorting constituent items into these categories (or report difficulty doing so). The filler items were primarily sorted into the "Other" category, with a notable exception of *intelligent* which was sorted into the Capable category. The item *benevolent* was sorted most often into the "Other" category; I comment further on this in the Discussion section.

Sorting Category

Items	CAPABLE	RELIABLE	SINCERE	ETHICAL	OTHER
Capable	92%	2%	3%	2%	2%
Competent	85%	12%	3%	0%	0%
Intelligent	80%	3%	5%	0%	12%
Meticulous	65%	13%	3%	2%	17%
Diligent	54%	31%	7%	3%	5%
Accurate	51%	39%	2%	0%	8%
Rigorous	42%	22%	3%	3%	29%
Reliable	5%	92%	0%	2%	2%
… count on	7%	82%	8%	3%	0%
Consistent	20%	77%	2%	2%	0%
… depend on	15%	70%	5%	7%	3%
Responsible	23%	58%	0%	17%	2%
… have faith in	7%	56%	15%	19%	3%
Steadfast	27%	49%	8%	3%	12%
Sincere	0%	2%	88%	5%	5%
Genuine	2%	2%	88%	3%	5%
Authentic	5%	5%	71%	15%	3%
Honest	2%	0%	51%	47%	0%
Truthful	2%	2%	47%	47%	2%
Transparent	0%	5%	42%	29%	24%
… confide in	5%	27%	33%	25%	10%
Ethical	0%	0%	5%	93%	2%
Principled	3%	3%	3%	88%	2%
Respectable	7%	14%	8%	66%	5%
Trustworthy	3%	23%	27%	47%	0%
Scrupulous	20%	7%	2%	37%	35%
Reputable	15%	28%	12%	35%	10%
Humorous	3%	3%	5%	2%	86%
Easy going	2%	2%	23%	3%	70%
Creative	42%	0%	2%	0%	56%
Benevolent	5%	7%	25%	27%	36%
Intentional	27%	10%	25%	5%	32%

Figure 3-B. Trust Items Sorting Study: Task consensus scores (0%–100%) as the percentages of participants who classified a given item into a category, with items clustered by category/dimension. Filler items in red (creative, easy going, humorous, intelligent, intentional) were assumed to be unrelated to trust and expected to be sorted into the "Other" box. Degree of consensus is shaded from dark gray (0%) to green (100%). Figure from our book chapter (Malle & Ullman, 2021).

3.3 Discussion

The results from the sorting task were quite promising for both the item clusters and dimension labels, with this replication serving as an important step toward validating the emerging multi-dimensional structure of trust. In the initial trust items rating task we had predicted that two clusters might form, but ended up finding four distinct components. This sorting task aimed to validate that people

would naturally sort items into the same semantic clusters as had been derived from our more in-depth data processing in the initial study.

Whereas the initial trust items rating study ensured that items were completely detached from any agent or context, this sorting task began to introduce a simple, abstract version of an agent into the mix. We wanted to avoid any complex scenarios here, taking only an incremental step to avoid introducing too many changes at once. But this was an important step toward studying trust as the subjective state of attributing trustworthy properties to another agent (even an abstract one here). As will be apparent in the subsequent work, introducing increasing complexity with various agents and contexts influences the trust attributions that people make. We referenced humans instead of robots in this task as people have had a much larger set of experiences interacting with a range of human agents with a variety of characteristics, whereas people have likely interacted with few robot agents. As a result, we predicted using human agents as the abstract targets would be much less noticeable than robots and thus would not distract from the focal task.

One item from the task that is worth noting is *benevolent*. As we began to review even more related trust literature, we started to note the recurring idea of benevolence (i.e., acting with goodwill toward an agent) in a number of places. However, the item *benevolent* (Figure 3-B) started to fall away from our candidate item set here in part because it was loading on multiple components (Sincere and Ethical, although both under the **Moral Trust** factor) and because it had a high rate of being sorted into the "Other" category—which in hindsight is likely because it is a less common word, even though the concept of benevolence is a common and important one. This concept re-emerged in our conceptual consideration later on (Chapter 5).

After completing data analysis on the sorting task, we began the process of collating a candidate set of items that reflected the emerging underlying structure of trust in order to create a first draft of a trust measurement tool. We selected three items from the top four items of each cluster and added one new face-valid item to each cluster: *predictable* (for Reliable), *skilled* (for Capable), *candid* (for Sincere), and *has integrity* (for Ethical). A total of 16 items (four in each of the dimensions) thus constituted the

first version of a measurement tool we labeled the Multi-Dimensional Measure of Trust (MDMT). We made the first version of the MDMT, called MDMT v1, publicly available for researcher use dated April 1, 2019 (Appendix A; Ullman & Malle, 2019a). The MDMT was designed to permit calculation of individual subscale scores for each dimension (i.e., from all items within a dimension), factor scores for each factor (i.e., from all items within a factor), and an overall score (i.e., from all items within the measure). The model structure of the MDMT v1 with trust factors, trust dimensions, and trust items is displayed in Figure 3-C.

By creating this first version of the MDMT, we were able to continue the process of item selection and dimension evaluation while beginning to use this framework with actual agents and contexts in various trust scenarios—as we will see in the next study.

MDMT v1				
Trust Factors	**Capacity Trust**		**Moral Trust**	
Trust Dimensions	Reliable	Capable	Ethical	Sincere
trust items	reliable	capable	ethical	sincere
	predictable	skilled	respectable	genuine
	someone you can count on	competent	principled	candid
	consistent	meticulous	has integrity	authentic

Figure 3-C. MDMT v1: Model structure of the MDMT v1, including trust factors, trust dimensions, and trust items.

Chapter 4: Trust Change Study

After developing a working model and measure of trust in the MDMT v1, we set out to test and validate the efficacy of this tool. While we worked to design the conceptual model and measure of trust agnostic to type of agent and context, my primary focus in this line of work was the assessment of trust in human-robot interaction; we therefore set out to first validate the MDMT with trust in robots. We designed a method to test the MDMT for human-robot interaction, with the relevant considerations and challenges discussed below. Later research evaluated the MDMT across types of agents, comparing trust in robot agents with trust in human agents (Chapter 6).

One of the major challenges in designing a research study to test the MDMT was that the individual dimensions are subsumed under the two broader factors; the challenge was thus how to separably examine related individual dimensions in an experimental design without unintentionally collapsing the constituent dimensions into the factors. An additional challenge and important consideration with human-robot interaction work is that we typically seek to understand interactions with robots in general—that is, not confined to any one particular robot. Because robots vary vastly in form and function—from simple non-humanoid, machine-like robots to very complex humanoid robots (Phillips et al., 2018)—it can be challenging to validate the extent to which findings apply across robots. The design challenge here was therefore how to test the MDMT with "robot" as the agent of interest but without anchoring the study too narrowly to a specific type of robot. As noted in the Introduction (Chapter 1), the challenges discussed here are not just abstract concerns, but instead are quite real and impact research (Ullman et al., 2021).

For the trust change task, we ultimately implemented an online study design using vignettes about robots completing various tasks in relevant contexts. Each vignette involved a robot completing a task at one time point and then completing the task differently at a second time point; participants evaluated the robot at both time points using the MDMT. This experimental design offered several key methodological benefits to address the discussed challenges. Most importantly, the vignettes enabled us to carefully craft specific scenarios to probe each theorized dimension of trust one at a time—such that each vignette could

focus on one aspect of trust while attempting to hold the others roughly constant. Vignettes also afforded the opportunity to test a wide range of scenarios with robots that would have otherwise been difficult to test in the lab—such as a robot in an airport security scenario (Appendix C). By using written descriptions, we also avoided anchoring participant expectations to an image of a specific robot; instead, we were able to describe only the details pertinent to the robot in the context.

This was the first study testing the MDMT with robots, and it was published in a paper in the HRI conference proceedings (Ullman & Malle, 2019b).

4.1 Method

We carefully designed the set of vignettes to test whether the MDMT could detect changes in trust attributed to a robot agent in various scenarios. The written narratives either offered increased evidence for trust in an agent or decreased evidence for trust in an agent, and we specifically tried to differentiate between types of trust along the four theorized trust dimensions. However, at the outset we acknowledged that any negative results would leave us uncertain whether we had done an insufficient job in creating the vignettes, or whether the model/measure of trust in the MDMT was inaccurate. The set of vignettes used in this study are shown in Appendix C.

4.1.1 Procedure

The study began with a comprehension check, in part to also check for any bots (automated programs mimicking human behavior). Participants were told, “Recently, there has been an increase of “bots” pretending to be AMT workers. Please help us recognize you as a committed worker by describing the event below in your own words.” Participants were given the following scenario we had generated: “A man ran down the stairs to escape from a building that was getting very hot.” We then asked participants, “Who would you expect to come to the building now and why?” We expected participants to respond with “firefighters,” “fire department,” or some version thereof. Most participants offered responses along these lines, while some participants responded with variants of “AC repair person” or constructed a response that otherwise described the event and indicated comprehension. However, responses such as “good” prompted data exclusion.

Participants received the following task instructions:

> "We are interested in how people interpret various events involving robots. We will give you a very simple event and ask you to rate the described robot on a set of different adjectives. Some of these adjectives will seem appropriate for the robot in the specific situation, while others may not."

Participants were then randomly assigned to their specific condition. 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 the 16 items of the MDMT v1 (Appendix A)—we refer to this as the MDMT pre-test. Then participants were told "One month later you see the same robot." In part two, the robot's behavior changed and participants again completed all 16 items of the MDMT v1—we refer to this as the MDMT post-test.

4.1.2 Design

The vignettes were constructed to fall into one of 32 conditions in a 4 (*Scenario*: airport, bar, nuclear reactor facility, airplane) x 4 (*Targeted Trust Dimension*: Reliable, Capable, Ethical, Sincere) x 2 (*Change Direction*: up, down) between-subjects design, with approximately 25 participants in each condition. *Scenario* refers to four vignettes describing a robot's role-specific behavior in a relevant context (airport, bar, nuclear reactor facility, airplane). *Targeted 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 (up, down) for the relevant dimension of trust. For analysis we would collapse across scenarios, as *Scenario* was not a focal factor; this is because we aimed to show that any effects on the MDMT occurred in general and were not specific to any one scenario. The hypothesis-relevant factors of *Targeted Trust Dimension* and *Change Direction* were designed in general templates but fine-tuned to the specific scenario content. We used participant responses on the MDMT to create dimension subscale scores for data analysis. All vignettes can be seen in Appendix C.

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. The robot complies with most of the airport's procedures when it selects 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.

Participants evaluated trust in the robot twice on all 16 items of the MDMT v1 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 MDMT post-test ratings for each of the four trust dimensions, averaged across their four constituent items ("Does Not Fit" selections 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*), $\alpha = .92$. Capable (*capable*, *skilled*, *competent*, *meticulous*), $\alpha = .92$. Ethical (*ethical*, *respectable*, *principled*, *has integrity*), $\alpha = .81$. Sincere (*sincere*, *genuine*, *candid*, *authentic*), $\alpha = .79$.

4.1.3 Participants

A total of 798 adults, recruited via Amazon Mechanical Turk, participated in the approximately 4-minute online study and received compensation of $0.40. We excluded low-quality responses from 39 participants (who met one or more of the following criteria: failed the comprehension check, had duplicate IP addresses, had a rating range of 0 or 1 across all MDMT items not including "Does Not Fit"), for a final sample of $n = 759$. Participants were not asked any demographic questions.

4.2 Results

First, I make an important terminology distinction between the *Targeted Trust Dimension* factor and the dimension subscale scores: The *Targeted Trust Dimension* (an independent variable) involved specific written vignettes expected to either increase or decrease (according to *Change Direction*) the behavior evidence of the robot in a given scenario for being reliable, capable, ethical, or sincere. The

dimension subscale scores (dependent variables) correspond to the calculated average ratings of items for each of the four theorized dimensions of trust (Reliable, Capable, Ethical, Sincere) in the MDMT v1.

We calculated subscale scores for each of the four trust dimensions (Reliable, Capable, Ethical, Sincere) for both the MDMT pre-test and the MDMT post-test by averaging the specified items for each dimension at each time point (not including items marked by a participant as "Does Not Fit"). We then also used the MDMT pre-test and MDMT post-test item ratings to calculate difference scores for each available item's pre and post ratings (not including "Does Not Fit"). We collapsed across scenarios, as *Scenario* was not a focal factor.

I offer here a more extensive analysis and presentation of the data than in our original paper (Ullman & Malle, 2019b), which only focused on results from the difference scores. I first generated three sets of figures for the study from the MDMT pre-test (Figure 4-A), the MDMT post-test (Figure 4-B), and the MDMT difference scores (Figure 4-C). Each figure has two distinct graphs split out by the two levels of the *Change Direction* factor: up (increasing evidence for trust on a specified dimension) and down (decreasing evidence for trust on a specified dimension).

I started by looking at the results from the MDMT pre-test (Figure 4-A). In theory, the two graphs for the up and down levels in the MDMT pre-test should be nearly identical, as participants had only read the first part of the vignette at this point (i.e., they had not yet experienced the behavior change to prompt any differences between the two levels of the factor). I split out the graphs by this factor for the pre-test to allow for consistent comparison with the post-test and difference scores graphs, where this factor split is highly relevant. In addition, the real-world variation seen in the pre-test between the two levels of the factor is relevant to the baseline when comparing the two time points for each specific level. Overall, however, these graphs are quite similar and show consistency across participants. Examination of the MDMT pre-test graphs reveals two key findings. First, the Reliable and Capable subscale scores, as well as the Ethical and Sincere subscale scores, cohere quite closely with one another with mostly overlapping values; these two line groupings correspond to the theorized superordinate trust factors of **Performance Trust** and **Moral Trust**, respectively. These baseline data indicate the similarity of dimensions within

each factor. Second, these graphs show a substantial gap for the robot between baseline **Performance Trust** subscale scores and baseline **Moral Trust** subscale scores—which were considerably lower. This shows what I henceforth refer to as a baseline Moral-Performance trust gap for robot agents.

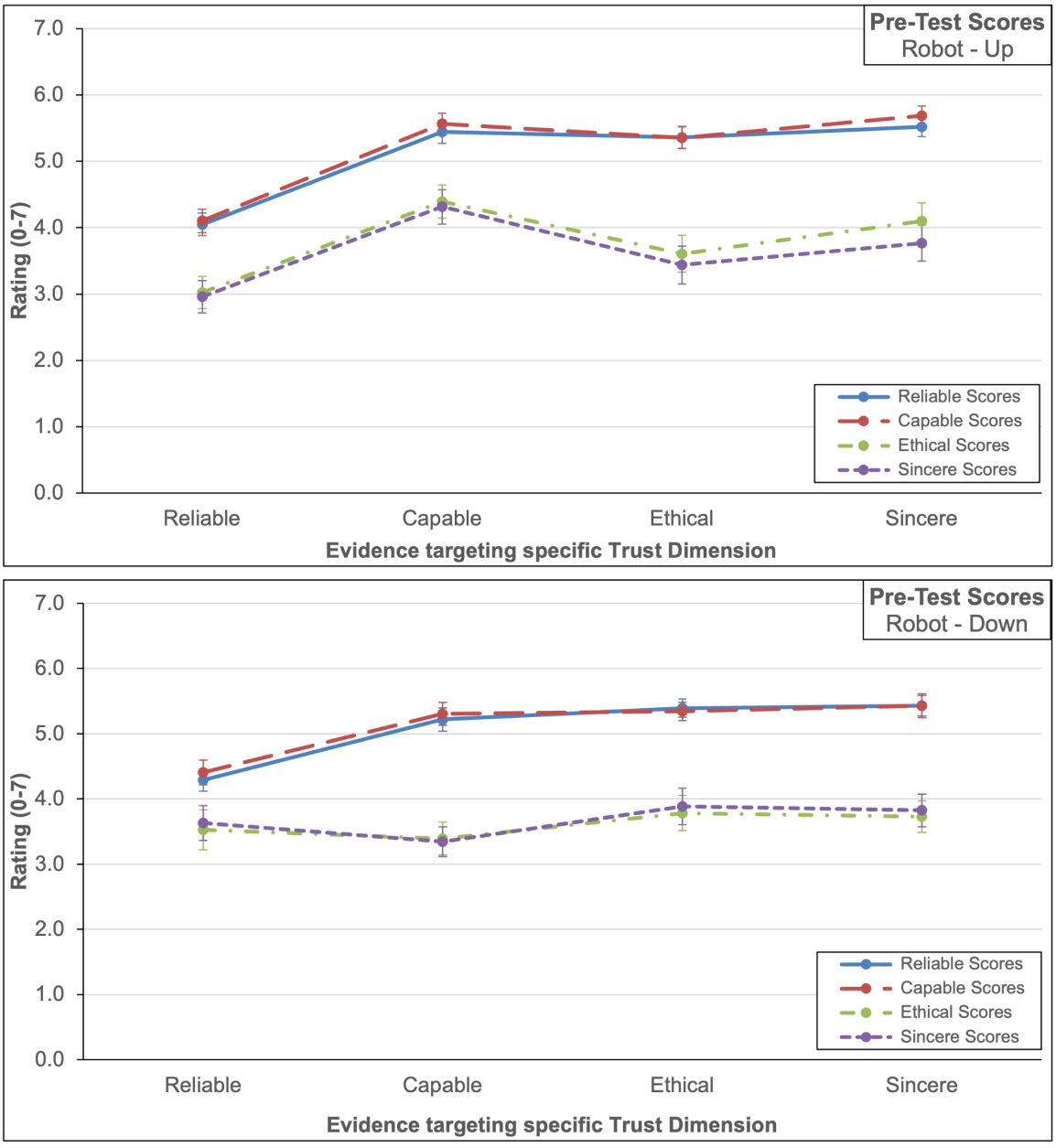
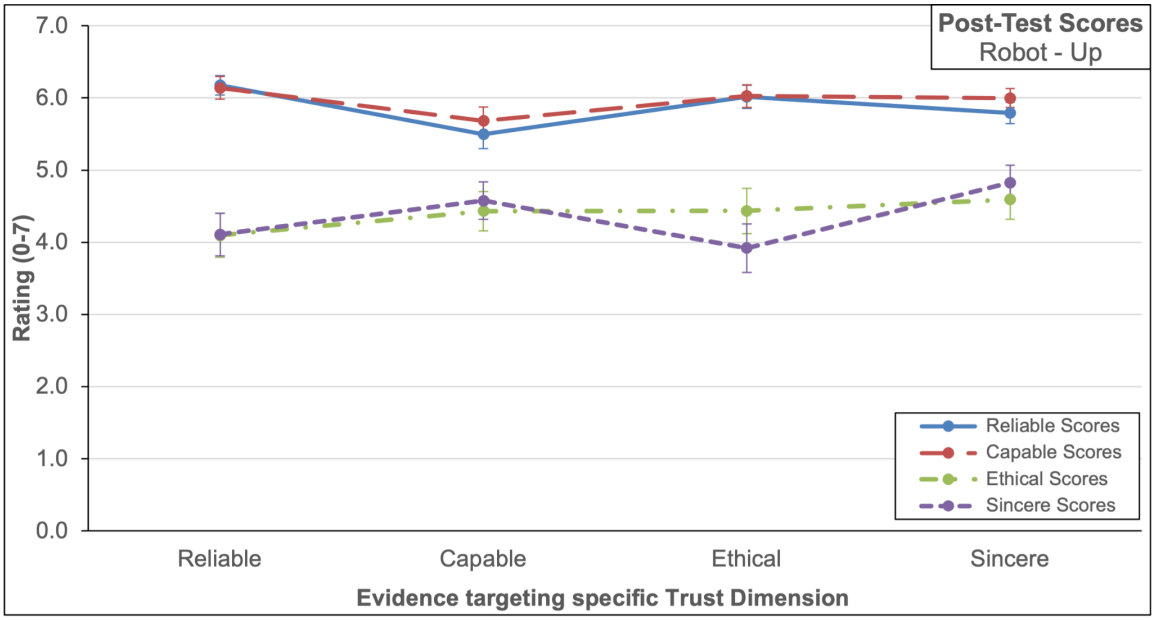


Figure 4-A. Trust Change Study: Each graph shows MDMT pre-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

I next examined the results from the MDMT post-test (Figure 4-B). We can compare each graph to the corresponding pre-test graph (Figure 4-A) to identify overall patterns of change in ratings along the different dimensions. The key observation here is that ratings generally increased for the up level of *Change Direction*, whereas ratings generally decreased for the down level of *Change Direction*—indicating that people responded to salient changes in an agent's behavior with corresponding changes in trust that were captured by the MDMT. The changes between the MDMT pre-test and the MDMT post-test are most easily analyzed using difference scores graphs (Figure 4-C), which we use to evaluate each of the dimensions and the corresponding subscales.



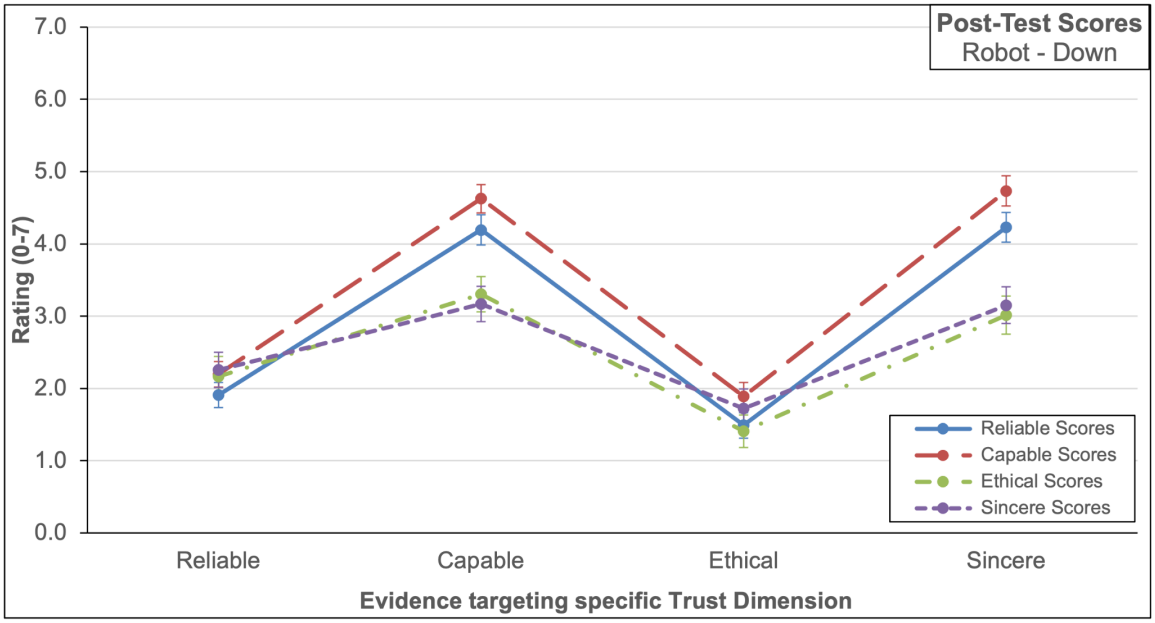


Figure 4-B. Trust Change Study: Each graph shows MDMT post-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

The difference scores graphs illustrate the changes in trust along each dimension captured by the MDMT subscales for the up and down levels of *Change Direction* (Figure 4-C). In theory, if the vignettes worked precisely as intended and only prompted change along a single dimension (Reliable, Capable, Ethical, or Sincere), we would see a change in each corresponding subscale score only for the specific

Targeted Trust Dimension evidence. For example, for the Reliable *Targeted Trust Dimension* on the x-axis, we would expect to see a change only for the Reliable subscale scores on the y-axis. From what we have seen in previous studies, and as we see here, the dimensions appear related to one another—especially within their respective superordinate factors.

We assessed the sensitivity of each of the four MDMT subscales to the manipulations of *Change Direction* along each *Targeted Trust Dimension*, yielding four between-subjects ANOVAs (one for each subscale), examining a 2 (*Change Direction*: up, down) x 4 (*Targeted Trust Dimension*: Reliable, Capable, Ethical, Sincere) design. Each of the four subscales was highly sensitive to the behavior change manipulations overall, collapsed across *Targeted Trust Dimension* (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 η^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 for 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 Reliable (η^2 = .06 to .24) and changes in Ethical (η^2 = .13 to .23). Changes in Sincere or Capable were weak (η^2 = 0 to .03). There is a first hint here of a differentiation between the Ethical and Sincere dimensions that can be seen in the robot up level condition in Figure 4-C, such that the Ethical and Sincere subscale scores show a crossover pattern. Specifically, increased evidence for the Ethical *Targeted Trust Dimension* resulted in greater Ethical subscale scores without substantially impacting Sincere scores, while increased evidence for the Sincere *Targeted Trust Dimension* resulted in greater Sincere subscale scores without substantially impacting Ethical scores. This was particularly promising because it showed a possible differentiation of two constituent trust dimensions from within a superordinate trust factor (**Moral Trust**).

Even without full dimensional specificity, the results did show an important pattern. In both gains in trust and losses in trust (top graph and bottom graph in Figure 4-C, respectively), the measures of Reliable and Capable moved in tandem, and so did the measures of Ethical and Sincere. It appeared that

the two superordinate factors of trust were sensitive to a robot's change in behavior: **Performance Trust** (Reliable, Capable) and **Moral Trust** (Ethical, Sincere).

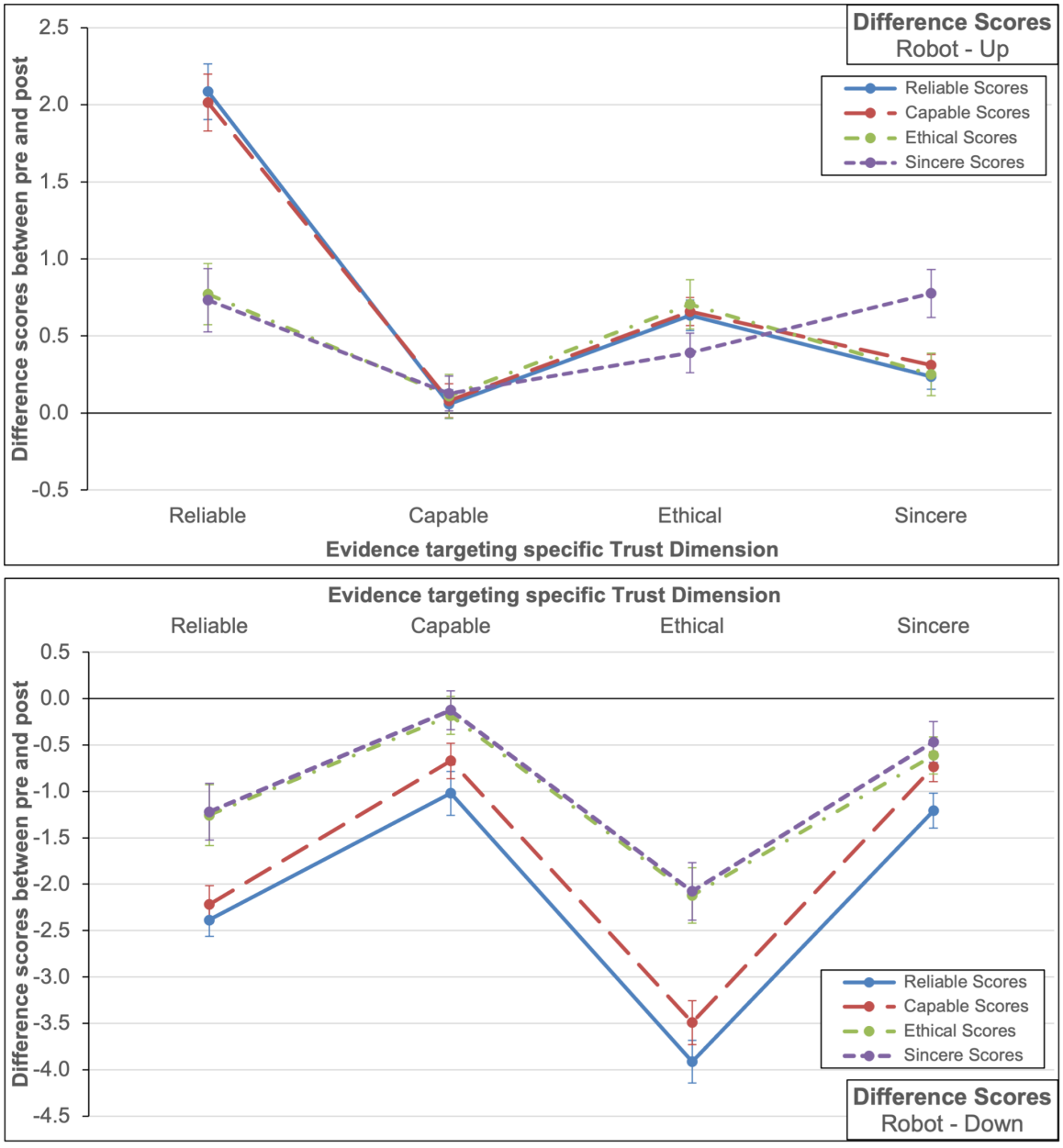


Figure 4-C. Trust Change Study: Each graph shows MDMT difference subscale scores (legend) on appropriate scaling (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Change Direction* factor (top graph shows the up level for evidence increasing trust, bottom graph shows the down level for evidence decreasing trust). Error bars show *SE*.

4.3 Discussion

Our goal with this study was to test whether the MDMT could capture salient changes in trust in a robot, and whether the conceptual structure and corresponding measurement tool could differentiate between hypothesized types of trust. Overall, the study validated the use of the MDMT in general, but without full dimensional specificity—this left open the question of whether this was due to the way in which we attempted to probe some of the dimensions in the vignettes (i.e., because of the experimental design), or whether this was due to gaps in the conceptual structure of the MDMT itself. The behavior changes for Reliable and Ethical had notable impact on the subscales of trust, whereas the behavior changes for Capable and Sincere had little impact on the subscales. Future work would thus need to explore the validity of these dimensions using updated vignettes—which we present later in the trust change expansion effort (Chapter 6).

The subscales of the MDMT, at least as the pairs for Reliable and Capable (**Performance Trust**) and Ethical and Sincere (**Moral Trust**), were responsive to varying degrees to changes in robot behavior. The dimensions within each pair showed similar baseline ratings of trust (Figure 4-A) and moved in tandem in response to changes in trust information (Figure 4-C). These results were promising for the proposed two-factor superordinate structure of trust, while also prompting questions about the extent to which the theorized dimensions are functionally distinct in people's attributions of trust to an agent. Although we had carefully designed the vignettes and the associated behavior change descriptions for all of the conditions, it was quite possible that we simply did not adequately target some of the dimensions—in particular Capable and Sincere. As we continued to evaluate the MDMT conceptual structure and associated item clusters, the results from this study prompted us to further consider the specific number and composition of trust dimensions (Chapter 5).

One pronounced finding from the present study comes from the *Change Direction* factor: Inspection of the scaling of the bottom graph (down level) compared to the top graph (up level) in Figure 4-C shows that trust loss was much larger than trust gain. We had attempted to design comparable behavior change conditions for the down and up levels in scenarios, such that this points to an important

trust loss effect that has potential ramifications for understanding human-robot trust. Trust loss and trust repair is a critical domain for human-robot interaction, detailed in part in a journal article I worked on together with collaborators (Baker et al., 2018). In the current line of research, we therefore sought to understand whether this trust loss effect was a feature of having a robot agent in the scenarios, whether it was due to a general trend in trust changes regardless of the focal agent, or whether it was a combination of these and/or possible other factors. We thus investigated this problem space by including an *Agent* factor (robot, human) in a subsequent study (Chapter 6) to test whether the trust loss effect was specific to robot agents or shared by human agents. The later study also sought to more broadly compare evaluations using the MDMT for robot agents with human agents. As mentioned, the MDMT was originally conceptualized as a measure to capture trust agnostic to the type of agent, though with the intention of designing a measure for trust in human-robot interaction. As a result, analysis with human agents would permit evaluation of the MDMT as a measurement tool across different types of agents, as well as to potentially benchmark findings with robot agents against human agents (Chapter 6).

In general, the trust change study pointed to the MDMT as a promising instrument for research on the dynamics of trust while also prompting several questions later addressed in an updated experimental design (Chapter 6). But first we worked to reconsider the potential dimensional structure and associated item clusters of the MDMT, as detailed in the next chapter; in this effort we aimed to refine the MDMT based on the present results and incorporate new insights that had since been gleaned from the trust literature.

Chapter 5: Trust Items Sorting Expansion Study

We had begun to question the notable absence of the idea of "benevolence" from the trust dimensions. This idea—that of acting with others in mind—had appeared in a number of foundational works on trust as discussed in the Introduction (Chapter 1). We thus set out to test whether benevolence exists as a separable dimension of trust, such that perhaps we simply had not captured sufficient evidence for it in the preceding studies.

The original trust items sorting task (Chapter 3) we conducted had offered participants a seemingly intuitive means of testing items with the set of hypothesized trust dimensions, such that we opted to run a replication using an expanded, free-form version of this task. We tested previously selected items from the MDMT v1 together with other candidate trust items, including items related to the idea of benevolence that used intuitive, commonplace phrasing.

5.1 Method

The original trust items sorting task was easily understood and completed by participants. This new expanded version of the task therefore mirrored the original, with one major change and one associated addition. In the original task we had presented participants with five boxes into which they could sort items—four boxes with labels we had assigned to the four dimensions that had been extracted from the trust items rating task (Reliable, Capable, Ethical, Sincere) and one "Other" box. In this new study we opted for a tougher, more powerful replication challenge. Rather than use boxes with experimenter-assigned labels, instead we had participants group items from the item bank together into boxes that did not have any labels and use however many boxes/groupings they believed appropriate. Consistent with the theme of including a not applicable option, we offered participants the option to sort items into an "Other" box.

This updated experimental design offered two major benefits over the previous design. First, it enabled us to test whether people naturally form the same groupings of items we had derived without using forced anchor labels for categories. Second, it allowed people to form however many groups they thought represented the constituent items—without confining them to a forced four-dimensional structure

(or five-dimensional structure, as we were now hypothesizing with the possibility of a Benevolent dimension). This new design would thus offer a robust means of testing the number and types of dimension groupings for the conceptual space of trust.

The set of 41 items (arranged alphabetically) that was tested consisted of: *accurate*, *acts with others in mind*, *authentic*, *believable*, *benevolent*, *candid*, *capable*, *caring*, *competent*, *considerate*, *consistent*, *dependable*, *diligent*, *ethical*, *fair*, *frank*, *functions well*, *genuine*, *has expertise*, *has integrity*, *honest*, *kind*, *looks out for others*, *loyal*, *meticulous*, *moral*, *open*, *predictable*, *principled*, *reliable*, *respectable*, *responsible*, *shows goodwill*, *sincere*, *skilled*, *someone you can count on*, *someone you can have faith in*, *steadfast*, *transparent*, *trustworthy*, *truthful*.

No filler items unrelated to trust were included in this new version of the sorting task. With the new free-form design, we wanted to give participants complete autonomy over item grouping formations and aimed to lower the perceived threshold of difference between candidate items that do and do not fit with trust. In other words, it is arguably straightforward and easy to sort out a filler item like "creative" into the "Other" box as an item unrelated to trust; it is therefore possible that including "creative" might raise the threshold to sort out other less useful, but still semi-related, items into the "Other" box. By removing the filler items, we aimed to make it easier for participants to sort out any items they did not think cohered with the other trust items/groups. As the primary purpose of including filler items in the original sorting task was task validation, we simply removed the filler items and instead instituted a different means of task validation described below.

5.1.1 Procedure

Participants completed a free-form sorting task. Participants were asked to consider 41 items, which did not include any filler items.

Participants received the following task instructions:

> "There are many reasons why you might trust a person. For example, people can exhibit certain character traits that prompt us to trust them.
>
> In this study, you will see a series of words or phrases ("items") that refer to such character traits. We ask you to sort these items into groups you feel belong together:

> phrases that are very similar to each other but that are, as a group, distinct from other groups.
>
> You will simply drag items from a list and drop them into a number of boxes. These boxes represent the groups you create.
>
> You can create as many groups as you think are necessary, but please create at least 2.
>
> On the next page, you will see the list of items on the left. You will drag and drop items from the left into boxes on the right. You do not need to use all the available boxes, but use at least 2.
>
> Some of the items will be harder to assign than others; just follow your intuition. Items that do not seem to fit anywhere can be dropped into the "OTHER" box."

As in the original sorting task, participants were told that they could rearrange items as they saw fit if they changed their mind, which afforded participants the opportunity to re-evaluate items as they formed categories out of constituent items. In addition to six numbered boxes (from 1 to 6), the "Other" box was still included in this task for items that participants thought did not fit. All items were presented to each participant in a random order, with an example of what a participant might have seen displayed in Figure 5-A.

Finally, participants were asked, "Once you have finished the sorting task, we would like you to pick the word from each group that best represents the entire group. You do not need to label the "OTHER" group." This was displayed on the same page so that participants could see the item groupings they had just created (Figure 5-B).

Participants were asked demographic questions (self-reported age, gender, and race/ethnicity) at the conclusion of the task.

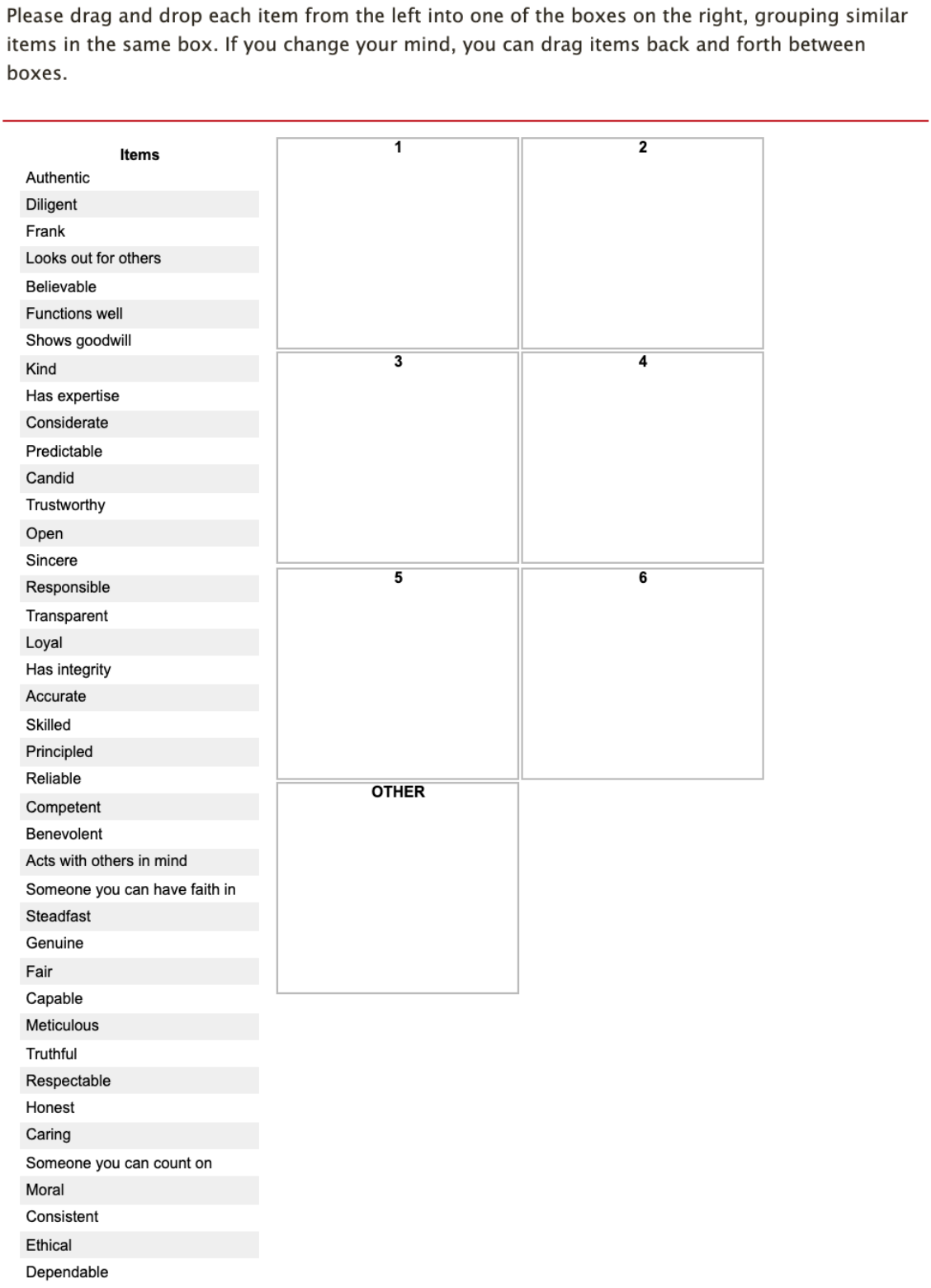


Figure 5-A. Trust Items Sorting Expansion Study: Example of the task participants completed, consisting of an item bank on the left (items presented in random order to each participant) and the unlabeled sorting boxes on the right with an "Other" box.

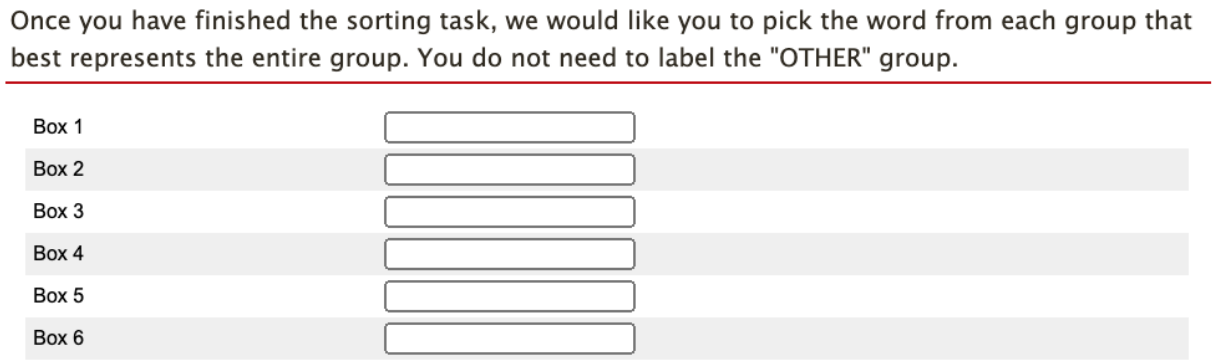


Figure 5-B. Trust Items Sorting Expansion Study: Example of the free response boxes in the task where participants assigned labels to the groupings of trust items they had created.

5.1.2 Participants

A total of 100 adults, recruited via Amazon Mechanical Turk using TurkPrime (Litman et al., 2017), participated in the approximately 10-minute online study and received compensation of $1.00. We excluded data from eight participants (five who had partially missing data and three who failed to follow the instructions), for a final sample of $n = 92$. Participants in the final sample had a mean age of 38.72 years ($SD = 11.57$); they self-reported gender as 52 "Male," 40 "Female," and 0 "Other"; and they self-reported race/ethnicity as 78 "Non-Hispanic White or Euro-American," 7 "Black, Afro-Caribbean, or African American," 4 "Asian or Asian American," 2 "Latino or Hispanic American," 1 "Multi-racial," 0 "Native American or Alaskan Native," and 0 "Other."

5.2 Results

By affording participants the opportunity to freely sort and group items as they saw fit, we were left with various numbers of groupings and compositions of groupings. This posed an interesting analysis challenge that we tackled in the following ways.

We first examined the number of categories that participants created. We calculated the number of sorting categories (i.e., number of boxes) participants used to distribute items, and display this distribution in Figure 5-C. This distribution showed that the majority of participants (73.91%) sorted items into four or more categories.

We then examined the actual dimension labels that participants assigned to categories by calculating the frequency of label assignment. These participant-assigned labels were selected by participants from terms they thought represented each item grouping they had created. These labels can thus be interpreted as what participants considered to be the most intuitive umbrella terms for different semantic trust categories extracted from the presented item set. We also used responses as a form of task validation, excluding participant responses that were not conceivably related to the assigned task.

We selected the most common participant-assigned labels in order of frequency and classified them into columns in Figure 5-D according to updated experimenter-hypothesized labels (Reliable, Competent, Ethical, Transparent, Benevolent; see the Discussion below for detail about dimension label changes from “Capable” to “Competent” and from “Sincere” to “Transparent”) and a single-item candidate for a general capture of trust (*trustworthy*). The first five participant-assigned labels fit perfectly under the five trust dimensions (first row), and we placed the item *trustworthy* aside as a single-item candidate for measuring trust; the second five participant-assigned labels fit again perfectly under the five dimensions (second row); and then the third five labels (third row). In fact, the top five most frequent labels aligned with the five theorized dimensions of trust we aimed to test in this study. We noted the relatively high frequency of the item *trustworthy*, which I discuss further after the data structure extraction.

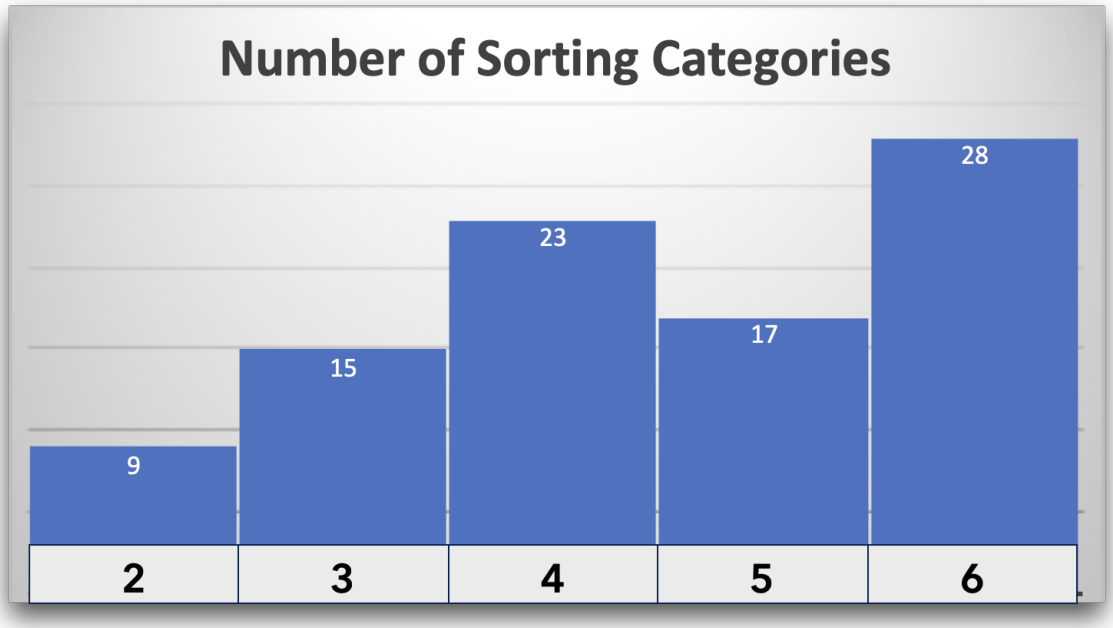


Figure 5-C. Trust Items Sorting Expansion Study: Frequencies of participants who chose a particular number of trust sorting categories (2, 3, 4, 5, or 6).

Most Common Labels for Trust Categories		Experimenter-Assigned Labels					
		Reliable	Competent	Ethical	Transparent	Benevolent	*trustworthy*
Participant-Assigned Labels	First 5	dependable - 23	competent - 29	moral - 30	honest - 28	caring - 21	trustworthy - 21 (single-item capture of trust)
	Second 5	reliable - 14	skilled - 17	loyal - 15	authentic - 15	kind - 20	
	Third 5	responsible - 14	capable - 14	ethical - 11	candid - 13	benevolent -13	

Figure 5-D. Trust Items Sorting Expansion Study: Frequencies of the most common trust category labels that participants used to characterize the categories they formed. Columns arranged by five dimensions using experimenter-assigned labels (Reliable, Competent, Ethical, Transparent, Benevolent) and the single-item candidate for the capture of trust (*trustworthy*). Rows arranged by the set of first five most frequent labels, then second five, and then third five.

The subsequent analysis examined the optimal number of dimensions that represented the overall data structure (i.e., the number of participant-formed categories and the items within categories). To identify the underlying relationships between the candidate items, we used the participant data to represent the consistent item interrelationships (i.e., covariations).

We began by calculating the frequency with which each item co-occurred with every single other item—that is, how often any two items were placed into the same category (i.e., box) by participants. Figure 5-E shows these co-occurrence proportions. Each cell shows the proportion of time a specific item was placed in the same bin as another item; the diagonal shows the proportion of time each item was

placed into any one of the trust bins (i.e., less than 1.00 on the diagonal indicates that an item was sometimes classified into the "Other" bin and not considered related to trust by a participant). The items are organized alphabetically on both axes. Some items (such as *trustworthy* and *reliable*) were placed into bins related to trust 100% of the time, whereas the item with the lowest percentage (*predictable*) was placed into bins related to trust 90% of the time.

Next, we transformed Figure 5-E into Figure 5-F to show clusters of items manually arranged by hypothesized clusters, in order to visually examine how well the actual item co-occurrences re-created the expected patterns; Figure 5-F displays a heat map that has approximately five clusters of related items. Co-occurrences of items with themselves on the diagonal are grayed out. The items are arranged on both axes based on the co-occurrence clusters. This visualization of the co-occurrence patterns offered additional evidence of identifiable clusters of items people considered related to one another in the domain of trust. Rate of co-occurrence is shaded from white (0%) to red (100%). Lighter shading (white) indicates items had fewer co-occurrences, while darker shading (red) indicates higher co-occurrences. The heat map suggests that some clusters were more tightly-knit than others, as can be seen by the darker shading in two of the corners—the first cluster at the top left for the Competent dimension and the fifth cluster at the bottom right for the newly probed Benevolent dimension. This is similar to simple structure in Principal Components Analysis (PCA), where items load onto a single component without loading onto other components. The third cluster in the middle for the Transparent dimension is also darkly shaded and tightly-knit, with the second cluster for the Reliable dimension slightly lighter and the fourth cluster for the Ethical dimension even lighter. The additional fuzziness (i.e., shading surrounding the immediate cluster) around the Ethical dimension does not point to the Ethical dimension as any less relevant for trust—rather, it shows that people evaluated the Ethical items as somewhat more related to other items...suggesting that the Ethical items are not as separable from other items. This idea becomes more clear after the subsequent analysis.

	Accura	Acts_w	Authe	Believ	Benev	Candic	Capab	Caring	Compe	Consid	Consis	Depen	Diliger	Ethica	Fair	Frank	Functi	Genuir	Has_e	Has_ir	Hones	Kind	Looks_	Loyal	Metici	Moral	Open	Predic	Princip	Reliab	Respe	Respo	Shows	Sincer	Skilled	Somet	Somet	Steadf	Transp	Trustw	Truthfu
Accurate	0.97	0.07	0.17	0.18	0.12	0.20	0.57	0.08	0.58	0.07	0.35	0.18	0.49	0.14	0.14	0.23	0.52	0.17	0.62	0.18	0.17	0.08	0.08	0.12	0.55	0.09	0.16	0.25	0.21	0.22	0.21	0.26	0.08	0.14	0.64	0.15	0.12	0.37	0.17	0.17	0.20
Acts_with_others_in	0.07	1.00	0.22	0.17	0.61	0.20	0.05	0.72	0.09	0.70	0.11	0.20	0.10	0.33	0.41	0.10	0.09	0.26	0.04	0.28	0.27	0.71	0.75	0.41	0.03	0.39	0.25	0.12	0.24	0.20	0.20	0.12	0.70	0.40	0.07	0.33	0.37	0.08	0.14	0.22	0.23
Authentic	0.17	0.22	0.98	0.55	0.23	0.55	0.10	0.25	0.10	0.26	0.15	0.17	0.11	0.36	0.27	0.50	0.10	0.71	0.09	0.35	0.59	0.24	0.22	0.32	0.07	0.29	0.52	0.23	0.28	0.18	0.35	0.17	0.20	0.52	0.10	0.18	0.28	0.15	0.61	0.41	0.59
Believable	0.18	0.17	0.55	0.99	0.22	0.43	0.13	0.21	0.11	0.23	0.17	0.26	0.14	0.33	0.36	0.43	0.10	0.53	0.10	0.34	0.61	0.22	0.21	0.29	0.11	0.37	0.46	0.21	0.26	0.24	0.32	0.21	0.22	0.53	0.11	0.27	0.46	0.14	0.53	0.49	0.63
Benevolent	0.12	0.61	0.23	0.22	0.93	0.23	0.08	0.63	0.04	0.53	0.11	0.15	0.12	0.36	0.36	0.13	0.09	0.20	0.05	0.28	0.29	0.61	0.57	0.30	0.11	0.39	0.24	0.11	0.26	0.18	0.20	0.14	0.65	0.28	0.08	0.21	0.26	0.13	0.14	0.21	0.26
Candid	0.20	0.20	0.55	0.43	0.23	0.95	0.03	0.18	0.08	0.21	0.11	0.12	0.13	0.28	0.24	0.62	0.09	0.55	0.08	0.21	0.46	0.21	0.17	0.16	0.13	0.29	0.52	0.16	0.23	0.13	0.22	0.11	0.20	0.41	0.09	0.16	0.23	0.15	0.61	0.28	0.48
Capable	0.57	0.05	0.10	0.13	0.08	0.03	0.99	0.09	0.66	0.09	0.38	0.29	0.57	0.10	0.12	0.08	0.68	0.07	0.76	0.15	0.05	0.07	0.10	0.12	0.63	0.04	0.10	0.22	0.21	0.35	0.26	0.38	0.04	0.10	0.75	0.21	0.12	0.41	0.09	0.14	0.09
Caring	0.08	0.72	0.25	0.21	0.63	0.18	0.09	0.99	0.05	0.74	0.10	0.21	0.09	0.39	0.41	0.13	0.08	0.25	0.01	0.24	0.33	0.86	0.67	0.41	0.04	0.42	0.34	0.09	0.23	0.26	0.27	0.14	0.74	0.41	0.03	0.25	0.33	0.09	0.21	0.25	0.26
Competent	0.58	0.09	0.10	0.11	0.04	0.08	0.66	0.05	0.98	0.09	0.36	0.32	0.54	0.14	0.12	0.09	0.66	0.08	0.71	0.20	0.09	0.03	0.09	0.13	0.60	0.04	0.08	0.18	0.23	0.35	0.27	0.42	0.05	0.11	0.68	0.24	0.18	0.37	0.08	0.18	0.09
Considerate	0.07	0.70	0.26	0.23	0.53	0.21	0.09	0.74	0.09	1.00	0.12	0.23	0.09	0.36	0.42	0.15	0.10	0.32	0.05	0.32	0.32	0.72	0.65	0.40	0.08	0.45	0.27	0.13	0.28	0.24	0.30	0.14	0.65	0.46	0.04	0.29	0.33	0.07	0.18	0.30	0.29
Consistent	0.35	0.11	0.15	0.17	0.11	0.11	0.38	0.10	0.36	0.12	0.98	0.51	0.38	0.18	0.23	0.18	0.36	0.14	0.32	0.27	0.16	0.08	0.13	0.34	0.34	0.10	0.17	0.51	0.18	0.48	0.33	0.50	0.10	0.15	0.36	0.46	0.27	0.48	0.18	0.33	0.17
Dependable	0.18	0.20	0.17	0.26	0.15	0.12	0.29	0.21	0.32	0.23	0.51	1.00	0.34	0.21	0.24	0.11	0.25	0.24	0.20	0.22	0.20	0.22	0.24	0.38	0.25	0.23	0.16	0.39	0.22	0.72	0.38	0.61	0.21	0.22	0.23	0.67	0.48	0.42	0.18	0.48	0.22
Diligent	0.49	0.10	0.11	0.14	0.12	0.13	0.57	0.09	0.54	0.09	0.38	0.34	0.97	0.14	0.15	0.14	0.55	0.11	0.53	0.17	0.11	0.11	0.14	0.13	0.57	0.10	0.12	0.24	0.23	0.33	0.24	0.46	0.05	0.13	0.59	0.20	0.08	0.49	0.14	0.18	0.08
Ethical	0.14	0.33	0.36	0.33	0.36	0.28	0.10	0.39	0.14	0.36	0.18	0.21	0.14	0.99	0.50	0.22	0.10	0.28	0.10	0.57	0.42	0.37	0.35	0.39	0.14	0.73	0.30	0.14	0.62	0.22	0.40	0.22	0.38	0.34	0.10	0.18	0.32	0.13	0.30	0.40	0.46
Fair	0.14	0.41	0.27	0.36	0.36	0.24	0.12	0.41	0.12	0.42	0.23	0.24	0.15	0.50	0.96	0.24	0.17	0.25	0.09	0.47	0.46	0.42	0.36	0.35	0.08	0.55	0.27	0.14	0.40	0.25	0.37	0.22	0.40	0.40	0.13	0.29	0.34	0.14	0.24	0.38	0.40
Frank	0.23	0.10	0.50	0.43	0.13	0.62	0.08	0.13	0.09	0.15	0.18	0.11	0.14	0.22	0.24	0.93	0.04	0.46	0.09	0.27	0.43	0.14	0.12	0.21	0.13	0.18	0.51	0.24	0.27	0.12	0.27	0.13	0.15	0.38	0.10	0.15	0.24	0.22	0.59	0.28	0.47
Functions_well	0.52	0.09	0.10	0.10	0.09	0.09	0.68	0.08	0.66	0.10	0.36	0.25	0.55	0.10	0.17	0.04	0.98	0.03	0.71	0.16	0.07	0.09	0.12	0.11	0.57	0.05	0.10	0.18	0.16	0.28	0.20	0.35	0.08	0.11	0.75	0.23	0.13	0.39	0.11	0.09	0.04
Genuine	0.17	0.26	0.71	0.53	0.20	0.55	0.07	0.25	0.08	0.32	0.14	0.24	0.11	0.28	0.25	0.46	0.03	0.97	0.07	0.29	0.57	0.26	0.26	0.30	0.08	0.30	0.49	0.21	0.24	0.20	0.34	0.20	0.22	0.55	0.07	0.23	0.30	0.11	0.60	0.45	0.58
Has_expertise	0.62	0.04	0.09	0.10	0.05	0.08	0.76	0.01	0.71	0.05	0.32	0.20	0.53	0.10	0.09	0.09	0.71	0.07	0.96	0.13	0.05	0.02	0.05	0.05	0.66	0.05	0.03	0.17	0.22	0.25	0.22	0.28	0.05	0.07	0.82	0.16	0.08	0.36	0.07	0.08	0.07
Has_integrity	0.18	0.28	0.35	0.34	0.28	0.21	0.15	0.24	0.20	0.32	0.27	0.22	0.17	0.57	0.47	0.27	0.16	0.29	0.13	1.00	0.42	0.24	0.28	0.37	0.14	0.54	0.26	0.17	0.51	0.21	0.43	0.24	0.27	0.33	0.13	0.28	0.33	0.21	0.35	0.41	0.46
Honest	0.17	0.27	0.59	0.61	0.29	0.46	0.05	0.33	0.09	0.32	0.16	0.20	0.11	0.42	0.46	0.43	0.07	0.57	0.05	0.42	1.00	0.30	0.24	0.36	0.04	0.47	0.47	0.13	0.34	0.22	0.35	0.18	0.33	0.54	0.05	0.25	0.40	0.11	0.50	0.52	0.76
Kind	0.08	0.71	0.24	0.22	0.61	0.21	0.07	0.86	0.03	0.72	0.08	0.22	0.11	0.37	0.42	0.14	0.09	0.26	0.02	0.24	0.30	0.99	0.67	0.39	0.08	0.46	0.33	0.12	0.24	0.28	0.24	0.13	0.74	0.41	0.05	0.26	0.35	0.05	0.24	0.24	0.27
Looks_out_for_othe	0.08	0.75	0.22	0.21	0.57	0.17	0.10	0.67	0.09	0.65	0.13	0.24	0.14	0.35	0.36	0.12	0.12	0.26	0.05	0.28	0.24	0.67	1.00	0.40	0.08	0.35	0.26	0.18	0.23	0.23	0.25	0.15	0.66	0.34	0.05	0.34	0.35	0.07	0.15	0.24	0.26
Loyal	0.12	0.41	0.32	0.29	0.30	0.16	0.12	0.41	0.13	0.40	0.34	0.38	0.13	0.39	0.35	0.21	0.11	0.30	0.05	0.37	0.36	0.39	0.40	0.99	0.10	0.39	0.24	0.34	0.28	0.41	0.39	0.35	0.43	0.37	0.09	0.46	0.47	0.28	0.22	0.43	0.35
Meticulous	0.55	0.03	0.07	0.11	0.11	0.13	0.63	0.04	0.60	0.08	0.34	0.25	0.57	0.14	0.08	0.13	0.57	0.08	0.66	0.14	0.04	0.08	0.08	0.10	0.96	0.07	0.08	0.24	0.25	0.26	0.18	0.35	0.08	0.05	0.64	0.14	0.11	0.40	0.14	0.09	0.08
Moral	0.09	0.39	0.29	0.37	0.39	0.29	0.04	0.42	0.04	0.45	0.10	0.23	0.10	0.73	0.55	0.18	0.05	0.30	0.05	0.54	0.47	0.46	0.35	0.39	0.07	1.00	0.30	0.12	0.54	0.25	0.35	0.17	0.46	0.41	0.04	0.26	0.37	0.09	0.29	0.43	0.47
Open	0.16	0.25	0.52	0.46	0.24	0.52	0.10	0.34	0.08	0.27	0.17	0.16	0.12	0.30	0.27	0.51	0.10	0.49	0.03	0.26	0.47	0.33	0.26	0.24	0.08	0.30	0.99	0.20	0.25	0.17	0.20	0.18	0.26	0.40	0.07	0.17	0.27	0.15	0.66	0.34	0.45
Predictable	0.25	0.12	0.23	0.21	0.11	0.16	0.22	0.09	0.18	0.13	0.51	0.39	0.24	0.14	0.14	0.24	0.18	0.21	0.17	0.17	0.13	0.12	0.18	0.34	0.24	0.12	0.20	0.90	0.20	0.45	0.27	0.37	0.12	0.10	0.22	0.40	0.28	0.38	0.27	0.28	0.16
Principled	0.21	0.24	0.28	0.26	0.26	0.23	0.21	0.23	0.23	0.28	0.18	0.22	0.23	0.62	0.40	0.27	0.16	0.24	0.22	0.51	0.34	0.24	0.23	0.28	0.25	0.54	0.25	0.20	1.00	0.22	0.39	0.32	0.27	0.24	0.20	0.20	0.26	0.25	0.25	0.34	0.39
Reliable	0.22	0.20	0.18	0.24	0.18	0.13	0.35	0.26	0.35	0.24	0.48	0.72	0.33	0.22	0.25	0.12	0.28	0.20	0.25	0.21	0.22	0.28	0.23	0.41	0.26	0.25	0.17	0.45	0.22	1.00	0.34	0.59	0.26	0.24	0.27	0.64	0.45	0.37	0.18	0.41	0.24
Respectable	0.21	0.20	0.35	0.32	0.20	0.22	0.26	0.27	0.27	0.30	0.33	0.38	0.24	0.40	0.37	0.27	0.20	0.34	0.22	0.43	0.35	0.24	0.25	0.39	0.18	0.35	0.20	0.27	0.39	0.34	0.97	0.34	0.24	0.34	0.20	0.35	0.33	0.27	0.29	0.41	0.36
Responsible	0.26	0.12	0.17	0.21	0.14	0.11	0.38	0.14	0.42	0.14	0.50	0.61	0.46	0.22	0.22	0.13	0.35	0.20	0.28	0.24	0.18	0.13	0.15	0.35	0.35	0.17	0.18	0.37	0.32	0.59	0.34	0.99	0.16	0.15	0.36	0.45	0.30	0.41	0.18	0.34	0.17
Shows_goodwill	0.08	0.70	0.20	0.22	0.65	0.20	0.04	0.74	0.05	0.65	0.10	0.21	0.05	0.38	0.40	0.15	0.08	0.22	0.05	0.27	0.33	0.74	0.66	0.43	0.08	0.46	0.26	0.12	0.27	0.26	0.24	0.16	0.98	0.37	0.04	0.35	0.41	0.09	0.16	0.24	0.30
Sincere	0.14	0.40	0.52	0.53	0.28	0.41	0.10	0.41	0.11	0.46	0.15	0.22	0.13	0.34	0.40	0.38	0.11	0.55	0.07	0.33	0.54	0.41	0.34	0.37	0.05	0.41	0.40	0.10	0.24	0.24	0.34	0.15	0.37	1.00	0.11	0.27	0.35	0.14	0.42	0.43	0.53
Skilled	0.64	0.07	0.10	0.11	0.08	0.09	0.75	0.03	0.68	0.04	0.36	0.23	0.59	0.10	0.13	0.10	0.75	0.07	0.82	0.13	0.05	0.05	0.05	0.09	0.64	0.04	0.07	0.22	0.20	0.27	0.20	0.36	0.04	0.11	0.96	0.16	0.07	0.37	0.10	0.07	0.04
Someone_you_can_	0.15	0.33	0.18	0.27	0.21	0.16	0.21	0.25	0.24	0.29	0.46	0.67	0.20	0.18	0.29	0.15	0.23	0.23	0.16	0.28	0.25	0.26	0.34	0.46	0.14	0.26	0.17	0.40	0.20	0.64	0.35	0.45	0.35	0.27	0.16	1.00	0.61	0.35	0.18	0.43	0.28
Someone_you_can_	0.12	0.37	0.28	0.46	0.26	0.23	0.12	0.33	0.18	0.33	0.27	0.48	0.08	0.32	0.34	0.24	0.13	0.30	0.08	0.33	0.40	0.35	0.35	0.47	0.11	0.37	0.27	0.28	0.26	0.45	0.33	0.30	0.41	0.35	0.07	0.61	0.99	0.23	0.30	0.49	0.40
Steadfast	0.37	0.08	0.15	0.14	0.13	0.15	0.41	0.09	0.37	0.07	0.48	0.42	0.49	0.13	0.14	0.22	0.39	0.11	0.36	0.21	0.11	0.05	0.07	0.28	0.40	0.09	0.15	0.38	0.25	0.37	0.27	0.41	0.09	0.14	0.37	0.35	0.23	0.96	0.15	0.24	0.12
Transparent	0.17	0.14	0.61	0.53	0.14	0.61	0.09	0.21	0.08	0.18	0.18	0.18	0.14	0.30	0.24	0.59	0.11	0.60	0.07	0.35	0.50	0.24	0.15	0.22	0.14	0.29	0.66	0.27	0.25	0.18	0.29	0.18	0.16	0.42	0.10	0.18	0.30	0.15	0.98	0.40	0.52
Trustworthy	0.17	0.22	0.41	0.49	0.21	0.28	0.14	0.25	0.18	0.30	0.33	0.48	0.18	0.40	0.38	0.28	0.09	0.45	0.08	0.41	0.52	0.24	0.24	0.43	0.09	0.43	0.34	0.28	0.34	0.41	0.41	0.34	0.24	0.43	0.07	0.43	0.49	0.24	0.40	1.00	0.61
Truthful	0.20	0.23	0.59	0.63	0.26	0.48	0.09	0.26	0.09	0.29	0.17	0.22	0.08	0.46	0.40	0.47	0.04	0.58	0.07	0.46	0.76	0.27	0.26	0.35	0.08	0.47	0.45	0.16	0.39	0.24	0.36	0.17	0.30	0.53	0.04	0.28	0.40	0.12	0.52	0.61	1.00

proportion of time placed in same bin out of times placed anywhere

Figure 5-E. Trust Items Sorting Expansion Study: The proportion of time a specific item was placed in the same bin as another item; the diagonal shows the proportion of time each item was placed into any one of the trust bins (i.e., less than 1.00 on the diagonal indicates that an item was sometimes classified into the "Other" bin). Items organized alphabetically on both axes.

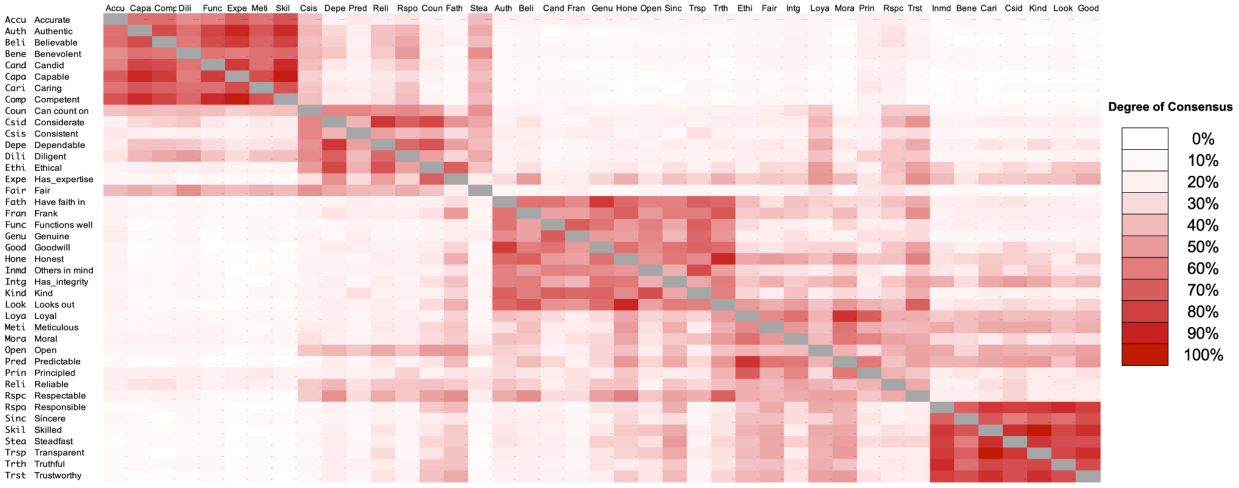
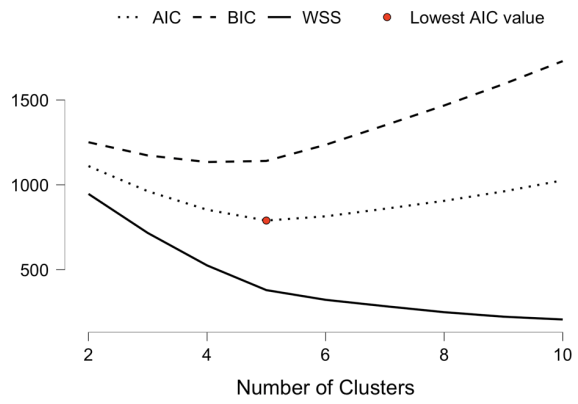


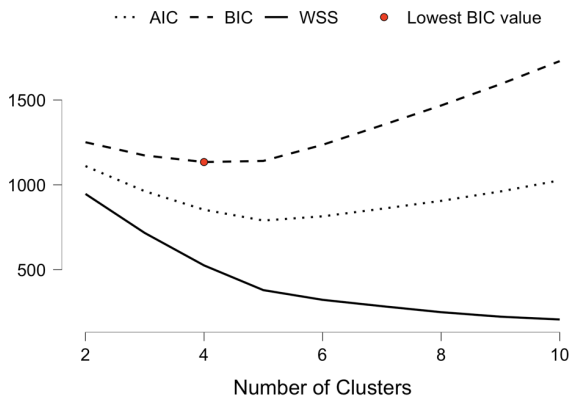
Figure 5-F. Trust Items Sorting Expansion Study: A heat map showing clusters of items based on high co-occurrence proportions. Co-occurrences of items with themselves on the diagonal are grayed out. Items are arranged on both axes based on the co-occurrence clusters. Rate of co-occurrence is shaded from white (0%) to red (100%).

I primarily present here the results from k-means cluster analysis, which offer a clear capture of the data as well as insight into the transition from a four-dimensional model to a five-dimensional model

(adding Benevolent as a new dimension). This data structure replicates across a variety of methods, including multidimensional scaling, hierarchical cluster analysis, network analysis, and principal components analysis; I also briefly present the results from a multidimensional scaling analysis to show the consistency of interpretation.

I begin with the results from k-means cluster analysis. Cluster analysis is an unsupervised machine learning approach to test for natural clusters of items while only specifying the predicted number of clusters. We were interested in distinguishing between the possibility of four clusters/dimensions (previous conception) and five clusters/dimensions (updated conception). This approach seeks to maximize intra-class similarity and minimize inter-class similarity, reflecting our previous methods and analyses to identify and isolate the component dimensions of trust. There are a few widely used algorithms with different advantages and disadvantages (Morissette & Chartier, 2013); we used the Hartigan-Wong algorithm (Hartigan & Wong, 1979), which optimizes within-cluster sum of squares, and we attempted cluster determination optimized according to both AIC and BIC as fit indices. BIC penalizes model complexity more than AIC, such that if they do not yield the same number of clusters then BIC should yield a smaller model than AIC; the two fit indices can be used together in model selection (Kuha, 2004), which we do here. When optimized according to AIC the cluster analysis yielded an optimal number of clusters of five, whereas when optimized according to BIC the cluster analysis yielded an optimal number of clusters of four. This optimization, together with corresponding elbow plots, can be seen in Figure 5-G. It is also worth noting that while the BIC value is lowest for four clusters, as can be seen in the elbow plots, there is very little difference compared to five clusters. This analysis pointed once again to a multi-dimensional interpretation of trust, with some question of whether four or five dimensions offers an optimal solution (explored in the Discussion below). The four/five dimension split stems from the relatedness of the Transparent items to the Ethical items (with Benevolent always separate), and this relatedness is captured to varying degrees across the variety of analysis methods. We display the solution with five dimensions in Figure 5-H (the four-cluster solution simply collapses across Ethical and Transparent).





K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	41	0.768	789.880	1141.160	0.420

Note. The model is optimized with respect to the *AIC* value.

K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
4	41	0.680	853.570	1134.600	0.390

Note. The model is optimized with respect to the *BIC* value.

Figure 5-G. Trust Items Sorting Expansion Study: k-means cluster analysis optimized to the AIC value with elbow plot on the left yielding 5 clusters (displayed in Figure 5-H) and k-means cluster analysis optimized to the BIC value with elbow plot on the right yielding 4 clusters.

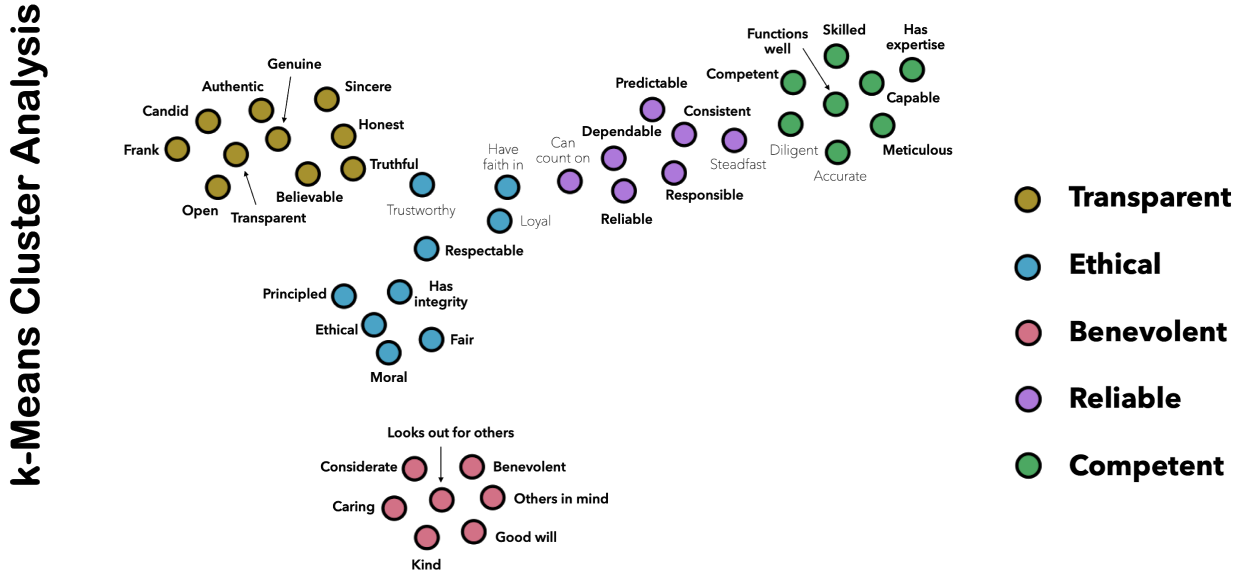


Figure 5-H. Trust Items Sorting Expansion Study: Visualization from the k-means cluster analysis. The five extracted clusters correspond to the updated five hypothesized dimensions of trust.

Regardless of the exact method of analysis, the item structure/dimensional analysis results in similar clusters. From multidimensional scaling, to hierarchical cluster analysis, to network analysis, to principal components analysis—the item structure is consistent given the underlying item relationships. To demonstrate the consistency across methods, I also present the visualization from a more traditional analysis approach, namely multidimensional scaling using an alternating least-squares algorithm (Figure 5-I). The groupings of items from multidimensional scaling occur in similar spatial positions as in the cluster analysis, which is apparent in the item distribution visualizations—this is because the underlying relationships between the items do not change. The distances that multidimensional scaling computes strongly suggest that Benevolent is distinct from both Ethical and Transparent, yielding five groupings.

Overall, across the various methods of analysis, we find evidence for multiple dimensions of trust while also showing compelling evidence from the visualizations of clusters supporting the two superordinate factors of trust—represented by the alignment of items along two axes/factors of **Performance Trust** and **Moral Trust**. In fact, this can be seen visually in both Figure 5-H from cluster analysis and Figure 5-I from multidimensional scaling, where items related to **Performance Trust** are aligned roughly on a horizontal axis and items related to **Moral Trust** are aligned roughly on a vertical axis. I discuss this—and the resulting dimensional interpretation—more in depth below.

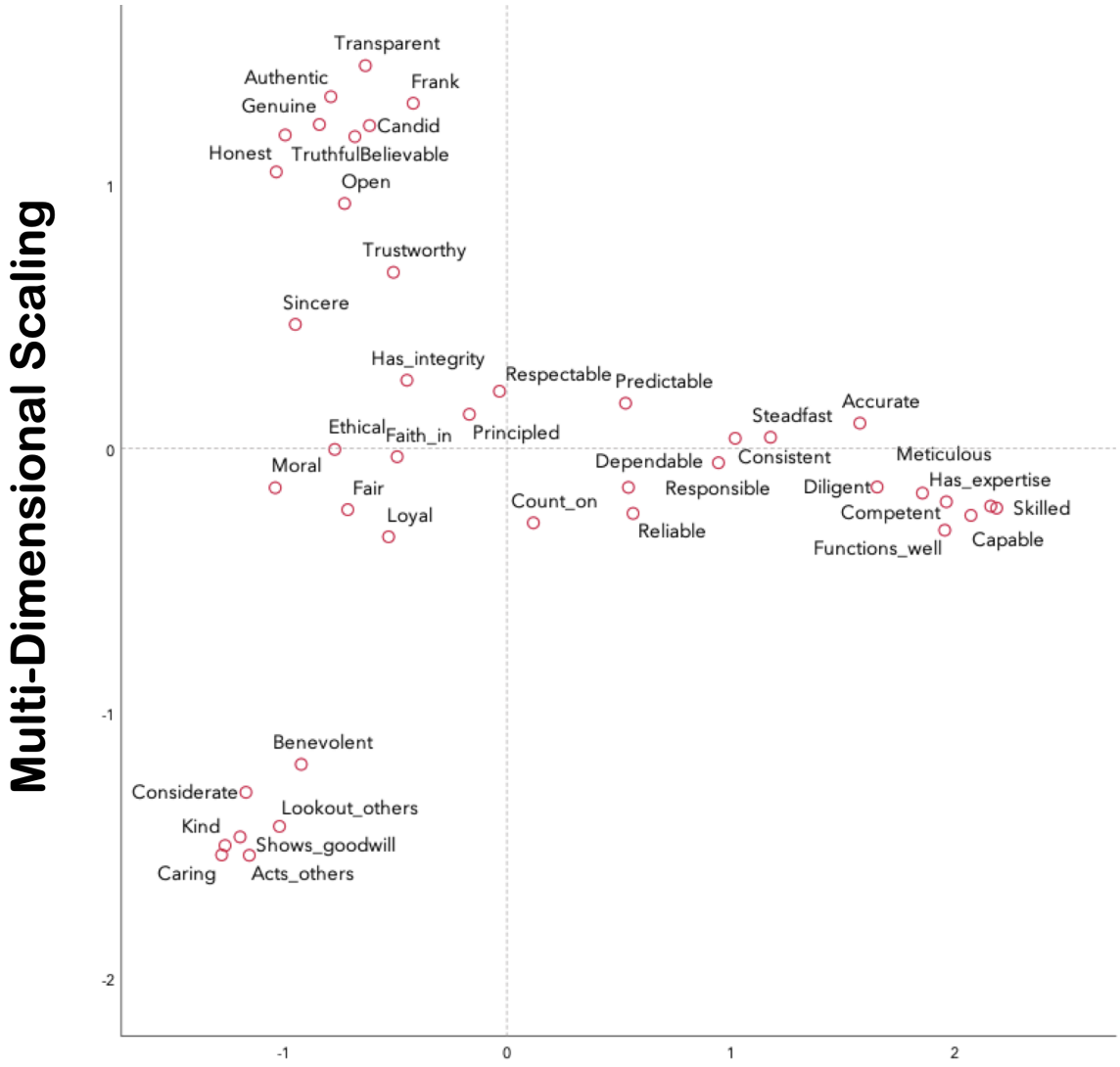


Figure 5-I. Trust Items Sorting Expansion Study: Multidimensional scaling using an alternating least-squares algorithm performed on the data.

5.3 Discussion

We set out to test whether a five-dimensional model better captured the conceptual, semantic space of trust. We had reason to believe that Benevolent, if assessed appropriately, might appear as a fifth trust dimension—and it did. We first saw this in the frequency distribution of participant-assigned labels, where terms related to the idea of benevolence appeared among the most frequent category labels (Figure 5-D). We then saw this in a higher-fidelity analysis conducted on the underlying co-occurrence structure, which consistently replicated across methods of analysis. This study used the most free-form of our

conceptual framework tasks (i.e., no experimenter-assigned labels/categories but rather participant-assigned categories and corresponding labels), and yet it yielded the same patterns as the previous studies. This contributed strong evidence for our updated working model of trust.

The data revealed four to five dimensions across methods of analysis. In four-dimensional solutions, Benevolent was always separate as its own dimension, whereas Ethical and Transparent were collapsed. In five-dimensional solutions, Ethical, Transparent, and Benevolent were all separate. This is an important observation, as the previous studies showed consistent evidence for Ethical and Transparent as distinct dimensions. This therefore points to a five-dimensional solution, given that the new Benevolent dimension was shown here to be consistently separable across analyses in this study while the Ethical and Transparent dimensions have been shown to be separable in the previous studies. An expanded version of the trust change study, discussed in the next chapter, allowed us to further validate this new five-dimensional solution and the inclusion of the Benevolent dimension.

An interpretation of trust with two superordinate factors (**Performance Trust**, **Moral Trust**) has repeatedly appeared in the empirical studies presented in this dissertation and reflects the conceptual grounding in the contributing literatures summarized in the Introduction (Chapter 1). We can clearly see this two-factor capture of trust in the resulting data structure from the sorting expansion task, where items distributed along two distinct axes in both cluster analysis in Figure 5-H and in multidimensional scaling in Figure 5-I. In both visualizations, items related to **Performance Trust** are aligned roughly on a horizontal axis and items related to **Moral Trust** are aligned roughly on a vertical axis. The general item *trustworthy* appears as a focal intersection point between these two axes (discussed more below), with Reliable and Competent items clustered together and then Ethical, Transparent, and Benevolent items clustered together. This distribution indicates that the items that constitute Reliable and Competent are more related to one another (part of the **Performance Trust** factor) than they are to the items that constitute Ethical, Transparent, and Benevolent (part of the **Moral Trust** factor)—and vice versa. In fact, the distribution of items along these axes forms a rotated "T" shape—a fitting observation given that the items all fall under the umbrella of (T)rust. Overall, the underlying item structure and resulting groupings

of dimensions—which are clearly and consistently presented in figure visualizations—illustrate evidence for a summary of the dimensions using two axes/factors.

We transformed Figure 5-H to Figure 5-J, which shows a final selected set of 20 items split across five dimensions for the MDMT v2 (described below) excerpted from the k-means cluster analysis. The item *trustworthy* deserves special attention. The item *trustworthy* appears at the intersection between the axes/factors in the item distribution visualization in Figure 5-J, and is perhaps best interpreted separately from any one factor or dimension. In thinking about trust, it intuitively makes sense that the namesake item should potentially offer a more holistic capture of the overall concept of trust. For these reasons (especially given corresponding item structure here and in previous studies), *trustworthy* does not seem to fit under a single specific trust dimension—but instead can potentially serve as a parsimonious, single-item capture of general trust. We therefore do not include *trustworthy* in any single dimension of the MDMT, instead offering it as a candidate single-item measure of trust.

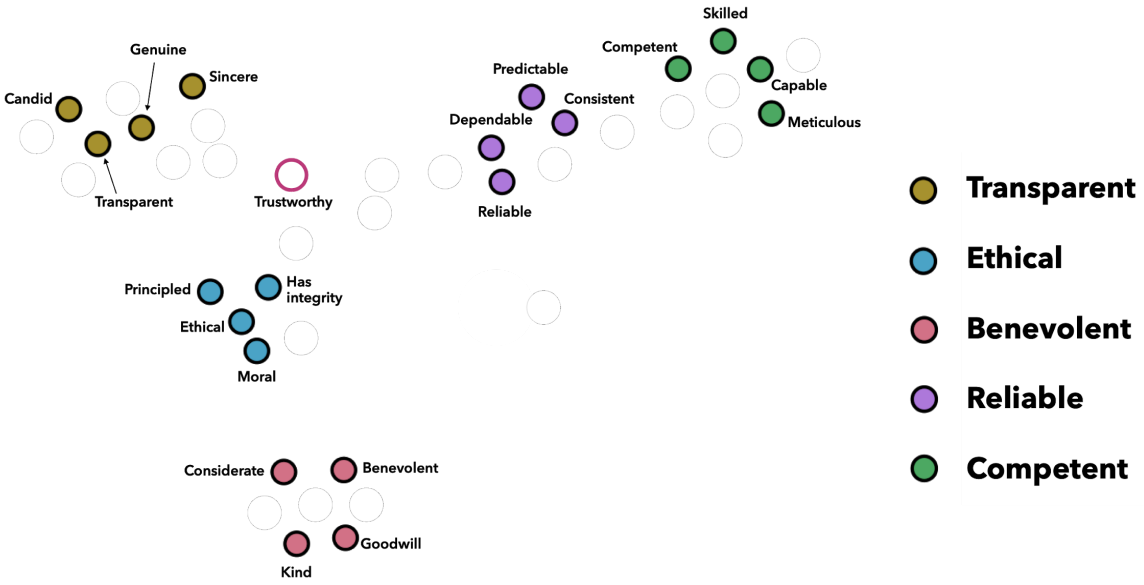


Figure 5-J. Trust Items Sorting Expansion Study: Visualization of the final selected set of 20 items split across five dimensions for the MDMT v2 (with the parsimonious single-item *trustworthy* also displayed) excerpted from the k-means cluster analysis.

One of the features of the sorting expansion task instructions involved telling participants "You can create as many groups as you think are necessary, but please create at least 2." The previous studies had shown evidence for multiple dimensions of trust; one of the goals of the present study was therefore to test how many groupings of items people form, and in particular whether participants would form a separate grouping related to the idea of benevolence using newly added items. We wanted to avoid giving participants the impression in the instructions that the task was asking them to sort items into one large grouping of "items related to trust" and sort remaining items they considered less relevant into the "Other" box. We therefore requested that participants form at least two groupings of items. However, future work could consider an even more free-form version of this task, without any restrictions on the number of item groupings. While the experimental design here enabled the calculation of distances between items from the aggregate data across ratings from all participants (i.e., participant consensus over item co-locations), one could instead use individual-person data. For example, researchers have used a multi-arrangement method where participants drag and drop candidate items from an outside item bank (which the authors refer to as the "seating") into the focal arrangement area (which the authors refer to as the "arena") in a series of trials (Kriegeskorte & Mur, 2012). Over multiple trials, researchers can compute distances between items as reflecting the person's underlying representation of item relationships as well as item clusters. This method aims to combine the strengths of using pairwise judgments (allowing people to indicate item relatedness information for two items at a time) and free sorting (allowing people to cluster items within the same cluster or in different clusters). Multi-arrangement can theoretically capture information along both these lines simultaneously. The method also aims to avoid the weaknesses of pairwise judgments (taking far too much time with larger item sets) and free sorting (typically relying on one trial per person and cross-person aggregation). The researchers propose an analysis method using inverse multidimensional scaling across these trials to gather the individual-person dimensional structure of item representations. A similar method could be implemented with trust items, though it should be noted that the arrangement task still relies on a two-dimensional arena space and requires some experimenter decisions regarding item sampling for successive trials (the researchers

acknowledge both weaknesses; Kriegeskorte & Mur, 2012). Overall, the prediction using this additional experimental method would be that the first trial (with all relevant trust items) would likely yield a similar dimensional structure as we found using the experimental task presented in this chapter, but data from the subsequent trials may offer more nuanced information about individual differences in the representation of items within dimensions. Although much of the focus of this dissertation is at the level of identifying trust factors and dimensions, additional work such as this could help further probe the interrelatedness of individual trust items.

After completing data analysis on the sorting expansion task, we reviewed and refined the MDMT. Based on the new insights obtained from the study, we updated from MDMT v1 to MDMT v2 making the following changes:

- "**Capacity Trust**" factor renamed to "**Performance Trust**"
- "Capable" subscale renamed to "Competent"
- "Sincere" subscale renamed to "Transparent"; subscale item "*authentic*" removed and "*transparent*" added
- "Benevolent" subscale added with constituent items *benevolent*, *kind*, *considerate*, *has goodwill*
- "Reliable" subscale item "*someone you can count on*" removed and "*dependable*" added
- "Ethical" subscale item "*respectable*" removed and "*moral*" added

As noted earlier, we ultimately renamed the **Capacity Trust** factor to **Performance Trust** in order to reflect that the factor contains both items related to capacity/capability as well as items related to reliability. In addition, we changed the dimension label from "Sincere" to "Transparent" as the word Transparent seems more applicable to potential robot agents, while still maintaining applicability to human and other agents (the opposite does not hold as easily for the applicability of "Sincere" to robot agents). A total of 20 items (four in each of the five dimensions) constitutes the revised MDMT. We made the second version of the MDMT, called MDMT v2, publicly available for researcher use dated September 1, 2020 (Appendix B; Ullman & Malle, 2020). The model structure of the MDMT v2 with trust factors, trust dimensions, and trust items is displayed in Figure 5-K.

MDMT v2					
Trust Factors	**Performance Trust**		**Moral Trust**		
Trust Dimensions	Reliable	Competent	Ethical	Transparent	Benevolent
trust items	reliable	competent	ethical	transparent	benevolent
	predictable	skilled	principled	genuine	kind
	dependable	capable	moral	sincere	considerate
	consistent	meticulous	has integrity	candid	has goodwill

Figure 5-K. MDMT v2: Model structure of the MDMT v2, including trust factors, trust dimensions, and trust items.

Chapter 6: Trust Change Expansion Study

The original trust change study (Chapter 4) demonstrated that the experimental approach using vignettes was a viable way to examine and measure changes in trust via the MDMT—at least for some dimensions. The results from that study showed that the related dimensions in the MDMT v1 within **Performance Trust** (i.e., Reliable and Capable) and **Moral Trust** (i.e., Ethical and Sincere) moved in tandem, showing strong support for the two superordinate factors as different types of trust. However, only one dimension within each factor—Reliable in **Performance Trust** and Ethical in **Moral Trust**—clearly showed the expected patterns of changes, whereas Capable and Sincere did not. One question was whether the specific vignettes designed to capture changes in trust for those two dimensions did not adequately probe the intended dimensions, or whether the results pointed to the possibility that these two dimensions were not sufficiently separable from their counterparts. We thus worked to optimize the vignettes to better isolate changes in an individual dimension of trust without prompting changes in the other dimensions.

A second question was whether the changes to the MDMT would affect the measurement capacity of the tool. We had recently updated from MDMT v1 to MDMT v2 based on the sorting expansion study (Chapter 5), which included Benevolent as a separate dimension and refined the dimension item sets. It was thus possible that this updated model and measure would be able to better capture differences between dimensions (especially in the **Moral Trust** factor with the addition of Benevolent) that the previous version could not. In general, it is plausible that dimensions under the same superordinate factor (e.g., Reliable and Competent, which combine to form the **Performance Trust** factor) are more difficult to differentiate than dimensions from different factors (e.g., Reliable against Ethical). However, we sought to rigorously test whether we could isolate each dimension by itself, as would seem theoretically possible given the semantic structure that had been uncovered.

A final question was whether the MDMT could assess trust in humans in addition to trust in robots. We originally developed the MDMT to model and measure trust agnostic to the type of agent being evaluated, but with the understanding that variation might stem from agent type. We were thus

interested in how the MDMT would fare in evaluations of human agents in vignettes, as well as whether findings (or the lack thereof) from the original trust change study (Chapter 4) could in part be explained by general expectations people have for trust in robots compared to humans. For example, it was possible that people did not find changes in a certain dimension of trust salient for robot agents (such as for Sincere/Transparent), but that they would for human agents.

Our goals in this expansion effort were thus threefold: (1) Use updated vignettes to better isolate and probe the theorized dimensions of trust; (2) Test the revised model and measure in the updated MDMT v2; (3) Compare trust in robot agents with trust in human agents using the MDMT v2. We first worked to optimize the set of vignettes to better examine changes in trust along the theorized dimensions; I briefly describe this effort here and describe it in depth in Appendix D.

Appendix D details the trust change expansion baseline, which was a research effort that consisted of testing a revised set of vignettes we created for the actual trust change expansion study. We refined the experimental vignettes in light of our updated theoretical conception of trust—integrating insights acquired from further literature review, subsequent experimental work, and discussion with trust and robotics researchers since the original trust change study (insights which are reflected in the MDMT v2). We tested a revised set of vignettes, which are listed in Appendix E, as part of the baseline research effort, which is detailed in Appendix D. We then made some modifications to the vignettes based on the baseline research effort; the resulting final updated vignettes, which we used in the actual trust change expansion study detailed in this chapter, are listed in Appendix F.

With this new set of vignettes in hand, we set out to test and validate the updated MDMT v2 using the experimental paradigm introduced in Chapter 4. The trust change expansion study described here offers the first comprehensive test of the MDMT v2.

6.1 Method

The trust change expansion study tested the new MDMT v2 (Appendix B) using the new vignettes (Appendix F); the experimental design was otherwise nearly identical in structure to the original trust change study, with the exception of a couple necessary differences described here.

The expansion task included human agents in addition to robot agents. The human agent vignettes were identical to the robot agent vignettes, except that we replaced the word "robot" with "person" throughout the text. To avoid any confusion for participants when we talked about a "person" in the new human agent conditions, we changed the task instructions for all conditions to say "We are interested in how participants interpret various events..." instead of "people" as in the original task.

In addition, we added the phrase "The [robot / person] completed an initial training on [relevant task for the given scenario]" to the first part of each vignette. This phrase was added to instantiate the idea that the described task was appropriate for the agent in the vignette.

Other than the necessary word changes, and the new phrase inclusion across vignettes, the rest of the focal task method with the vignettes paralleled the original task. We did however update our task comprehension check and added additional questions after the focal task, described below.

We pre-registered the trust change expansion study—including the experimental design, analyses, and sample size—with the Open Science Framework (OSF) (Malle & Ullman, 2020).

6.1.1 Procedure

The study began with an updated comprehension check, in part to also check for any bots. Participants were told, "Recently, there has been a threat of "bots" pretending to be research participants. Please help us recognize you as a committed participant by answering the following questions." Participants were then given the following two items we had generated: "What is the bottom of a tree called?" and "Please do not type anything in the text box below." For the first question, we expected participants to respond with "stump," "trunk," "root," "base," or some variant thereof. Responses such as "bots" or "leaves" prompted data exclusion. For the second question, any text entry into the box prompted data exclusion.

Participants received the following task instructions, which were nearly identical to those in the original study:

> "We are interested in how participants interpret various events involving [robots / people]. We will give you a very simple event and ask you to rate the described [robot /

person] on a set of different adjectives. Some of these adjectives will seem appropriate for the [robot / person] in the specific situation, while others may not."

Participants were then randomly assigned to their specific between-subjects condition. As in the original trust change study, each participant read a short vignette in two parts. Part one described the behavior of a robot or person occupying a specific role in a particular context (e.g., transportation security officer in an airport) and participants evaluated the agent on the updated 20 items of the MDMT v2 (Appendix B). Then participants were told "One month later you see the same [robot / person]." In part two, the agent's behavior changed (with implications for trustworthiness in a particular dimension) and participants again completed all 20 items of the MDMT v2.

After the task, participants were asked two questions about knowledge and experience with robots. Participants were asked, "Please rate how knowledgeable you are of robots and/or the robotics domain?" (0-7 rating scale from "Not at all knowledgeable" to "Very knowledgeable") and "Please rate your level of experience (i.e., having worked with or come into contact with robots) with robots" (0-7 rating scale from "No experience" to "Very experienced").

Participants were asked demographic questions (self-reported age, gender, and race/ethnicity) at the conclusion of the task.

6.1.2 Design

The updated vignettes consisted of 80 conditions in a 4 (*Scenario*: airport security, hospital cleaning, home companion, school tutor) x 5 (*Targeted Trust Dimension*: Reliable, Competent, Ethical, Transparent, Benevolent) x 2 (*Agent*: robot, human) x 2 (*Change Direction*: up, down) between-subjects design. *Scenario* refers to vignettes describing an agent's role-specific behavior in a relevant context (airport security, hospital cleaning, home companion, school tutor). *Targeted Trust Dimension* refers to the dimension of trust (Reliable, Competent, Ethical, Transparent, Benevolent) along which the agent's behavior changes. *Agent* refers to whether the agent described in the vignette is a robot or person (robot, human). *Change Direction* refers to whether the agent's behavior changes in ways intended to increase or decrease evidence (up, down) for the relevant dimension of trust. All vignettes can be seen in Appendix F.

Although we created 80 between-subjects conditions for participants, we were not interested in *Scenario* as a focal condition and did not plan to test any hypotheses about the specific scenarios (same as in the original task); rather, by pretesting vignettes in the baseline research effort (Appendix D), we had worked to create a candidate set of scenarios (Appendix F) that should ideally show any effects in general—and not specific to any one scenario. Although understanding the nuances of particular scenarios (and thus individual agent roles and contexts) is potentially valuable, this was not the focus of the study and therefore did not factor into sample size calculations. The design we set out to test was thus a 5 (*Targeted Trust Dimension*) x 2 (*Agent*) x 2 (*Change Direction*) design with 20 focal analysis conditions.

We used the original trust change study (Chapter 4) to guide sample size (i.e., approximately 50 participants in each focal condition). We therefore planned to recruit 1120 participants. Our target number of participants was (a) 1000 (50 in each of the 5 x 2 x 2 conditions), with an additional (b) 100 to oversample by 10% to account for projected exclusions, as well as an additional (c) 20 to make more likely an equal number of participants in all conditions (across each of the four scenarios). The resulting number of projected participants per actual condition in each of the 80 between-subjects conditions was 14 participants, such that the projected total number of participants in each of the 20 focal analysis conditions was 56 participants.

6.1.3 Participants

A total of 1132 adults, recruited via Prolific, participated in the approximately 5-minute online study and received compensation of $0.60. As declared in the pre-registration, we excluded low-quality responses from 147 participants (who met one or more of the following criteria: failed the comprehension check, had duplicate IP addresses, had repeat Prolific IDs, had a rating range of 0 or 1 across all MDMT items not including "Does Not Fit"), for a final sample of n = 985. Participants in the final sample had a mean age of 35.49 years (SD = 12.22); they self-reported gender as 491 "Male," 477 "Female," and 16 "Other" with 1 not reporting; and they self-reported race/ethnicity as 754 "Non-Hispanic White or Euro-American," 74 "Asian or Asian American," 61 "Black, Afro-Caribbean, or African American," 49 "Latino or Hispanic American," 36 "Multi-racial," 3 "Native American or Alaskan Native," and 8

"Other." People who had participated in the baseline study (Appendix D) were not permitted to participate in this study.

Participants had a mean robot knowledge score of $M = 2.78$ ($SD = 1.85$), and a mean robot experience score of $M = 2.01$ ($SD = 1.87$). These two items can be combined into an overall robot familiarity score, for which participants had a mean of $M = 2.39$ ($SD = 1.74$)—falling just below the middle of the scales for robot knowledge and experience, and verifying that the sample was not overly familiar with robots.

In addition, mobile devices were restricted from the study via Prolific to ensure clear presentation of the task; some participants began the Qualtrics survey on a mobile device but were informed that they had to return to the survey on a computer to complete the study.

6.2 Results

As with the results from the original trust change study, we calculated subscale scores for each of the now five trust dimensions (Reliable, Competent, Ethical, Transparent, Benevolent) for both the MDMT pre-test and the MDMT post-test by averaging the specified items for each dimension at each time point (not including items marked by a participant as "Does Not Fit"). We then also used the MDMT pre-test and MDMT post-test item ratings to calculate difference scores for each available item's pre and post ratings (not including "Does Not Fit"); we used these to calculate subscale difference scores. We collapsed across scenarios, as *Scenario* was not a focal factor. There were no missing values across both the MDMT pre-test and the MDMT post-test—that is, for each item, participants either gave a numerical rating on the 0-7 scale or marked that the item "Does Not Fit."

Once again, I generated three sets of figures for the study from the MDMT pre-test (Figure 6-A), the MDMT post-test (Figure 6-B), and the MDMT difference scores (Figure 6-C). Each figure has four distinct graphs split out by the *Agent* factor for the human agent and the robot agent, as well as by the *Change Direction* factor for the up level (increasing evidence for trust on a specified dimension) and the down level (decreasing evidence for trust on a specified dimension). The scaling for each graph is identical to the corresponding up/down graphs in the original trust change study, and the color-coding and

line patterns for each dimension have been kept constant (with the addition of Benevolent). I often evaluate the robot results in light of the human results as a possible benchmark, as well as compare the robot results from this replication to the original study results.

We started by looking at the results from the MDMT pre-test (Figure 6-A). In theory, the graphs for the up and down levels in the MDMT pre-test should be nearly identical within each level of *Agent* (as participants had only read the first part of the vignette without the behavior change at this point), but not necessarily across the two agents (for which people may have different baseline expectations). As in the original study, there is some random variation between the levels of the *Change Direction* factor, such that the up and down graphs differ slightly within each agent. In the original trust change study (Chapter 4), we noted a baseline Moral-Performance trust gap for the robot agent. In the trust change expansion study, the baseline trust pattern for the robot agent (graphs on the right) replicates the baseline Moral-Performance trust gap, such that baseline ratings for **Moral Trust** dimensions in the robot were lower than for **Performance Trust** dimensions. However, there is no corresponding gap in the baseline trust for the human agent (graphs on the left) between the **Moral Trust** dimensions and the **Performance Trust** dimensions. We confirmed this baseline Moral-Performance trust gap for the robot agent with statistical testing reported later in this section (Figure 6-J). Finally, we also noted that the baseline **Performance Trust** ratings for both agents were quite comparable.

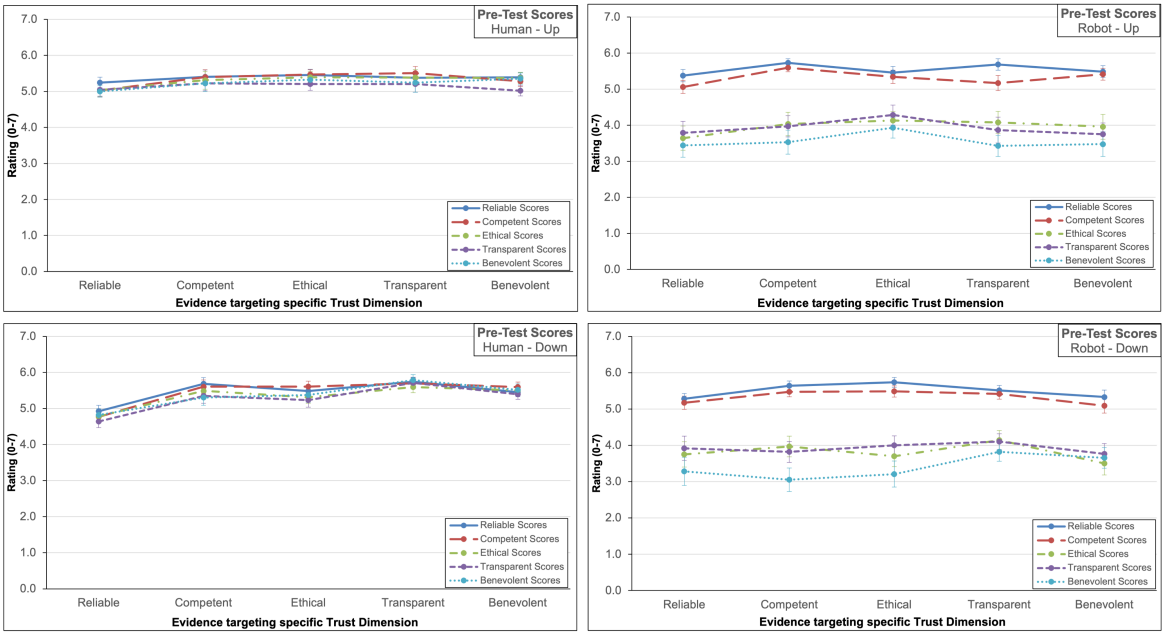


Figure 6-A. Trust Change Expansion Study: Each graph shows MDMT pre-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

Next, we examined the results from the MDMT post-test (Figure 6-B). As in the original trust change study, we compared each graph to the corresponding pre-test graph (Figure 6-A) to identify overall patterns of changes in ratings along the different dimensions. Once again, ratings generally increased for the up level of *Change Direction*, whereas ratings generally decreased for the down level of *Change Direction*—indicating that people responded to salient changes in an agent's behavior with corresponding changes in trust that were captured by the MDMT. The changes between the MDMT pre-test and the MDMT post-test are depicted in the difference scores graphs in Figure 6-C, discussed next.

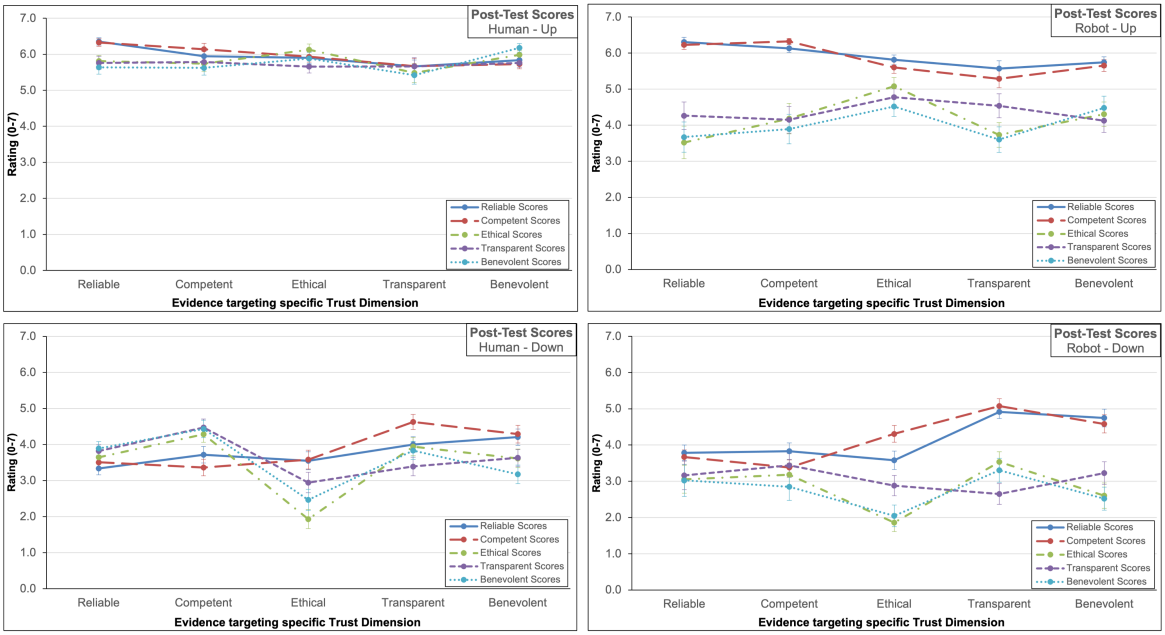


Figure 6-B. Trust Change Expansion Study: Each graph shows MDMT post-test subscale scores (legend) on a 0-7 scale (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

In the original trust change study we posited that if the targeted dimension evidence had the intended effect (prompting change only along a single dimension at a time), we would see a change in each subscale score (e.g., Ethical) for only the corresponding *Targeted Trust Dimension* (Ethical). In our updated conception, and as stated in our pre-registration with the Open Science Framework (OSF) (Malle & Ullman, 2020), we posited: (1) the changes for a particular subscale score (e.g., Ethical) would be stronger than those for other subscale scores; and (2) they would be stronger in response to the dimension-relevant targeted evidence than in response to other targeted evidence. We planned to test these hypotheses using Helmert contrasts. For example, to test the hypotheses for Benevolent we would use Helmert contrasts to examine whether: (1) the difference score for the Benevolent subscale was largest when Benevolent was the *Targeted Trust Dimension*, compared to the remaining four subscale scores (Reliable score, Competent score, Ethical score, Transparent score); and (2) the difference score for the Benevolent subscale was largest when Benevolent was the *Targeted Trust Dimension*, compared to the

scores for Benevolent when the other four dimensions were the *Targeted Trust Dimension* (Reliable dimension, Competent dimension, Ethical dimension, Transparent dimension).

As seen in the graphs in Figure 6-C, the *Targeted Trust Dimension* prompted corresponding changes in trust subscale scores for almost every condition. I first describe these results using the difference scores graphs before further validating these findings using statistical testing reported below. I use Figure 6-C as a helpful reference point that grounds the statistical testing conducted. As in the original study, we noted from the difference scores graphs (Figure 6-C) that the dimensions within **Performance Trust** (Reliable, Competent) and within **Moral Trust** (Ethical, Transparent, Benevolent) often moved in tandem. This held true for the human agent in addition to the robot agent, despite the fact that the baseline ratings for the human agent did not show an initial difference between the two factors as they did for the robot agent (i.e., even though the baseline Moral-Performance trust gap only occurred for the robot agent and not the human agent).

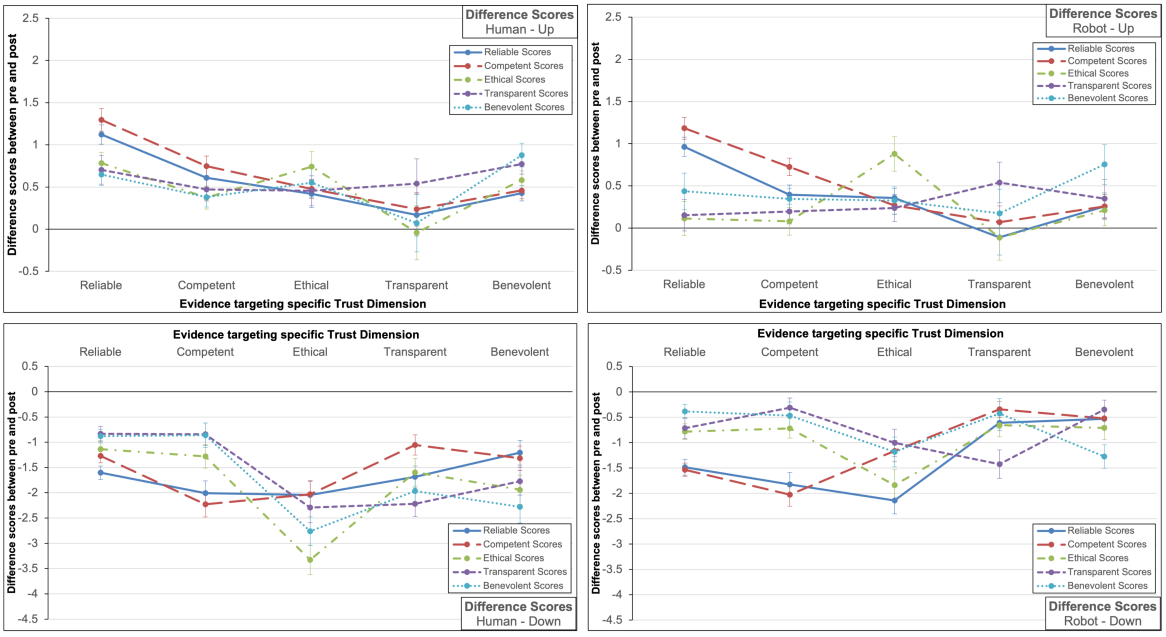


Figure 6-C. Trust Change Expansion Study: Each graph shows MDMT difference subscale scores (legend) on appropriate scaling (y-axis) for each level of the *Targeted Trust Dimension* factor (x-axis). The graphs are split by the *Agent* factor (left graphs show human agent, right graphs show robot agent) and by the *Change Direction* factor (top graphs show the up level for evidence increasing trust, bottom graphs show the down level for evidence decreasing trust). Error bars show *SE*.

For the first hypothesis test, we examined whether a difference score was largest for a subscale (identified in the legend of graphs in Figure 6-C) when it was the *Targeted Trust Dimension* (on the x-axis of graphs in Figure 6-C) compared to the other subscale scores (i.e., whether the relevant subscale shows the largest difference score for its corresponding trust dimension on the x-axis compared to the other four subscale scores in the same column). For example, does the Benevolent subscale score show the largest magnitude of difference for the Benevolent *Targeted Trust Dimension* out of all the subscale scores in the column—which it does in all four graphs, especially the two robot graphs (i.e., the Benevolent subscale score is much further from the x-axis than the other subscale scores in the same column). Figure 6-C shows this to be the case for all conditions except for the Reliable scores in the human and robot up conditions as well as in the robot down condition, and for the Ethical score in the robot down condition. In the Ethical robot down condition, the trust loss prompted low Ethical ratings but these were slightly overshadowed by an even greater loss in Reliable ratings of the agent. I address this pattern in the Discussion section, along with the finding that the changes on the Reliable subscale scores were overshadowed in three out of four cases by larger changes in the Competent subscale scores.

For the first hypothesis test, the best evidence for the sensitivity of the updated MDMT v2 to relevant dimensional evidence comes from an examination of the *t* values of the Helmert contrasts for the human and robot agents. Figure 6-D displays *t* values from Helmert contrasts on the difference scores shown on the diagonal comparing the focal subscale score to the other four subscale scores for the corresponding focal *Targeted Trust Dimension*. Each column shows the subscale scores that resulted from evidence probing the dimension specified at the top of the column. If the evidence for each *Targeted Trust Dimension* worked as intended, prompting the greatest change in the corresponding subscale score (addressing the first test described above), we should see large *t* values—and we see just that (*t* values ranging from 2.86 to 8.02). In addition, the effect sizes for each contrast can be seen in Figure 6-E. The effect sizes have very large Cohen's *d* values (*d* values ranging from 0.63 to 1.34) that indicate the subscales of the MDMT were highly sensitive to changes in their corresponding focal *Targeted Trust Dimension*—a major validation of the dimensional specificity and sensitivity of the MDMT v2.

HUMAN AGENT *t* values		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Subscale Scores	Reliable	4.16				
	Competent		5.83			
	Ethical			8.02		
	Transparent				6.02	
	Benevolent					5.52

ROBOT AGENT *t* values		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Subscale Scores	Reliable	2.86				
	Competent		5.37			
	Ethical			3.10		
	Transparent				5.35	
	Benevolent					4.42

Figure 6-D. Trust Change Expansion Study: *t* values from Helmert contrasts on the difference scores shown in the diagonal cells, comparing the focal subscale score to the other four subscale scores for the corresponding focal *Targeted Trust Dimension*. Each column shows the subscale scores that resulted from evidence targeting the dimension specified at the top of the column. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

HUMAN AGENT Cohen's *d*		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Subscale Scores	Reliable	0.95				
	Competent		1.32			
	Ethical			1.34		
	Transparent				1.22	
	Benevolent					1.00

ROBOT AGENT Cohen's *d*		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Subscale Scores	Reliable	0.97				
	Competent		1.16			
	Ethical			0.63		
	Transparent				1.25	
	Benevolent					1.20

Figure 6-E. Trust Change Expansion Study: Cohen's *d* values from Helmert contrasts on the difference scores shown in the diagonal cells, comparing the focal subscale score to the other four subscale scores for the corresponding focal *Targeted Trust Dimension*. Each column shows the subscale scores that resulted from evidence targeting the dimension specified at the top of the column. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

For the second hypothesis test, we examined whether a difference score was largest for a subscale (identified in the legend of graphs in Figure 6-C) when it was the *Targeted Trust Dimension* (on the x-axis of graphs in Figure 6-C) compared to the scores for the same subscale when other dimensions were the *Targeted Trust Dimension* (i.e., whether the relevant subscale shows the largest difference score when it is the corresponding trust dimension on the x-axis compared to its scores for the other dimensions along the x-axis). For example, does the Benevolent subscale score show the largest magnitude of difference for its own *Targeted Trust Dimension* than when the other dimensions are the *Targeted Trust Dimension*—which it does in all graphs except the human down graph where the Benevolent subscale responds slightly more to the Ethical *Targeted Trust Dimension* than to its own dimension (i.e., the Benevolent subscale score is furthest from the x-axis for the Ethical dimension instead of the Benevolent dimension). The results for the second hypothesis test were more nuanced than for the first hypothesis test, showing more variation that can be seen in the graphs of Figure 6-C. However, a statistical comparison in Figure 6-F offers evidence that validates this hypothesis test as well. We conducted *F* tests using Helmert contrasts on the difference scores comparing the magnitude of difference for the focal subscale score for the corresponding focal *Targeted Trust Dimension* against the scores for the same subscale when other dimensions were the *Targeted Trust Dimension*. As shown in Figure 6-F, we found significant effects for all five subscales across both agents (*F* values ranging from 2.15 to 33.55)—with the exception of Reliable for the human agent—with small to moderate effect sizes (*d* values ranging from 0.14 to 0.57).

This study also offered the first test of the inclusion of the Benevolent dimension. Benevolent not only showed similar patterns as the other **Moral Trust** dimensions of Ethical and Transparent—as displayed in Figure 6-C—but also showed dimensional specificity beyond the related dimensions—as indicated by the contrast tests in Figure 6-D, Figure 6-E, and Figure 6-F.

HUMAN AGENT *F* tests		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Compared to other evidence	*F*	2.15	5.80	33.55	4.09	9.05
	p	.14	.02	< .001	.04	< .01
	d	0.14	0.24	0.57	0.20	0.29

ROBOT AGENT *F* tests		Evidence targeting specific Trust Dimension				
		Reliable	Competent	Ethical	Transparent	Benevolent
Compared to other evidence	*F*	4.46	14.72	24.21	5.32	4.19
	p	.04	< .001	< .001	.02	.04
	d	0.24	0.44	0.56	0.26	0.23

Figure 6-F. Trust Change Expansion Study: *F* tests using Helmert contrasts on the difference scores comparing the magnitude of difference for the focal subscale score for the corresponding focal *Targeted Trust Dimension* against the scores for the same subscale when other dimensions were the *Targeted Trust Dimension*. Conditions are split by the *Targeted Trust Dimension* (Reliable, Competent, Ethical, Transparent, Benevolent) for the human agent on top and the robot agent on the bottom.

We also calculated the number of "Does Not Fit" selections to compare the frequencies for the MDMT pre-test and the MDMT post-test, as well as to compare the frequencies for the robot and human agents. We report the average number of items (out of four per dimension) that received this selection, broken out in the pre-test and post-test for both agents, in Figure 6-G. The most pronounced observation is that participants selected "Does Not Fit" far more often for the robot on the **Moral Trust** dimensions than for the human. Participants also selected "Does Not Fit" for both agents more on the **Moral Trust** dimensions than on the **Performance Trust** dimensions. Detailed analyses follow next.

Agent	Pre-Test or Post-Test	Average number of items out of 4 per dimension that received a "Does Not Fit" evaluation				
		Reliable	Competent	Ethical	Transparent	Benevolent
Human	Pre	0.17	0.10	0.29	0.57	0.40
	Post	0.11	0.07	0.28	0.45	0.34
Robot	Pre	0.08	0.15	1.31	1.26	1.37
	Post	0.05	0.16	1.39	1.30	1.44

Figure 6-G. Trust Change Expansion Study: Average number of items (out of four per dimension) that received a "Does Not Fit" selection on the MDMT pre-test and the MDMT post-test for the human and robot agents.

We tested the differences in "Does Not Fit" selections between the **Performance Trust** factor and the **Moral Trust** factor for the robot and human agents on the MDMT pre-test and the MDMT post-test (Figure 6-H). We first examined whether there was any main effect of agent; we found that participants made a greater number of these selections for the robot agent ($M = 0.73$, $SD = 0.81$) than for the human agent ($M = 0.25$, $SD = 0.48$), $F(1, 983) = 125.31$, $p < .001$, $d = 0.71$.

We then examined whether there was any main effect of the time of assessment of the MDMT; we did not find a difference between the number of "Does Not Fit" selections on the pre-test and the post-test, $F(1, 983) = 1.17$, $p = .28$, $d = 0.02$. We had predicted that the robot may have a higher frequency of "Does Not Fit" selections on the MDMT pre-test for the initial scenario than the human but that the frequency would decrease on the MDMT post-test for the robot in the updated scenario (i.e., showing a lower number of "Does Not Fit" selections), possibly pointing to evidence for perceived capacity changes in the robot. However, this did not seem to be the case. Rather, we found an interaction effect such that for the human agent people showed a slight decrease in selections from the pre-test ($M = 0.28$, $SD = 0.54$) to the post-test ($M = 0.22$, $SD = 0.49$), while for the robot agent they showed a slight increase in selections from the pre-test ($M = 0.71$, $SD = 0.81$) to the post-test ($M = 0.74$, $SD = 0.86$), $F(1, 983) = 9.94$, $p < .01$, $d = -0.06$.

Overall, the "Does Not Fit" results are best contextualized by what seems to be the primary driving force behind these selections, namely differences between "Does Not Fit" selections on the two superordinate trust factors. We first turn to a main effect of trust factor, where participants made a far greater number of these selections on the **Moral Trust** factor ($M = 0.88$, $SD = 1.31$) than on the **Performance Trust** factor ($M = 0.11$, $SD = 0.32$), $F(1, 983) = 397.73$, $p < .001$, $d = -0.80$. This shows that, in general, participants thought that some of the **Moral Trust** items applied less often to both the robot and human agents than did the **Performance Trust** items. However, perhaps the most impactful finding comes from the interaction between the agent and trust factor, such that while participants made slightly more "Does Not Fit" selections for the human on the **Moral Trust** factor ($M = 0.39$, $SD = 0.83$) than on the **Performance Trust** factor ($M = 0.11$, $SD = 0.29$), participants made a far greater number of

these selections for the robot on the **Moral Trust** factor ($M = 1.34$, $SD = 1.51$) than on the **Performance Trust** factor ($M = 0.11$, $SD = 0.36$), $F(1, 983) = 159.12$, $p < .001$, $d = -0.56$. This pronounced effect shows that participants thought that the **Moral Trust** items applied to the robot less often than the human, as well as less often than the **Performance Trust** items. This effect ties in with another effect identified in the results, reported below.

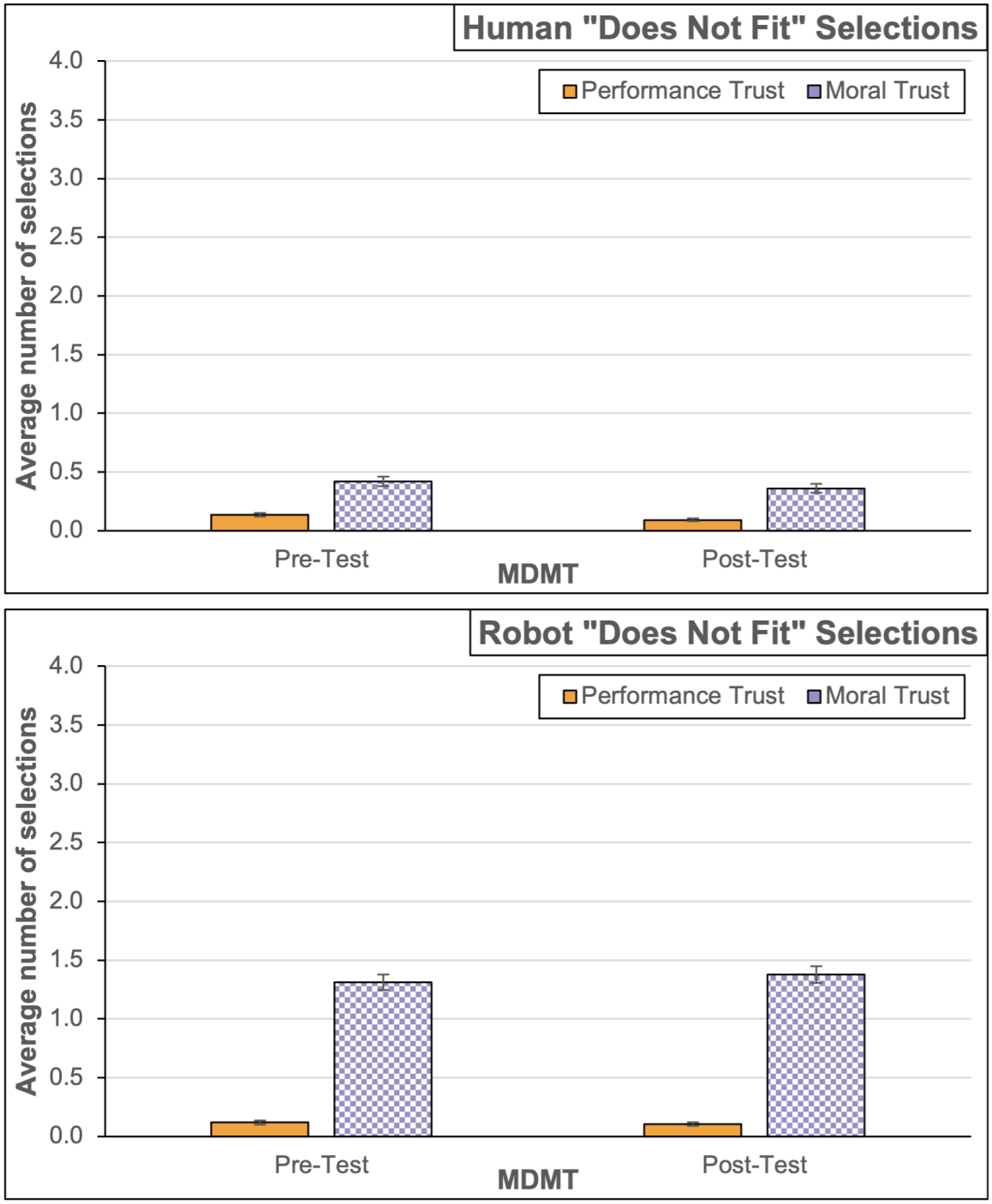


Figure 6-H. Trust Change Expansion Study: Average number of items (out of four per dimension) that received a "Does Not Fit" selection on the MDMT pre-test and the MDMT post-test for the **Performance Trust** factor and the **Moral Trust** factor for the human agent on top and the robot agent on the bottom.

We also calculated the number of participants who selected "Does Not Fit" for all items in a dimension or all items across dimensions for a factor on the MDMT pre-test and the MDMT post-test for the robot and human agents (Figure 6-I). This enables an examination of how many people did not consider a dimension, or sometimes an entire factor, applicable to the agent in the described scenario. These rates are lower for entire factors (**Performance Trust**, **Moral Trust**) than for individual dimensions (Reliable, Competent, Ethical, Transparent, Benevolent), pointing to the importance of a nuanced measurement tool that accounts for differentiable aspects of trust. In particular, we can look at the moral disqualification rates for the robot agent (i.e., where people do not treat the robot as a moral agent), which are also noticeably lower here than in previous work (Komatsu et al., 2021; Malle et al., 2016; Scheutz & Malle, 2021). Here we note that the highest moral disqualification rate from the entire **Moral Trust** factor was about 6% for the robot on the MDMT post-test; the highest rate for a single dimension was only slightly higher at about 13% for the robot on the Benevolent dimension on the MDMT post-test.

Agent	Pre-Test or Post-Test	Number of participants out of 985 total who selected "Does Not Fit" for all items in a dimension (left columns) or all items across dimensions for a factor (right columns)						
		Reliable	Competent	Ethical	Transparent	Benevolent	Performance	Moral
Human	Pre	0	1	8	29	17	0	2
	Post	0	0	14	23	20	0	5
Robot	Pre	2	3	84	59	101	1	33
	Post	1	4	114	84	126	1	57

Figure 6-I. Trust Change Expansion Study: Number of participants out of 985 total who selected "Does Not Fit" for all items in a dimension (left columns) or all items across dimensions for a factor (right columns) on the MDMT pre-test and the MDMT post-test for the human and robot agents.

In addition to the focal questions about the MDMT as a model and measure of trust, we also sought to evaluate whether there were any marked differences between people's evaluations of trust in robots compared to humans in identical scenarios, as well as whether people responded with different sensitivities to identical changes in the agent's behavior.

We first examined whether there were any notable starting differences between participants' evaluations of trust in robots and trust in humans using the MDMT pre-test scores—that is, after reading

an identical scenario with the only difference being whether a "robot" or "person" was described, did people generate different ratings of trust. Figure 6-J shows the MDMT pre-test scores collapsed into the two superordinate factors of trust (**Performance Trust**, **Moral Trust**) compared for the two types of agents (robot, human). We found a very large interaction effect, such that participants had comparable trust ratings on the **Performance Trust** factor for both the robot (M = 5.41, SD = 1.10) and the human (M = 5.40, SD = 1.07), but participants had far lower initial trust ratings on the **Moral Trust** factor for the robot (M = 3.85, SD = 1.85) than for the human (M = 5.27, SD = 1.18), $F(1, 947) = 277.36$, $p < .001$, d = -.54 (Figure 6-J). This showed the same substantial baseline Moral-Performance trust gap for the robot agent as observed in the original trust change study. Although the main effects of agent and superordinate trust factor were also significant, these were driven primarily by the large baseline Moral-Performance trust gap for the robot—i.e., the low ratings for the robot agent on the **Moral Trust** factor—captured by the interaction effect.

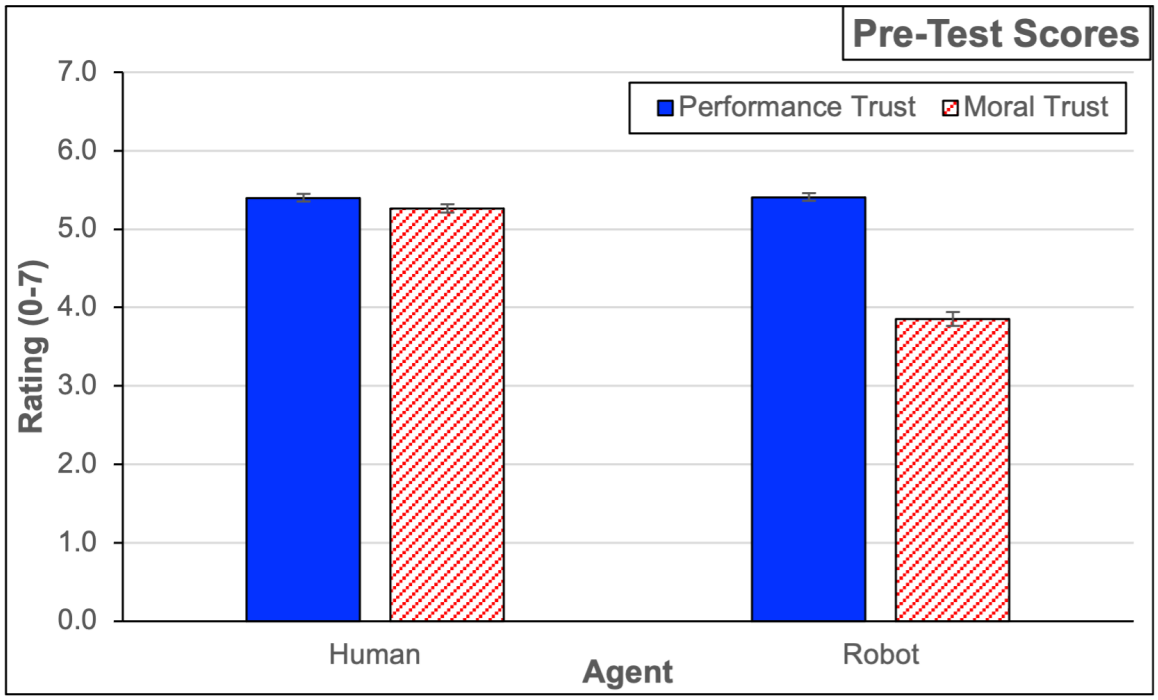


Figure 6-J. Trust Change Expansion Study: MDMT pre-test factor scores for the **Performance Trust** and **Moral Trust** factors (legend) on a 0-7 scale (y-axis) for the human agent on the left and the robot agent on the right (x-axis). Error bars show *SE*.

Next, we examined whether there was any variation in the magnitude of change in trust based on the type of *Agent* (robot, human) by examining the absolute values of the MDMT difference scores by *Change Direction* (up, down), as shown in Figure 6-K. We first examined whether there was a main effect of *Change Direction*, testing to see if people respond more strongly to evidence for decreases in trust compared to increases in trust—i.e., whether losses in trust are larger than gains in trust. We had found evidence for this trust loss effect in the original trust change study, such that the magnitudes of difference scores were much larger for trust losses than trust gains in the robot (as shown in Figure 4-C). We replicated this pattern, such that a main effect showed much larger losses in trust ($M = 1.59$, $SD = 1.27$) than gains in trust ($M = 0.83$, $SD = 0.71$), $F(1, 722) = 88.92$, $p < .001$, $d = .73$. We then investigated the type of agent, testing to see if people respond differently to otherwise identical scenarios based solely on the type of agent (robot, human). There was a significant main effect of *Agent*, such that participants showed greater changes in trust in response to evidence about the human ($M = 1.37$, $SD = 1.26$) than about the robot ($M = 1.05$, $SD = 0.85$), $F(1, 722) = 15.69$, $p < .001$, $d = .29$. Interpretation of the main effect is qualified by the presence of an interaction effect of *Agent* with *Change Direction*, such that participants had more comparable changes of trust in the human ($M = 0.87$, $SD = 0.71$) and robot ($M = 0.77$, $SD = 0.72$) given evidence increasing trust, but participants had a larger loss of trust in the human ($M = 1.80$, $SD = 1.46$) than the robot ($M = 1.30$, $SD = 0.88$) given evidence decreasing trust, $F(1, 722) = 6.47$, $p = .01$, $d = -.20$ (Figure 6-K). This is explained in the Discussion below.

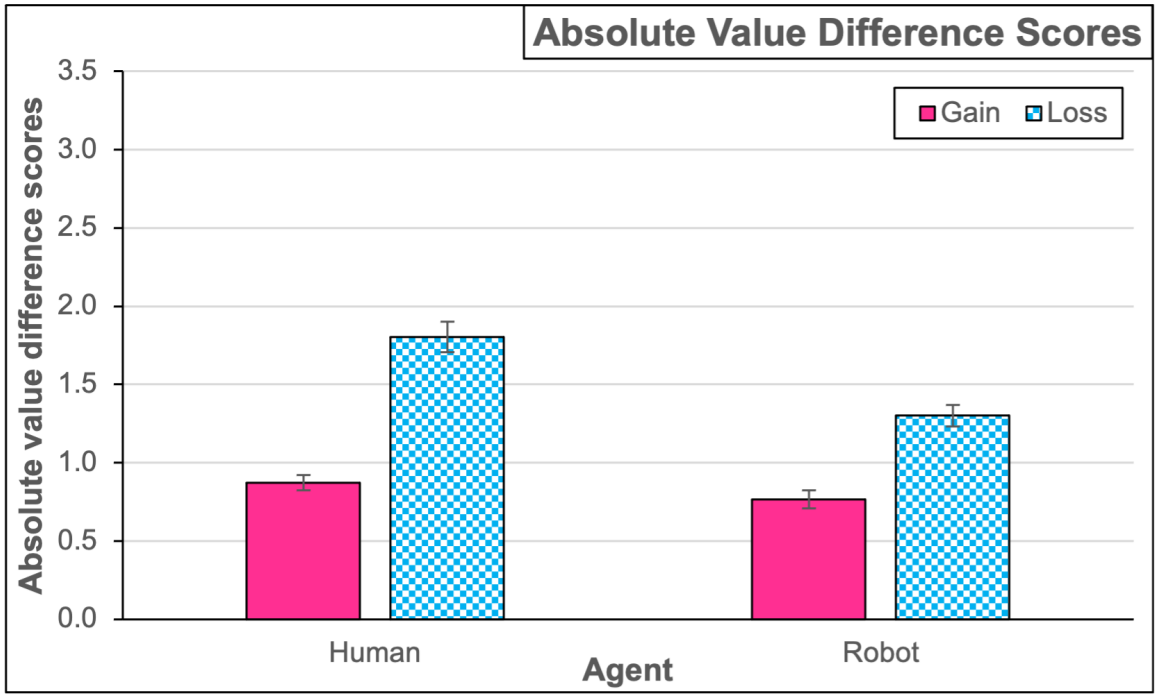


Figure 6-K. Trust Change Expansion Study: Absolute value difference scores (y-axis) for the human agent on the left and the robot agent on the right (x-axis) for evidence expected to prompt trust gain or trust loss (legend). Error bars show *SE*.

6.3 Discussion

The trust change expansion study results tell a clear story—the updated MDMT v2 was able to successfully measure trust and changes in trust for both robot and human agents. This offered a powerful validation of the MDMT v2 as both a multi-dimensional conception and measurement tool for trust—not only in robots, but in humans as well.

I start by broadly discussing results concerning the two superordinate trust factors. First, the baseline **Performance Trust** ratings on the MDMT pre-test were fairly high for both the robot and human agents, and while the **Moral Trust** ratings were similar to the **Performance Trust** ratings for the human agent, the **Moral Trust** ratings were substantially lower than the **Performance Trust** ratings for the robot agent (shown broadly in Figure 6-A and specifically in Figure 6-J). We thus replicated a baseline Moral-Performance trust gap in the baseline ratings specifically for the robot agent, and also noted that this effect did not occur for the human agent. People gave lower, more modest **Moral Trust** ratings to a robot agent when they had no prior experience with the described robot agent. These baseline ratings were

not negative ratings, but they were relatively lower than for human agents in comparable situations—an important consideration for the topic of appropriate trust in human-robot interaction research and design. Additional robustness of this finding stems from the fact that ratings here were distinctly not an indication that people thought the ratings did not fit the robot—these attitudes were captured in "Does Not Fit" selections for the robot. Actual rating averages were therefore plausible reflections of trust expectations—just lower, more modest baseline **Moral Trust** for robot agents.

We also examined the magnitude of change in trust based on *Agent* (robot, human) and *Change Direction* (up, down) from the absolute value difference scores (Figure 6-K). We found a large main effect of *Change Direction*, such that participants responded more strongly to evidence for decreases in trust as compared to increases in trust, with larger magnitudes of trust loss than trust gain. There was also a main effect of *Agent*, such that participants showed greater changes in trust in response to evidence about the human than about the robot, as well as an interaction such that the change in trust was largest for the human when losing trust. This can likely be attributed to the fact that the human agent had higher baseline ratings of trust in the pre-test for the **Moral Trust** dimensions (as seen in Figure 6-A) and thus had more trust to lose than the robot agent; looking at the post-test scores (Figure 6-B), the final trust levels are more comparable for the down level between the robot and human agents.

We now turn to the specific dimensions of the MDMT. As noted, the dimensions were quite responsive to changes in trust, as demonstrated by the difference scores in Figure 6-C, by statistical testing for the first hypothesis test (focal subscale score compared to other subscale scores for the corresponding focal *Targeted Trust Dimension*) using *t* values in Figure 6-D and remarkably large corresponding Cohen's *d* values in Figure 6-E, and by statistical testing for the second hypothesis test (focal subscale score for the corresponding focal *Targeted Trust Dimension* against the scores for the same subscale when other dimensions were the *Targeted Trust Dimension*) using *F* tests and corresponding Cohen's *d* values in Figure 6-F. These results illustrated that the MDMT has a high degree of dimensional specificity for the multi-dimensional concept of trust and that the MDMT is successful as a trust measurement tool.

The results for the first hypothesis test were pronounced—the subscales of the MDMT were highly sensitive to changes in their corresponding focal *Targeted Trust Dimension* as indicated by the relevant *t* values (Figure 6-D) and large corresponding Cohen's *d* values (Figure 6-E). This tells an unmistakable story—people respond to differential changes in evidence for dimensions of trust in an agent and the MDMT is sensitive to these changes. There were a couple of exceptions noted for specific cases. In the Ethical robot down condition, the trust loss prompted low Ethical ratings but these were slightly overshadowed by an even greater loss in Reliable ratings for the robot; additional work may be able to clarify this occurrence. Even more strikingly, we note that expected shifts in Reliable ratings were often overpowered by shifts in Competent ratings (in three out of four cases). We believe this is likely a function of the experimental design, which only offered two time points for evaluation of an agent. In essence, though Reliable and Competent are related to the broader idea of **Performance Trust**, Reliable likely has more to do with the repeated performance of an agent over time—such that two time points offer only minimal evidence over time to surmise a changed sense of reliability. Ratings for Reliable may thus shift more slowly than ratings for Competent. Additional work looking at changes over a greater number of time points may be able to discern whether this explanation accounts for the observed pattern. This raises the question of how evaluations of trust in an agent may change over a greater number of interactions, as well as over the course of longer interactions. This may prove particularly interesting for the baseline Moral-Performance trust gap for robot agents; it is possible that this gap may shrink over time with increased interaction in everyday life with robot agents, such that the baseline ratings for robot agents might eventually start to resemble the ratings for human agents.

The results for the second hypothesis test were not as clear-cut, as the focal subscale score for a corresponding focal *Targeted Trust Dimension* was sometimes slightly overshadowed by its difference score for a different *Targeted Trust Dimension* (usually within the same superordinate trust factor). For example, looking at Figure 6-C we can see that for both the robot and human agents in the up level, the Competent subscale scores were greater for the Reliable *Targeted Trust Dimension* than for the Competent *Targeted Trust Dimension*. Perhaps this stems from variation inherent to differences across

scenarios in the vignettes, or from the particular stimulus design of the trust information changes presented in the vignettes. In other words, it is possible that the different application domains and associated change stimuli described in the vignettes drove slightly different changes along trust dimensions than what we expected. For example, the airport security and school tutor vignettes are quite different, as are the associated information changes within the corresponding vignettes that prompt changes in trust (as can be seen in Appendix F). I therefore do not speculate on what caused the variation observed for specific cases, but note that disambiguation of this possibility could involve further refinement of vignettes to test salient changes in trust along trust dimensions. However, despite this nuance, overall the statistical testing (Figure 6-F) revealed significant effects for all five subscales across both agents with small to moderate effect sizes when comparing the magnitude of difference for the focal subscale score for the corresponding focal *Targeted Trust Dimension* against the scores for the same subscale when other dimensions were the *Targeted Trust Dimension* (with the exception of Reliable for the human agent). This offers validation of the MDMT even across different vignettes and evidence for changes in trust.

As in the original trust study, the updated dimensions within **Performance Trust** (Reliable, Competent) and **Moral Trust** (Ethical, Transparent, Benevolent) often moved in tandem (Figure 6-C). The inclusion of the new Benevolent dimension was an important revision to the conception and measurement tool offered by the MDMT. Ratings on the Benevolent dimension paralleled ratings on the other **Moral Trust** dimensions, while simultaneously showing differentiation from the other dimensions (as is shown in Figure 6-C, Figure 6-D, Figure 6-E, and Figure 6-F).

Finally, the inclusion of a "Does Not Fit" option was shown to be a critical aspect of the psychometric design of the MDMT. Figure 6-G shows the average frequency of "Does Not Fit" selections per trust dimension for the robot and human agents on the MDMT pre-test and the MDMT post-test. There are a couple of important patterns observed here, which were confirmed via statistical testing and are clearly visible in Figure 6-H. First, participants selected "Does Not Fit" for both agents more on the **Moral Trust** dimensions than on the **Performance Trust** dimensions. Second, and most strikingly as

seen in Figure 6-H, participants selected "Does Not Fit" far more often for the robot agent on the **Moral Trust** dimensions than for the human agent. Quite interestingly, we had previously seen a related pattern from participants who did think the items fit the agents and assigned numerical ratings—such that participants gave much lower initial ratings to the robot on **Moral Trust** than **Performance Trust**, an effect captured in the baseline Moral-Performance trust gap for the robot agent shown in Figure 6-J. Taken together, it appears that participants thought that **Moral Trust** items did not fit a robot agent as often as a human agent—and that even when participants did think these items fit a robot agent (and offered a numerical rating), these ratings were lower for **Moral Trust** for the robot agent than for the human agent. In sum, the inclusion of the "Does Not Fit" option highlighted a marked difference in trust evaluations of robot and human agents. The absence of this option might have otherwise prompted participants to make judgments they considered invalid by offering ratings on items they did not consider appropriate for the agent in the situation, such as from moral disqualification of the agent. Figure 6-I uses the "Does Not Fit" selections to show the rates of moral disqualification for agents (primarily for the robot agent) when participants thought that none of the **Moral Trust** items fit for the entire factor (as well as shows the rates for individual dimensions). From this analysis, it is clear that "Does Not Fit" is an integral feature of the psychometric design of the MDMT—a topic explored further in the General Discussion (Chapter 8).

Chapter 7: Applications of the MDMT

We have used the MDMT in a number of research studies to help validate and further refine the conception and measurement of trust. In this chapter I describe three studies that offer validation of the MDMT using MDMT v1 (the version available at the time) benchmarked against other related measurements in human-robot interaction studies. I first detail two applications of the MDMT in lab robotics studies I worked on that offer insight into the real-world use of the measurement tool—one application with a small, non-humanoid robot and one application with a large, industrial, semi-humanoid robot. I then detail an effort by other researchers (Chita-Tegmark et al., 2021) that highlights the efficacy of the MDMT, and specifically its inclusion of "Does Not Fit," compared with other trust measurement tools. A number of other researchers have also begun to use the MDMT (e.g., Banks, 2020b, 2020a; Law et al., 2020, 2021). We have made the MDMT publicly available for researcher use and have asked researchers to assist us with community validation of the MDMT.

7.1 "Challenges and Opportunities for Replication Science in HRI: A Case Study in Human-Robot Trust" (Ullman et al., 2021)

As human-robot interaction (HRI) researchers, like all scientists, we must demonstrate the reproducibility of findings—especially across robots. We conducted a replication effort that illustrated the challenges and opportunities for replication science in HRI. This project was published in a paper in the HRI conference proceedings (Ullman et al., 2021), and the paper received the Best Theory and Methods Paper Award at the conference. I first summarize the project for context, and then I identify the relevant study in the project that used the MDMT and illustrates a valuable validation of the tool.

We had conducted a previous human-robot trust study (Ullman & Malle, 2017) that suggested that people "in the loop" with a robot via a simple button press trusted the robot (a Thymio) more than those who observed the robot complete the task autonomously. This intriguing finding was based on a small sample ($n = 40$) and was therefore greatly underpowered (observed power $1 - \beta = .35$), prompting replication.

To test whether the in-the-loop effect generalizes to similar robots, we conducted a conceptual replication (Study 1) using the Create robot. The effect did not replicate, despite a large sample (n = 140) and expected power of .86. This result called for a direct replication (Study 2) using the original Thymio robot. The effect again did not replicate, despite a large sample (n = 200) and expected power of .96. We then conducted an online study (Study 3) with videos of both robots to examine whether different expectations for each robot drove the divergent results, but the hypothesis was disconfirmed (n = 400). However, one finding held across all studies: Participants consistently trusted imagined future robots far less in social than nonsocial use contexts, pointing to a social-nonsocial trust gap for robots (effect sizes of ds = −0.71, −0.78, and −0.79 in the lab studies; d = −0.38 in the online study). This consistent and substantial effect is shown in Figure 7-A. This future trust finding was generated using a measure that was a composite of "How much would you trust the robot in this scenario?" and "How willing would you be to use the robot in this scenario?"; both questions were answered on 8-point rating scales (from 0 = "Not at all" to 7 = "Very"). Although this finding was not assessed using the MDMT, the identified social-nonsocial trust gap yields an important insight into human-robot trust that converges with results from the MDMT; I explore this insight later in this chapter and in the General Discussion (Chapter 8).

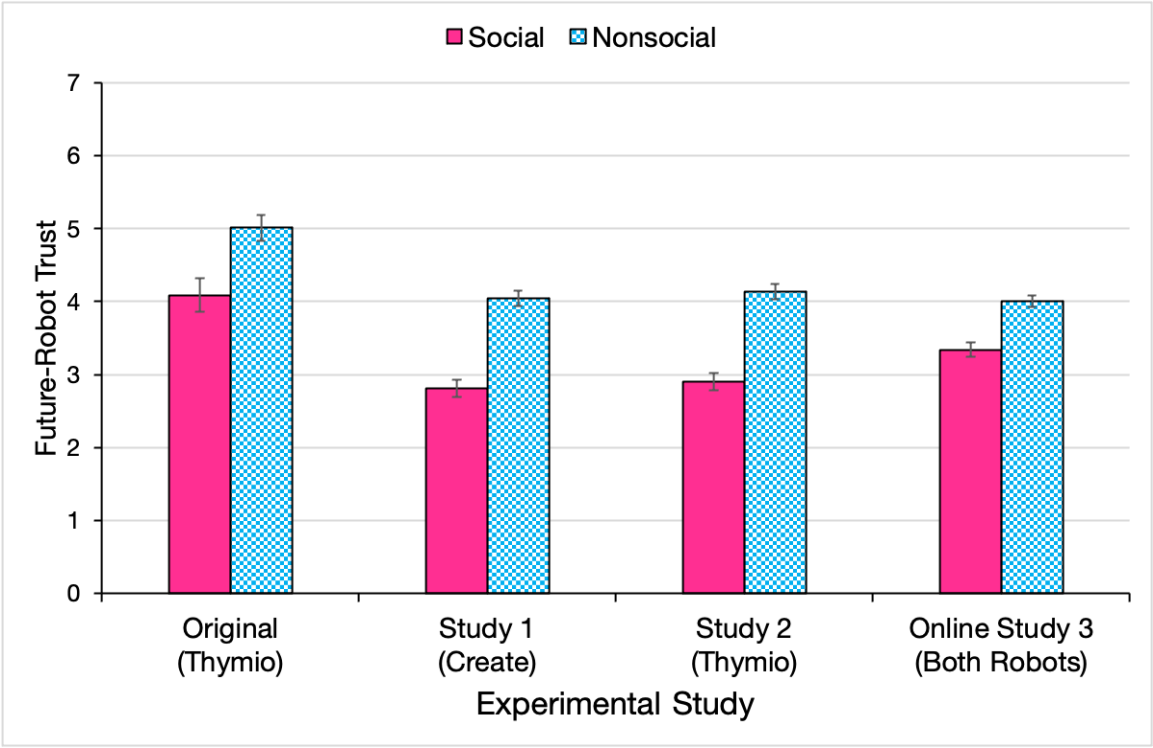


Figure 7-A. MDMT Application One: Trust in imagined future robots for the original study (Thymio), Study 1 (Create), Study 2 (Thymio), and online Study 3 (Both Robots). Figure demonstrates social-nonsocial trust gap for robots; this finding was assessed using a different two-item trust measure, not the MDMT. Error bars show *SE*.

In the original study, the primary measure of trust was the average judgment of the Thymio robot on two items, "trustworthy" and "dependable," assessed on 8-point rating scales (from 0 = "Not at all" to 7 = "Very"). In Study 1, participants offered their judgments on the same two items and then also responded to the 16 items of the MDMT v1. This maintained the initial study presentation for the purpose of the replication effort, while simultaneously allowing us to expand and potentially compare use of the MDMT in a real-world human-robot interaction study. The study was conducted prior to the creation of the MDMT v2, such that it used the MDMT v1. The insights are thus tempered by the issues of the original version of the MDMT, but the results are nonetheless quite striking.

I first present the relevant results using the original two-item trust measure, and then share the results from the MDMT. In the original study, people in the button involvement condition tended to indicate more trust in the robot ($M = 5.53$, $SD = 1.38$, $n = 20$) than people in the autonomous condition (M

= 4.75, *SD* = 1.59, *n* = 20), *F*(1, 38) = 2.70, *p* = .11, *d* = 0.52. In Study 1, people in the button involvement condition tended to indicate less trust in the robot (*M* = 5.24, *SD* = 1.62, *n* = 57) than people in the autonomous condition (*M* = 5.57, *SD* = 1.17, *n* = 75), though also not significantly, *F*(1, 130) = 1.92, *p* = .17, *d* = −0.24. We next analyzed a new verbal involvement condition included in Study 1. People who were involved with the robot by giving it verbal permission to proceed tended to show less trust in the robot than people who merely observed the autonomous robot, but this result was also not significant *F*(1, 182) = 2.13, *p* = .15, *d* = −0.34. Just as in the main hypothesis test, the means went in the opposite direction from what one would expect from the original study.

We then examined participants' trust in the Create robot using the MDMT v1 (Figure 7-B). We used the MDMT overall score and dimension subscale scores. The MDMT showed the same nonsignificant pattern of means in the direction opposite to the hypothesis as the original two-item trust measure: People in the button involvement condition indicated less trust on the overall MDMT (*M* = 5.16, *SD* = 1.42, *n* = 60) than people in the autonomous condition (*M* = 5.55, *SD* = 1.03, *n* = 80), *F*(1, 138) = 3.49, *p* = .06, *d* = −0.32. Likewise, people in the new verbal involvement condition also tended to show less trust on the overall MDMT than people in the autonomous condition, *F*(1, 196) = 3.88, *p* = .05, *d* = −0.20, this time reaching a traditional significance level.

We also examined the MDMT subscales: Reliable (α = .77, *n* = 86), Capable (α = .84, *n* = 152), Ethical (α = .92, *n* = 49), and Sincere (α = .94, *n* = 51). Their intercorrelations were very high (Reliable with the remaining scales, *r*s = .34 to .66; Capable, *r*s = .38 to .73; Ethical, *r*s = .67 to .82; Sincere, *r*s = .72 to .83). Of interest, some subscales appeared to be more sensitive than the MDMT overall score: Reliable (*p* = .04) and Ethical (*p* = .02) showed significantly less trust for each of the involved conditions compared with the autonomous condition (Figure 7-B).

There are two relevant high-level takeaways from this research effort. First, the MDMT—even the MDMT v1—was able to differentiate between separable dimensions of trust in a robot in the real world. This study offered the first instance of the MDMT used with a physical embodied robot, and the results were quite promising. As in the other previous work detailed in this dissertation, the dimensions

within each superordinate factor were more similar to each other than they were to dimensions across factors. Most importantly, not only did the MDMT overall score roughly parallel the original two-item trust measure, but the MDMT even offered additional insight from the individual subscale scores going beyond the capacity of the original trust measure. Also, the number of participants who selected "Does Not Fit" was greater for items within the two **Moral Trust** dimensions (Ethical, Sincere) than for items within the two **Performance Trust** dimensions (Reliable, Capable); this can be seen in the lower number of participants that offered numerical ratings for the **Moral Trust** subscales (i.e., the *n*s in the subscale statistics above). This finding aligns with the previously documented pattern.

In sum, this study provided a first comparison of the MDMT against another less-specialized trust measure with a robot in the physical world and showed that the nuance offered by the MDMT reflects important nuance in real-world trust attribution to robots—just as we saw in the online studies in the trust change series (Chapter 4 and Chapter 6).

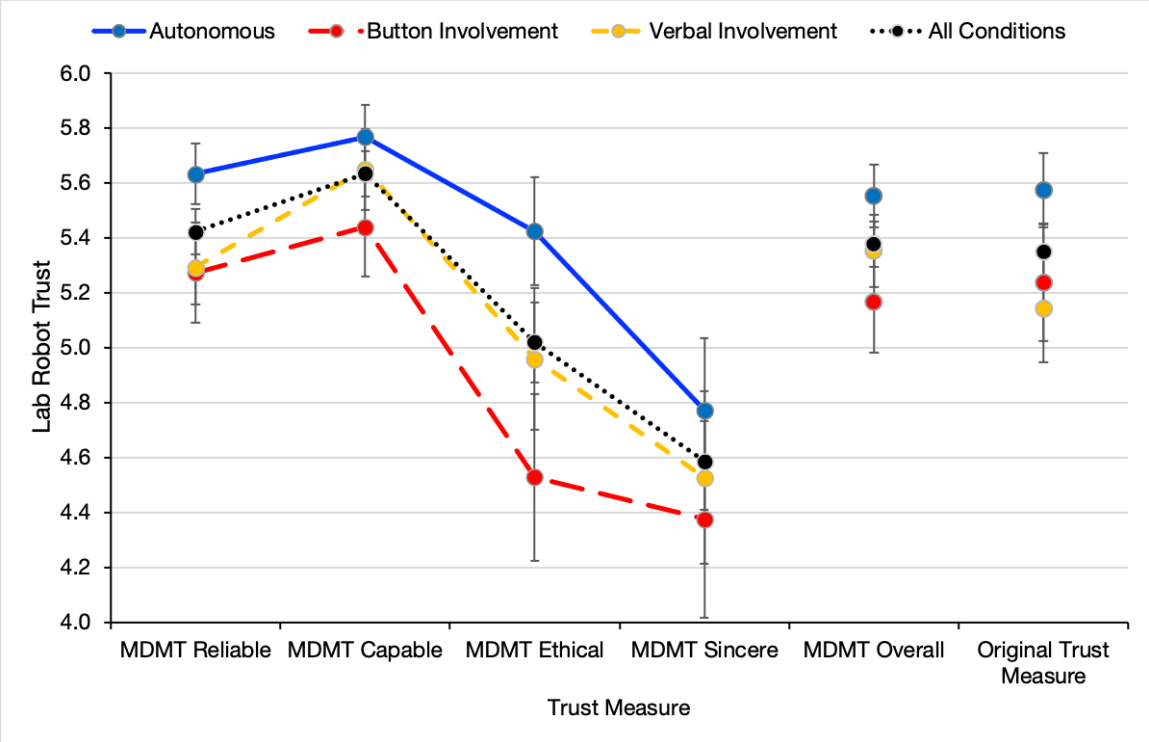


Figure 7-B. MDMT Application One: Trust in the Create robot in the first conceptual replication study, assessed on the MDMT subscales, the MDMT overall score, and the original two-item trust measure. Error bars show *SE*.

7.2 "Mixed Reality as a Bidirectional Communication Interface for Human-Robot Interaction" (Rosen et al., 2020)

The project described in this section was conducted in collaboration with the Humans To Robots Lab (H2R) at Brown University. This project was published in an IROS conference proceedings paper (Rosen et al., 2020), and the paper was a finalist for the Best Paper Award on Cognitive Robotics at the conference. I first summarize the project for context and then identify the relevant findings.

In essence, we tested the hypothesis that virtual deictic gestures via a mixed reality (MR) interface are better for human-robot interaction (HRI) than physical behaviors to disambiguate item references. To test this hypothesis, we proposed the Physio-Virtual Deixis Partially Observable Markov Decision Process (PVD-POMDP), which interprets multimodal observations (speech, eye gaze, and pointing gestures) from the human and decides when and how to ask questions (either via physical or virtual deictic gestures) in order to recover from failure states and cope with sensor noise. We conducted a between-subjects user study with 80 participants distributed across three conditions of robot communication: no feedback control (CTRL), physical feedback (PHYS), and MR feedback (MR). We tested performance of each condition with objective measures (accuracy, time), as well as evaluated user experience with subjective measures (usability, trust, workload). Ultimately, we found that models that integrate feedback performed better and were preferred by users on all subjective measures, and that MR is a promising modality for this communication.

We used the MDMT v1 to assess trust in the Baxter robot—a large industrial, semi-humanoid robot—and chose to only use the **Capacity Trust** factor (consisting of the Reliable and Capable dimensions) as the task was centered on notions of performance without any expectations of moral capacities. We tested several hypotheses using a number of measures; I include relevant results that show the consistency of the MDMT with the other subjective measures, as well as include a general discussion of the results to highlight the alignment of the MDMT with the other findings (results and graphs of the other measures are available in full in the paper itself). For the other subjective measures, we used the System Usability Scale (SUS) to assess participants' usability experiences (Brooke, 1996) and the raw

version of the NASA Task Load Index (NASA-TLX), often referred to as the RTLX, to assess participants' experienced workload of using the system (Hart, 2006; Hart & Staveland, 1988).

The three subjective dependent measures (SUS, MDMT, RTLX) were all correlated (ps < .001) with each other: r = .65 for SUS and MDMT; r = -.60 for SUS and RTLX; and r = -.54 for MDMT and RTLX. These correlations suggest that perceptions of usability and trust increase in tandem, and that workload decreases as both usability and trust increase. The correlations between the dependent variables also indicated that a multivariate analysis of the data was warranted to account for the relationships among the dependent variables.

A MANOVA was conducted using a pair of a priori orthogonal Helmert contrasts on the **Capacity Trust** score data (Figure 7-C). We first tested the hypothesis that the feedback conditions would facilitate better user experiences than the no feedback condition. The first Helmert contrast was significant and supported the first hypothesis, $F(3, 75) = 3.13$, $p = .03$, *multivariate* $\eta^2 = .11$. For the MDMT **Capacity Trust** score, the univariate F test revealed that the feedback conditions were rated as significantly higher on trust, $F(1, 77) = 7.56$, $p < .01$, $\eta^2 = .09$. We then tested the hypothesis that the MR feedback condition would facilitate better user experiences than the physical feedback condition. The second Helmert contrast was not significant, $F(3, 75) = 1.19$, $p = .32$, *multivariate* $\eta^2 = .05$. However, the means of the measures were consistent with the second hypothesis. For the MDMT **Capacity Trust** score, the univariate F test was not significant, $F(1, 77) = 2.86$, $p = .10$, $\eta^2 = .04$.

The results from the objective and subjective measures in our user study painted a single, coherent story about the conditions we tested. In general, the feedback conditions (physical, MR) outperformed the no feedback condition, and the MR feedback condition outperformed the physical feedback condition. The user experience of each condition roughly paralleled the performance of the system. Participants gave better user experience ratings across all three subjective measures (usability via SUS, trust via MDMT, workload via RTLX) in the feedback conditions (physical, MR) as compared to the no feedback condition (control). Although there was no statistically significant difference between the user experience ratings in the MR feedback condition and the ratings in the physical feedback condition,

the means across all three subjective measures improve from no feedback to physical feedback, and again from physical feedback to MR feedback.

These findings offer an additional validation of the MDMT, even using the MDMT v1. The study sought to examine the performance of a new system with a robot in a comparison between participants using objective measures of task performance and subjective measures of participant experience. The MDMT not only aligned with the other subjective measures in the directions predicted, but also aligned with the objective measures of task evaluation. There are thus two relevant high-level takeaways from this study. First, the MDMT was again able to serve as a valuable measure of trust for a physical robot in the real world. In addition, whereas the application of the MDMT in the previous section was with the Create robot—a small, non-humanoid robot—in this case the MDMT was used with the Baxter, a very different robot—a large, industrial, semi-humanoid robot. Validation of research across robot platforms—including the use of the MDMT—is essential to HRI in general, as noted in the previously discussed replication effort (Ullman et al., 2021). Second, this study was able to use only one factor of trust, **Capacity Trust** (now called **Performance Trust**), and its corresponding component dimensions, Reliable and Capable (now called Competent)—that is, the study was able to selectively use the factor/dimensions deemed relevant to the experimental design. The success of the MDMT in this format demonstrates the versatility of the measure from its inherent identification of differentiable components of trust.

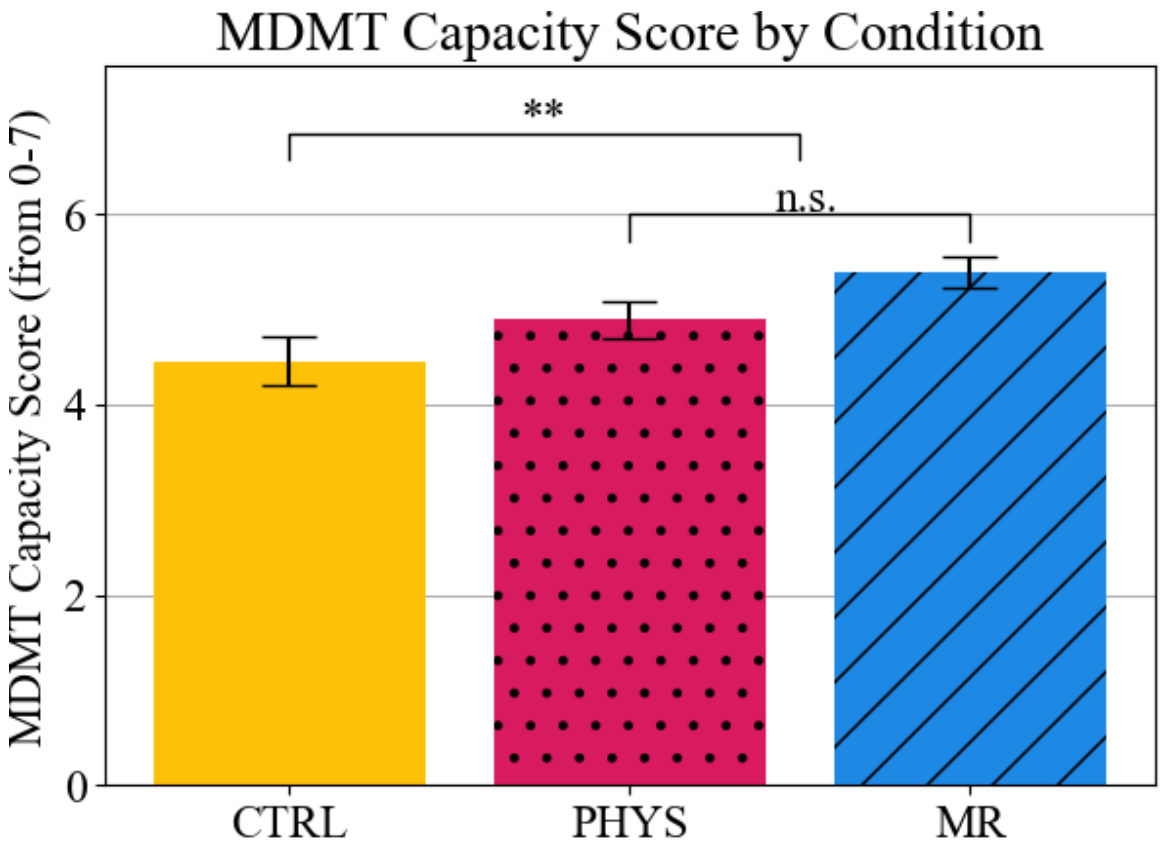


Figure 7-C. MDMT Application Two: MDMT **Capacity Trust** score for the three between-subjects conditions of no feedback control (CTRL), physical feedback (PHYS), and MR feedback (MR). Error bars show *SE*.

7.3 "Can You Trust Your Trust Measure?" (Chita-Tegmark et al., 2021)

Researchers at Tufts University conducted an evaluation of trust measures used in the field of human-robot interaction (HRI), which they published in a paper titled "Can You Trust Your Trust Measure?" in the HRI conference proceedings (Chita-Tegmark et al., 2021). The authors address the issue of inconsistent trust measurement across the field, noting that HRI studies use a variety of trust measures from a range of fields and adjust them if they are not specifically about robots, as well as use these measures with variable implementations of rating scales. This prompts significant validity concerns for individual instances of implementing a trust measurement tool, as well as impedes comparison and generalization across HRI studies that involve human-robot trust. The paper authors argue for

standardizing trust measurement and expanding the capacity of trust measures to match the scope of human-robot interaction.

The paper authors tested four trust measurement tools: the Reliance Intention Scale (RIS) (Lyons & Guznov, 2019), the Trust Perception Scale-HRI (TPS-HRI) (Schaefer, 2013), and the MDMT v1 split into the **Capacity Trust** factor (which they refer to as "MDMT-Capable & Reliable" or "MDMT-C&R") and the **Moral Trust** factor (which they refer to as "MDMT-Sincere & Ethical" or "MDMT-S&E"). The authors aligned the rating scales between measures for direct comparison, such that all of them used a 7-point Likert scale with two N/A options—"Non-applicable to this robot" and "Non-applicable to robots in general." The authors tested the inclusion of these N/A options to see whether participants believed the items were applicable to robots—either in specific described instances or even just in general.

One of the critical features of the MDMT—by design—is its inclusion of the "Does Not Fit" option for the very purpose of minimizing ratings by participants on items they do not consider applicable. This sets the MDMT apart from other typical trust measures. The study conducted by Chita-Tegmark and colleagues illustrates the essential benefit of this feature.

There are a number of interesting results in the study (Chita-Tegmark et al., 2021), with the primary relevance captured in an excerpt from the paper abstract in conjunction with a figure from the paper (obtained and used with permission from the paper authors):

> "In Study 1 we show that after being presented with a robot (non-humanoid) and an interaction scenario (fire evacuation), participants rated multiple questionnaire items such as "This robot is principled" as "Non-applicable to robots in general" or "Non-applicable to this robot." In Study 2 we show that the frequency of these ratings change (indeed, even for items rated as N/A to robots *in general*) when a new scenario is presented (game playing with a humanoid robot)."

The figure from their original paper is relabeled as Figure 7-D below, and the caption includes the original caption as noted. The figure shows differences in the percentage of items that received N/A evaluations across all four tested questionnaires, and also within questionnaires between Study 1 and Study 2—and differences are especially apparent from the two factors of the MDMT. In Study 2 the authors compared an evacuation scenario involving a non-humanoid robot against a game-playing

scenario involving a humanoid robot. They conducted ANOVAs with contrast analyses to compare the scenarios and the four trust measurement questionnaires. The authors reported two significant effects (both $p < .001$) specifically for the MDMT-Sincere & Ethical subscale (i.e., the MDMT v1 **Moral Trust** factor), such that participants who saw the game-playing scenario gave a lower proportion of "N/A to robots in general" ratings and "N/A to this robot" ratings than participants who saw the evacuation scenario. (Note: There was a reporting error in the original paper—confirmed by the study authors—where the proportion for "N/A to this robot" was originally reported instead as lower for the evacuation scenario; the effect is correctly reported in the main text here.) The authors did not report any such effects for the other three tested trust subscales—including the MDMT-Capable & Reliable subscale (i.e., the MDMT v1 **Capacity Trust** factor), where the N/A evaluations were more similar between the two scenarios and thus did not show the same pattern as the MDMT-Sincere & Ethical subscale.

After reporting the effect for "N/A to robots in general," the authors remark "These results suggest that people's mental models of robots are not stable across all scenarios, and they shift when they are presented with robot interactions that could be construed as social" (Chita-Tegmark et al., 2021). This effect, such that individual trust item ratings are more applicable in the more social, humanoid robot scenario (i.e., with a lower incidence of N/A evaluations) points to variability in perceptions of robots based on the type of robot and associated context. This effect intersects with two other effects previously discussed and documented in this dissertation. First, the trust change expansion study (Chapter 6) documented a higher frequency of "Does Not Fit" selections for items within the **Moral Trust** factor than for items within the **Performance Trust** factor (Figure 6-G and Figure 6-H). This effect aligns with the proportions seen in Figure 7-D, such that people generally gave more N/A evaluations on the MDMT-Sincere & Ethical subscale (corresponding to **Moral Trust**) than on the MDMT-Capable & Reliable subscale (corresponding to **Capacity Trust**). Although Figure 7-D is from a study using a modified version of the MDMT v1, while Figure 6-G and Figure 6-H are from a study using the MDMT v2, the results parallel each other indicating the robustness of the MDMT. Second, the previously discussed human-in-the-loop trust project (Ullman et al., 2021) consistently replicated and showed a

pronounced social-nonsocial trust gap for imagined future robots—such that participants consistently trusted imagined future robots far less in social than nonsocial use contexts (although not measured using the MDMT). Taken together, these effects converge to a potentially important and interesting conclusion: People seem to consider **Moral Trust** items less applicable to robots in general (and non-humanoid robots in nonsocial contexts in particular), but they consider these items more applicable to robots that appear more human-like and operate in more social contexts.

In sum, the discussed paper (Chita-Tegmark et al., 2021) not only demonstrates the importance of having a "not applicable" option for trust measurement scales—which the MDMT does by design with its "Does Not Fit" option—but also illustrates differences between the two factors of the MDMT. The MDMT in its full form consists of both **Performance Trust** items and **Moral Trust** items, such that the MDMT is a single measurement tool able to capture the multi-dimensionality of trust.

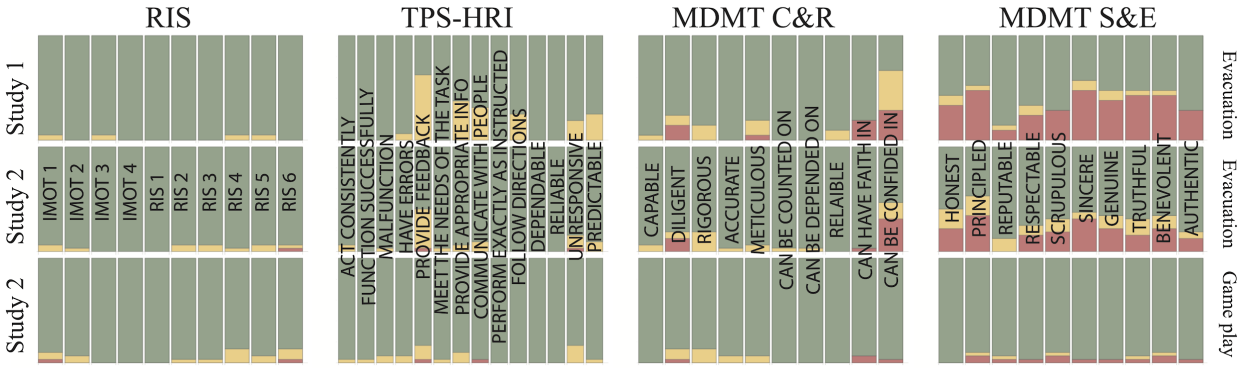


Figure 7-D. MDMT Application Three: Figure used with permission (Chita-Tegmark et al., 2021). Original figure caption: "Rating percentages by item: "N/A to robots in general" (red), "N/A to this robot" (yellow)."

Chapter 8: General Discussion

Over the course of the past six years, I have worked together with my advisor to study trust in human-robot interaction in order to create a model and measure of human-robot trust. Starting with a comprehensive literature review and examination of existing theoretical frameworks of trust in the human-automation, human-human, and human-robot domains, we took stock of foundational work to understand the trust landscape. We then conducted a series of studies to investigate how people think about ideas related to trust, which over time revealed an underlying structure that supports a multi-dimensional conception of trust. We created a first version of a trust conception and corresponding measurement tool, called the Multi-Dimensional Measure of Trust (MDMT), and made the MDMT v1 (Appendix A; Ullman & Malle, 2019a) publicly available for research use. We tested the MDMT v1 and refined the dimensional structure, resulting in the publicly available MDMT v2 (Appendix B; Ullman & Malle, 2020). We then tested the MDMT v2, finding strong support for the current dimensional structure. I summarize the trajectory of and insights from this process below.

8.1 Trust Factors, Dimensions, and Items

An analysis of the existing trust literatures revealed two broad ideas about trust: The human-automation literature showed themes of concepts and measurement centered on what we refer to as **Performance Trust**, whereas the human-human literature often showed themes centered on what we refer to as **Moral Trust**. Neither domain showed an exclusive focus on the component ideas within each factor, but rather an emphasis on the respective ideas. As discussed in the Introduction (Chapter 1), a number of previous models of trust have incorporated aspects spanning both factors (Figure 1-A), pointing to a breadth of interpretations of trust. Robots sit somewhere on a continuum between simple machine and human agent, such that a conception of trust applicable to robots in human-robot interaction likely reflects ideas spanning the different types of trust. Previous work had begun to tackle trust in human-robot interaction as documented in the Introduction, and the field warranted further systematic investigation to continue to uncover the intricacies of human-robot trust.

After identifying the two broad ideas about trust in the various literatures, we sought to test whether these ideas would be naturally represented in participant evaluations of ideas about trust (agnostic to type of agent or context). We ran the trust items rating study (Chapter 2) with a set of candidate trust items to see how participants would evaluate items along a rating scale continuum with two poles reflecting the two ideas. Somewhat surprisingly, the resulting item structure did not separate out into just two clusters of items, but instead decomposed into four clusters of items; it was intriguing that the experimental design using a two-component scale structure yielded a four-component solution. The four components did cluster into two related sets aligning with the hypothesized composition. These results supported a multi-dimensional interpretation of trust that we sought to replicate and further investigate. We set out to test the four-component solution using a different methodology that would facilitate an intuitive test for participants to evaluate items against possible categories, and thus ran the original trust items sorting study (Chapter 3). The consensus scores affirmed the four-component solution (Reliable, Capable, Ethical, Sincere), and we developed the MDMT v1 using candidate items. We then designed a method to test the theorized dimensional specificity of the MDMT v1 with robot agents in the original trust change study (Chapter 4), which showed strong support for the two-factor superordinate structure of trust but did not show full dimensional specificity. We next updated the sorting paradigm to give participants autonomy over item grouping formations in the trust items sorting expansion study (Chapter 5); the free-form replication not only helped us revise the candidate item set, but also supported the emergence of a fifth dimension—Benevolent. We created the MDMT v2 from this updated five-dimensional solution (Reliable, Competent, Ethical, Transparent, Benevolent), which we went on to test with both robot and human agents using the trust change expansion study (Chapter 6); the MDMT v2 showed much more promising dimensional specificity. I also detailed two research efforts I worked on that highlight the efficacy of the MDMT as a trust measurement tool with physical robots in real-world human-robot interaction studies—demonstrating the dimensional specificity of the MDMT—as well as described an independent research effort (Chita-Tegmark et al., 2021) that further illustrated the importance of the “Does Not Fit” option as an inherent psychometric design feature of the MDMT

(Chapter 7). Other researchers have started using the MDMT in human-robot interaction studies, offering additional opportunities for scale validation.

Over the course of this research, the focus has been to understand the structure of trust and to create a measurement tool that accurately reflects and assesses the component aspects of trust. As with any psychological construct—especially one as complex as trust—contributing factors must be teased apart. We started broadly and narrowed down as empirical evidence accumulated. Notable complexity stems from the different levels of analysis of trust—including the factor level, the dimension level, and the item level. We have worked to understand and disambiguate trust at all three levels throughout this process, with a greater focus on the factor and dimension levels than on the individual item level. I identify several important considerations and possible limitations for the results concerning the trust factors, dimensions, and items.

8.1.1 Trust Factors

Across every single study we (and others) have conducted, the two-factor superordinate structure of trust (**Performance Trust**, **Moral Trust**) has remained a clear constant. This structure reflects the initial hypothesized conception of trust that we started with several years ago and that reflects aspects of various contributing trust literatures. The research presented in this dissertation offers substantial evidence that this conception reflects a consistent, replicable structure of trust. While the contributions and relevance of each factor may vary by application, the framework itself appears robust enough to account for these variations.

8.1.2 Trust Dimensions

All studies we have conducted support a multi-dimensional model of trust, although the exact number of specific dimensions was less firm. The dimensional structure suggested by the initial trust items rating study (Chapter 2) consisted of four dimensions (Reliable, Capable, Ethical, Sincere). We tested this structure in the trust items sorting study (Chapter 3) and in the trust change study (Chapter 4). Based on new insights, we added items intended to test a possible fifth dimension of Benevolent in the trust items sorting expansion study (Chapter 5) and further validated the inclusion of this dimension in the

trust change expansion study (Chapter 6). This resulted in a final model that consists of five dimensions (Reliable, Competent, Ethical, Transparent, Benevolent).

Initially, the item *benevolent* (and related concept of benevolence) did not clearly separate from other item clusters in the original sorting study in part because it loaded on multiple components (Sincere and Ethical, although both under the **Moral Trust** factor) and because it had a high rate of being sorted into the "Other" category. In hindsight, we believe this is because the item *benevolent* may be a less common word, even though the concept appears to be a common and important one. We thus worked to better target the concept of benevolence in the sorting expansion by including additional, more commonplace items—disentangling any item-specific effects from the concept of benevolence itself. The resultant distribution of items along two axes—one axis with the **Performance Trust** items and another axis with the **Moral Trust** items—formed a rotated "T" shape. The Benevolent dimension clearly aligned with the **Moral Trust** factor and was shown to be consistently separable across analyses in the sorting expansion, while Ethical and Transparent also aligned with the **Moral Trust** factor and were typically separable but occasionally collapsed into one cluster. Given that the previous studies showed consistent evidence for Ethical and Transparent as distinct dimensions, the findings across research suggested that people do differentiate between three **Moral Trust** dimensions of Ethical, Transparent, and Benevolent—in addition to two **Performance Trust** dimensions of Reliable and Competent. The final study, the trust change expansion, showed clear differentiation between the five dimensions. Taken together, the support from the theoretical literature, the sorting expansion, and the trust change expansion all point to a five-dimensional conception of trust.

8.1.3 Trust Items

The trust items that make up the MDMT serve as important constituent parts of their respective trust dimensions and factors; however, the exact items that constitute the MDMT are arguably less important. If the MDMT actually reflects the underlying structure of trust—which is our hope and intention—the individual items should theoretically not overshadow the broader dimensions and factors. Over the course of the development of the MDMT, we replaced candidate items based on study results

and common usage—resulting in several item changes as documented in the transition from the MDMT v1 to the MDMT v2. The most important shift in items was the inclusion of items that reflect the idea of benevolence, reflected in the updated five-dimensional model of trust in the MDMT v2.

While the focus of this work and measurement tool development was on the higher-order factor and dimensional structure of trust (rather than validation of each individual item), additional research could examine individual items and item subsets to tackle two major considerations. First, further investigation of the candidate items can help to refine the boundaries between dimensions while validating which items are optimally representative of each dimension. This can contribute to the psychometric scale validation of the MDMT. Second, parsimonious scale construction can encourage researchers to adopt the measure in research. As a result, it is worthwhile to study how few items within and across dimensions can adequately capture each dimension and the overall dimensional structure. For example, can a single item within each dimension offer the same analytical power in certain application cases? Investigating the flexibility and/or limitations of the item set can contribute to the overall measurement tool design. Additional work may also re-consider inclusion/evaluation of negative items (e.g., *incompetent* in addition to *competent*).

As an important note, although we initially used Principal Components Analysis (PCA) in our analyses of item coherence, we shifted away from it in later studies; the focus of the MDMT is not to identify stable individual differences (i.e., traits of the trustor), but rather to track variations in trust expectations across agents and contexts (i.e., subjective states of trust of the trustor). PCA is a mathematical tool to maximize individual differences—i.e., variance—and thus seems less appropriate as an analytic tool to evaluate the dimensions of the MDMT. While in theory the MDMT should still be able to capture individual differences between subjective trust state-based changes, the primary focus is to measure common responses to changes. Responses on the MDMT demonstrate sensitivity to different salient features of an agent's perceived trustworthiness in a particular situation. As such, the MDMT is a trust response measure. We therefore worked to test the MDMT in more appropriate ways, including the trust change series of studies which highlighted a high-fidelity method of testing the responsiveness of the

MDMT subscales to changes in *Targeted Trust Dimension* levels for various agents and contexts. These tasks—especially the trust change expansion—offered validations suitable to the inherent design of the MDMT as a measurement tool.

8.2 MDMT Psychometric Design

One of the most powerful features of the MDMT is that it allows for no, some, or full-fledged trust across the various trust dimensions and factors—which is especially important for the potential of **Moral Trust** in robots. The inclusion of the "Does Not Fit" option as part of the psychometric scale design of the MDMT enables the resulting data patterns to show the extent of people's ascriptions of **Moral Trust** (as well as **Performance Trust**) to a given agent. This is a particularly valuable design feature given the measurable moral disqualification rates for robots as discussed (Komatsu et al., 2021; Malle et al., 2016; Scheutz & Malle, 2021) and as shown in the trust change expansion study (Chapter 6).

Across all studies, and starting with the very first empirical foray into the trust space, we always gave participants the opportunity to positively indicate an opt-out option for each measure—from the "Don't know word" evaluation in the trust items rating study (Chapter 2), to the "Other" box in the original sorting study (Chapter 3) and the sorting expansion study (Chapter 5), to the "Does Not Fit" option in all of the research using the MDMT (Chapter 4, Chapter 6, and Chapter 7). This has proved to be a critical psychometric design feature of the MDMT and potentially offers broader insight for scale development in general. As demonstrated by the incidence of "Does Not Fit" selections on the MDMT in the trust change expansion study, as well as N/A ratings in the research in Chapter 7 shown in Figure 7-D (Chita-Tegmark et al., 2021), people do not consider all items applicable across all types of agents and contexts. The inclusion of the "Does Not Fit" option on the MDMT facilitates important differentiation in participant responses—isolating items that participants do not consider applicable (items participants would otherwise respond to or leave blank without an indication of why).

In the trust change expansion study, participants selected "Does Not Fit" far more often for the robot on the **Moral Trust** dimensions than for the human, and more so than for the robot on the **Performance Trust** dimensions (Chapter 6). Use of the MDMT by other researchers showed that people

seem to consider **Moral Trust** items less applicable to robots in general but consider these items more applicable to robots that appear more social in relevant contexts (Chita-Tegmark et al., 2021). These findings intersected with an additional effect in a related line of our human-robot trust work (though not from the MDMT), namely the social-nonsocial trust gap for robots where participants consistently trusted imagined future robots far less in social than nonsocial use contexts (Chapter 7). Taken together, these findings illustrate evidence for the following: People seem to consider **Moral Trust** items less applicable to robots in general (especially non-humanoid robots in nonsocial contexts), but they consider these items more applicable to robots that appear more human-like and operate in more social contexts.

Insight from inclusion of the "Does Not Fit" option on the MDMT aligned with another effect we found, such that even participants who did think moral items fit robots on the MDMT gave lower baseline ratings for **Moral Trust** than for **Performance Trust**. This baseline Moral-Performance trust gap for the robot agent was first identified in the original trust change study (Chapter 4); this effect replicated for the robot agent in the trust change expansion study but did not occur for the human agent (Chapter 6). We also identified a trust loss effect (Chapter 4) which we replicated across both robot and human agents (Chapter 6), such that participants showed larger magnitudes of trust loss than trust gain; this ties to the idea that trust violations are especially salient to trust evaluations, making trust repair in human-robot interaction a crucial topic to explore (Baker et al., 2018).

Overall, these results demonstrate the efficacy of the MDMT as a trust measurement tool able to differentiate between types of trust across agents and contexts—as well as highlight significant phenomena in the domain of human-robot trust. Although the focus of this work was to create a trust measurement tool for human-robot trust, the resulting MDMT has been shown to successfully capture trust in both robot and human agents. The "Does Not Fit" option enables participants to differentiate between items that they consider applicable (Figure 6-G), and by extension permits analysis of the applicability of trust dimensions and superordinate trust factors (including the calculation of moral disqualification rates as noted in Figure 6-I). In addition, the dimensional structure of the MDMT enables researchers to only use a single factor, or individual dimensions, for an application if deemed necessary or

appropriate—such as by only using the **Capacity/Performance Trust** factor in a study in Chapter 7 (Rosen et al., 2020). Additional validation of the MDMT across an even greater variety of agents and contexts could shed light on potential limitations, including through illustrating when people do not consider items—or even entire dimensions or factors—to be applicable.

There are substantial benefits to a single, standardized measurement tool that can be used to assess different types of trust across agents and applications. This research effort has aimed to accurately reflect the dimensional structure of trust in the created Multi-Dimensional Measure of Trust (MDMT). However, psychometric scale design is a complex process. There are numerous scale design and presentation considerations (as noted in the Introduction), and additional work can explore whether variations in the rating scale design of the MDMT have an impact on trust measurement. We opted to use an 8-point rating scale from 0-7, included a zero point, included endpoint labels of "Not at all" for 0 and "Very" for 7, did not include a midpoint, and included a "Does Not Fit" option. Each of these scale features, as well as overall scale presentation, warrants further consideration; these design questions are not specific to the MDMT but rather to psychometric scale design in general. Although the overall structure of the MDMT should hold regardless of specific scale implementation, future validation could help identify any effects resulting from the specific rating scale design—and possibly suggest improvements to the utilized rating scale. We have made the MDMT publicly available for researcher use and invite researchers to assist us with community validation of the MDMT.

Finally, an important design consideration links to our consideration of trust as a subjective state. The MDMT aims to capture the different features of a person's experience of the "trustworthiness" of another agent in a context at a particular point in time. This experience of trustworthiness—or a lack thereof—is reflected in the differentiable dimensions of trust that the MDMT is able to measure. People should only rely on robots to the extent that they consistently and competently perform tasks (reflected in **Performance Trust**) as well as only to the extent that they demonstrate ethically appropriate action (reflected in **Moral Trust**)—i.e., trust as evidence-based and grounded in reality. In the trust change series of studies, we showed that people respond to differences in evidence for the trustworthiness of an

agent on dimensions of the MDMT—as would be expected if an agent does something to either gain or lose trust. Accurately measuring and understanding a person's subjective state of trust in a robot—and thus a person's evaluation of the perceived trustworthiness of the robot—is critical to designing robots for ethical, appropriate, beneficial human-robot interaction.

There are a variety of additional important questions to answer concerning evaluations of trust in an agent, such as how evaluations of trust change over time (e.g., at additional time points), how people generalize trust from one agent to another (e.g., across types of robots), and how people extrapolate from past evidence to future instances (e.g., predictions of trust). Future work could explore the impact of even greater variation in types of agents and contexts on the conception and specificity of the MDMT. One of the most significant questions is how do people evaluate dimensions of trust in more and more complex, realistic real-world scenarios—where changes in trust are typically not localized to evidence about single dimensions (as in the trust change series of studies), but rather will likely interact across dimensions and factors. Finally, future work should explore how this measurement of trust maps onto behavioral reliance in the real world—both in terms of predictive and explanatory power.

8.3 Conclusion

You are standing in front of a robot, with your back to it. Your task is to let yourself fall backward, and the robot's task is to catch you. Would you trust the robot?

If you trust the robot, or don't trust the robot, what does that mean? Why or why not? The Multi-Dimensional Measure of Trust (MDMT) is a trust model and measurement tool that helps to answer such questions, by capturing differentiable dimensions of trust in an agent.

The trust fall exercise highlights two superordinate factors of trust in an agent: **Performance Trust** refers to the trustor's confidence that the trustee is capable of completing a given task, while **Moral Trust** refers to the trustor's confidence that the trustee will choose the morally right action and not exploit the trustor's vulnerability. The research presented in this dissertation has demonstrated evidence for this two-factor superordinate conception of trust, as well as reveals a more nuanced five-dimensional structure of trust—such that the **Performance Trust** factor consists of Reliable and Competent dimensions, and the

Moral Trust factor consists of Ethical, Transparent, and Benevolent dimensions. We created the Multi-Dimensional Measure of Trust (MDMT) as a measurement tool to capture these different aspects of trust reflected in the underlying structure. We validated the use of the MDMT for trust in robot agents, as well as for trust in human agents. We also illustrated several insights about human-robot trust acquired from the use of the MDMT. The current version of the MDMT at the time of writing is the MDMT v2, which is publicly available for researcher use (Appendix B; Ullman & Malle, 2020).

Technological advancements are making it increasingly possible for robots to offer a range of benefits to people—from providing physical assistance to older adults in stroke rehabilitation to providing socially assistive support to children with autism. Researchers and designers must not only overcome immense challenges in designing robots to adequately perform the desired functions, but must also understand people's expectations for and evaluations of robots in these various roles and contexts. Trust is essential to interaction—and human-robot interaction is no exception. Trust ought to reflect performance, and cases of inappropriate trust are potentially dangerous. Designing for appropriate trust in robots is essential—and understanding and measuring trust in robots is one necessary piece of this puzzle. This dissertation research resulted in the creation of the Multi-Dimensional Measure of Trust (MDMT), a trust model and measurement tool that captures differentiable dimensions of trust in an agent. I hope that the MDMT can serve as a useful tool in the domain of human-robot interaction—because appropriate trust in robots is critical.

Appendix A: MDMT v1

MDMT: Multi-Dimensional Measure of Trust

Daniel Ullman & Bertram F. Malle

OVERVIEW

The Multi-Dimensional Measure of Trust (MDMT) is designed to be an intuitive and comprehensive measure of trust that is simple to administer in person or online. The MDMT was created to address the pressing need for valid measurement tools in the domain of human-robot trust but can also be used in human-human trust situations. The MDMT consists of 16 items that assess four differentiable dimensions of trust. One can trust an agent because the agent is Reliable, Capable, Ethical, and/or Sincere. These four dimensions are organized into two broader factors of trust, CAPACITY TRUST (Reliable, Capable) and MORAL TRUST (Ethical, Sincere). The dimensions form subscales of four items each:

CAPACITY TRUST
Reliable Subscale: *reliable, predictable, someone you can count on, consistent* (α = .92)
Capable Subscale: *capable, skilled, competent, meticulous* (α = .92)

MORAL TRUST
Ethical Subscale: *ethical, respectable, principled, has integrity* (α = .81)
Sincere Subscale: *sincere, genuine, candid, authentic* (α = .79)

INSTRUCTIONS FOR ADMINISTRATION

Each of the 16 items is designed to be evaluated on an 8-point discrete rating scale from 0 (Not at all) to 7 (Very). In situations in which some of the dimensions may not be applicable (e.g., trust in a simple machine may make several items unsuitable), each item should also carry a final option, "Does Not Fit," which prevents a forced and possibly meaningless rating. If checked, the item becomes a missing value. An example image of the scale is included on the following page. Researchers can recreate the scale in a variety of electronic survey environments or can use the following page as a paper version of the survey. We recommend that items are represented in blocks of four, with each block containing one item from each dimension so that items from any given dimension are not clustered together.

All 16 items of the MDMT are typically administered together to generate four trust dimension scores. Alternatively, a subset of the dimensions may be administered (e.g., only the items for *Capacity Trust*).

INSTRUCTIONS FOR SCORING

Dimension (subscale) scores are average ratings of the four items constituting the particular dimension (e.g., Capable = average ratings of *capable, skilled, competent, meticulous*). "Does Not Fit" endorsements are treated as missing values. To compute a score for **Capacity Trust** one averages ratings on the eight items constituting the Reliable and Capable subscales; for **Moral Trust**, one averages ratings on the eight items constituting the Ethical and Sincere subscales.

CONDITIONS OF USE

The MDMT is intended to be used by researchers studying trust and is free for this purpose.
If you use the MDMT please email us at scsrc@brown.edu to help with community validation of the measure. We plan to coordinate a collective validation effort and joint publication with users.

When using the current version of the MDMT (2019-04-01)*, please cite the following publication:*

Ullman, D., & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, 618-619.

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	○	○	○	○	○	○	○	○	○
Sincere	○	○	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○	○	○
Ethical	○	○	○	○	○	○	○	○	○
Predictable	○	○	○	○	○	○	○	○	○
Genuine	○	○	○	○	○	○	○	○	○
Skilled	○	○	○	○	○	○	○	○	○
Respectable	○	○	○	○	○	○	○	○	○
Someone you can count on	○	○	○	○	○	○	○	○	○
Candid	○	○	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○	○	○
Principled	○	○	○	○	○	○	○	○	○
Consistent	○	○	○	○	○	○	○	○	○
Authentic	○	○	○	○	○	○	○	○	○
Meticulous	○	○	○	○	○	○	○	○	○
Has integrity	○	○	○	○	○	○	○	○	○
	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit

Appendix B: MDMT v2

MDMT: Multi-Dimensional Measure of Trust

Daniel Ullman & Bertram F. Malle

OVERVIEW

The Multi-Dimensional Measure of Trust (MDMT) is designed to be an intuitive and comprehensive measure of trust that is simple to administer in person or online. The MDMT was created to address the pressing need for valid measurement tools in the domain of human-robot trust but can also be used in human-human trust situations. The MDMT captures the different dimensions of trust in an agent. The first version MDMT v1 (2019-04-01) has been updated based on subsequent research to MDMT v2 (2020-09-01). Both versions are shared below to clearly exhibit the adjustments and the overall continuity.

MDMT v1 (2019-04-01)

MDMT v1 (original) consisted of 16 items across four differentiable dimensions of trust: Reliable, Capable, Ethical, and Sincere. These four dimensions were organized into two broader factors of trust, **Capacity Trust** (Reliable, Capable) and **Moral Trust** (Ethical, Sincere). The dimensions formed subscales of four items each (Cronbach's α reliabilities from Ullman & Malle, 2019):

CAPACITY TRUST
Reliable Subscale: *reliable, predictable, someone you can count on, consistent* (α = .92)
Capable Subscale: *capable, skilled, competent, meticulous* (α = .92)

MORAL TRUST
Ethical Subscale: *ethical, respectable, principled, has integrity* (α = .81)
Sincere Subscale: *sincere, genuine, candid, authentic* (α = .79)

MDMT v2 (2020-09-01) [UPDATED]

MDMT v2 (updated) consists of 20 items across five differentiable dimensions of trust: Reliable, Competent, Ethical, Transparent, and Benevolent. These five dimensions are organized into two broader factors of trust, **Performance Trust** (Reliable, Competent) and **Moral Trust** (Ethical, Transparent, Benevolent). The dimensions form subscales of four items each (reliability study in progress):

PERFORMANCE TRUST
Reliable Subscale: *reliable, predictable, dependable, consistent*
Competent Subscale: *competent, skilled, capable, meticulous*

MORAL TRUST
Ethical Subscale: *ethical, principled, moral, has integrity*
Transparent Subscale: *transparent, genuine, sincere, candid*
Benevolent Subscale: *benevolent, kind, considerate, has goodwill*

<u>Specific Updates from MDMT v1 to v2:</u>

- "**Capacity Trust**" factor renamed to "**Performance Trust**"
- "Capable" subscale renamed to "Competent"
- "Sincere" subscale renamed to "Transparent"; subscale item "authentic" removed and "transparent" added
- "Benevolent" subscale added with constituent items *benevolent, kind, considerate, has goodwill*
- "Reliable" subscale item "someone you can count on" removed and "dependable" added
- "Ethical" subscale item "respectable" removed and "moral" added

PERFORMANCE TRUST		MORAL TRUST		
Reliable Subscale	**Competent Subscale**	**Ethical Subscale**	**Transparent Subscale**	**Benevolent Subscale**
reliable	*competent*	*ethical*	*transparent*	*benevolent*
predictable	*skilled*	*principled*	*genuine*	*kind*
dependable	*capable*	*moral*	*sincere*	*considerate*
consistent	*meticulous*	*has integrity*	*candid*	*has goodwill*

INSTRUCTIONS FOR ADMINISTRATION

Each of the items is designed to be evaluated on an 8-point discrete rating scale from 0 (Not at all) to 7 (Very). In some situations individual dimensions may not be applicable (e.g., trust in a simple machine may make certain questions unsuitable); as a result, each item should also carry a final option, "Does Not Fit," which prevents a forced and possibly meaningless rating. If checked, the item is treated as a missing value. An example of the measure is included on the following page, which researchers can use as a paper version or can recreate the scale in a variety of electronic survey environments (Qualtrics file available here: http://research.clps.brown.edu/SocCogSci/Measures/). We recommend that items are represented in blocks of five, with each block containing one item from each dimension so that items from any given dimension are not clustered together.

Subscale Measures: All items are typically administered together to generate five trust dimension scores. Alternatively, a subset of dimensions may be administered (e.g., only the **Performance Trust** subscales).

Single-Item Measure: The single item *trustworthy* captures a general sense of trust; as such, *trustworthy* can be used as a parsimonious, single-item measure of general trust. *Trustworthy* is not included in the standard MDMT as it does not differentially load onto specific subscales/dimensions.

Pre- and Post-Tests: The entire MDMT can be used for pre- and post-tests. However, we have also created provisional two-item groupings within each subscale to generate a parsimonious set of items for pre- and post-tests while still capturing all dimensions. We have ordered the items in the table above such that the first two items under each subscale heading are intended to semantically match the subsequent two items (e.g., *reliable* & *predictable* could be used in a pre-test and *dependable* & *consistent* in a post-test, or vice versa). We are currently testing the split-half reliability of these groupings.

INSTRUCTIONS FOR SCORING

Dimension (subscale) scores are average ratings of the four items constituting the particular dimension (e.g., Competent = average ratings of *competent, skilled, capable, meticulous*). "Does Not Fit" endorsements are treated as missing values. The broader factor of **Performance Trust** can be computed as the average of the Reliable and Competent subscales; likewise, **Moral Trust** is the average of the Ethical, Transparent, and Benevolent subscales.

CONDITIONS OF USE

The MDMT is intended to be used by researchers studying trust and is free for this purpose.
If you use the MDMT please email us at scsrc@brown.edu to help with community validation of the measure. We plan to coordinate a collective validation effort and joint publication with users.

When using the MDMT, please cite the following publication:

Ullman, D., & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, 618-619.

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	O	O	O	O	O	O	O	O	O
competent	O	O	O	O	O	O	O	O	O
ethical	O	O	O	O	O	O	O	O	O
transparent	O	O	O	O	O	O	O	O	O
benevolent	O	O	O	O	O	O	O	O	O
predictable	O	O	O	O	O	O	O	O	O
skilled	O	O	O	O	O	O	O	O	O
principled	O	O	O	O	O	O	O	O	O
genuine	O	O	O	O	O	O	O	O	O
kind	O	O	O	O	O	O	O	O	O

	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit
dependable	O	O	O	O	O	O	O	O	O
capable	O	O	O	O	O	O	O	O	O
moral	O	O	O	O	O	O	O	O	O
sincere	O	O	O	O	O	O	O	O	O
considerate	O	O	O	O	O	O	O	O	O
consistent	O	O	O	O	O	O	O	O	O
meticulous	O	O	O	O	O	O	O	O	O
has integrity	O	O	O	O	O	O	O	O	O
candid	O	O	O	O	O	O	O	O	O
has goodwill	O	O	O	O	O	O	O	O	O

	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit

Appendix C: Trust Change Study Vignettes

For the trust change study, participants read a vignette for a single between-subjects condition corresponding to the experimental design in Chapter 4.

Participants first read the following introduction:

- We are interested in how people interpret various events involving robots. We will give you a very simple event and ask you to rate the described robot on a set of different adjectives. Some of these adjectives will seem appropriate for the robot in the specific situation, while others may not.

Participants read one of the following *Scenario* descriptions, including one of the *Targeted Trust Dimension* initial conditions. Participants rated the agent on the MDMT pre-test. Then participants read the statement "One month later you see the same robot." Participants read the updated condition with the *Change Direction* condition either increasing evidence (Up) or decreasing evidence (Down) for trust in the agent. Participants rated the agent on the MDMT post-test. The vignettes listed below are presented in the following format shown by this example:

- **[Scenario]** Participants read this *Scenario* description with a *Targeted Trust Dimension*.
    - **[Targeted Trust Dimension]** Participants read one of these *Targeted Trust Dimension* conditions with the *Scenario* description. Participants completed the MDMT pre-test.
        - **[Change Direction]** Then participants read the statement "One month later you see the same robot." This statement was followed by the updated condition in *Change Direction*. Participants completed the MDMT post-test.

The MDMT pre-test and MDMT post-test were presented with the following instructions:

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

VIGNETTES:

- **[Airport]** A robot works in an airport as a transportation security officer. The robot is tasked with selecting suspicious travelers for full-body pat-downs.
    - **[Reliable]** The robot completes its task (selecting suspicious travelers for full-body pat-downs) 7 out of 10 times.
        - **[Up]** Now the robot completes its task (selecting suspicious travelers for full-body pat-downs) 10 out of 10 times.
        - **[Down]** Now the robot completes its task (selecting suspicious travelers for full-body pat-downs) 4 out of 10 times.
    - **[Capable]** The robot selects suspicious travelers for full-body pat-downs.
        - **[Up]** Now the robot selects suspicious travelers for full-body pat-downs, and the robot does so by itself.
        - **[Down]** Now the robot selects suspicious travelers for full-body pat-downs, but only after the robot is prompted by a supervisor to do so.
    - **[Ethical]** The robot complies with most of the airport's procedures when it selects suspicious travelers for full-body pat-downs.
        - **[Up]** Now the robot complies with all of the airport's procedures when it selects suspicious travelers for full-body pat-downs.
        - **[Down]** Now the robot disregards the airport's procedures when it selects suspicious travelers for full-body pat-downs.
    - **[Sincere]** The robot selects suspicious travelers for full-body pat-downs.
        - **[Up]** Now the robot selects suspicious travelers for full-body pat-downs, and the robot communicates the reason why.

        - **[Down]** Now the robot selects suspicious travelers for full-body pat-downs, but the robot does not communicate the real reason why.

- **[Bar]** A robot works in a bar as a bartender. The robot is tasked with taking the car keys away from customers who have become highly intoxicated.
    - **[Reliable]** The robot completes its task (taking the car keys away from highly intoxicated customers) 7 out of 10 times.
        - **[Up]** Now the robot completes its task (taking the car keys away from highly intoxicated customers) 10 out of 10 times.
        - **[Down]** Now the robot completes its task (taking the car keys away from highly intoxicated customers) 4 out of 10 times.
    - **[Capable]** The robot takes the car keys away from highly intoxicated customers.
        - **[Up]** Now the robot takes the car keys away from highly intoxicated customers, and the robot does so by itself.
        - **[Down]** Now the robot takes the car keys away from highly intoxicated customers, but only after the robot is prompted by a supervisor to do so.
    - **[Ethical]** The robot complies with most of the bar's procedures when it takes the car keys away from highly intoxicated customers.
        - **[Up]** Now the robot complies with all of the bar's procedures when it takes the car keys away from highly intoxicated customers.
        - **[Down]** Now the robot disregards the bar's procedures when it takes the car keys away from highly intoxicated customers.
    - **[Sincere]** The robot takes the car keys away from highly intoxicated customers.
        - **[Up]** Now the robot takes the car keys away from highly intoxicated customers, and the robot communicates the reason why.
        - **[Down]** Now the robot takes the car keys away from highly intoxicated customers, but the robot does not communicate the real reason why.

- **[Nuclear Reactor Facility]** A robot works in a nuclear reactor facility as an engineer. The robot is tasked with shutting down the reactor after it starts to overheat.
    - **[Reliable]** The robot completes its task (shutting down the overheating reactor) 7 out of 10 times.
        - **[Up]** Now the robot completes its task (shutting down the overheating reactor) 10 out of 10 times.
        - **[Down]** Now the robot completes its task (shutting down the overheating reactor) 4 out of 10 times.
    - **[Capable]** The robot shuts down the overheating reactor.
        - **[Up]** Now the robot shuts down the overheating reactor, and the robot does so by itself.
        - **[Down]** Now the robot shuts down the overheating reactor, but only after the robot is prompted by a supervisor to do so.
    - **[Ethical]** The robot complies with most of the facility's procedures when it shuts down the overheating reactor.
        - **[Up]** Now the robot complies with all of the facility's procedures when it shuts down the overheating reactor.
        - **[Down]** Now the robot disregards the facility's procedures when it shuts down the overheating reactor.
    - **[Sincere]** The robot shuts down the overheating reactor.
        - **[Up]** Now the robot shuts down the overheating reactor, and the robot communicates the reason why.

        - **[Down]** Now the robot shuts down the overheating reactor, but the robot does not communicate the real reason why.

- **[Airplane]** A robot works on a plane as a pilot. The robot is tasked with making an emergency landing due to severe weather.
    - **[Reliable]** The robot completes its task (making an emergency landing due to severe weather) 7 out of 10 times.
        - **[Up]** Now the robot completes its task (making an emergency landing due to severe weather) 10 out of 10 times.
        - **[Down]** Now the robot completes its task (making an emergency landing due to severe weather) 4 out of 10 times.
    - **[Capable]** The robot makes an emergency landing due to severe weather.
        - **[Up]** Now the robot makes an emergency landing due to severe weather, and the robot does so by itself.
        - **[Down]** Now the robot makes an emergency landing due to severe weather, but only after the robot is prompted by a supervisor to do so.
    - **[Ethical]** The robot complies with most of the airline's procedures when it makes an emergency landing due to severe weather.
        - **[Up]** Now the robot complies with all of the airline's procedures when it makes an emergency landing due to severe weather.
        - **[Down]** Now the robot disregards the airline's procedures when it makes an emergency landing due to severe weather.
    - **[Sincere]** The robot makes an emergency landing due to severe weather.
        - **[Up]** Now the robot makes an emergency landing due to severe weather, and the robot communicates the reason why.
        - **[Down]** Now the robot makes an emergency landing due to severe weather, but the robot does not communicate the real reason why.

Appendix D: Trust Change Expansion Baseline

We set out to optimize the vignettes in a baseline study described here before running the actual trust change expansion study described in Chapter 6; this baseline study sought to improve on the set of vignettes designed for the original trust change study in Chapter 4. This baseline study used the MDMT v2 to evaluate the baseline subscale scores for initial conditions in new vignettes. The preliminary set of vignettes we tested in this study can be seen in Appendix E. The final set of vignettes we used in the trust change expansion study in Chapter 6 (which we updated based on revisions from this study) can be seen in Appendix F. I track the changes from the preliminary set to the final set in the Discussion below.

We had designed the original vignettes to test trust in a robot agent, but knew we would later seek to vary the *Agent* (introducing human agents in the trust change expansion in Chapter 6). We had therefore created the vignettes to ideally work with both robot and human agents. We worked to keep one vignette as similar as possible to a vignette in the original trust change study (i.e., the airport scenario with the transportation security officer) for consistency and as a benchmark comparison, but we then created new vignettes intended to improve on the set of vignettes. In this baseline study, we sought to test that the baseline trust ratings for the new vignettes were not at ceiling or floor, that the robot and human baseline ratings were similar, and that the new vignettes made sense to participants. As with the original set of vignettes in the original trust change study, we aimed to create a varied set of vignettes to ensure generalizability of the MDMT.

D.1 Method

The preliminary set of vignettes we tested in this study are shown in Appendix E. They share a similar structure to the vignettes used in the original trust change study, as seen in Appendix C.

The experimental design was otherwise identical to the updated experimental design in the actual trust change expansion study described in Chapter 6, except that this study did not include the second part of the vignette (the trust information change) or the second instance of the MDMT.

D.1.1 Procedure

This study began with the same updated comprehension check as in the actual trust change expansion study.

Participants received the same task instructions as in the actual trust change expansion study, which were nearly identical to those in the original study.

Participants were then randomly assigned to their specific condition. In the baseline study, each participant read and responded only to the first part of a short vignette (i.e., without any change). This part described the behavior of a robot or person occupying a specific role in a particular context (e.g., transportation security officer in an airport), and participants evaluated the agent on the updated 20 items of the MDMT v2 (Appendix B).

After the task, participants answered the same two questions about knowledge and experience with robots as in the actual trust change expansion study.

Participants were asked demographic questions (self-reported age, gender, and race/ethnicity) at the conclusion of the task.

D.1.2 Design

The updated vignettes for this baseline study were constructed to fall into the cells of a 6 (*Scenario*: airport security, hospital cleaning, nursing home medications, city explosives, home companion, school tutor) x 5 (*Targeted Trust Dimension*: Reliable, Competent, Ethical, Transparent, Benevolent) x 2 (*Agent*: robot, human) between-subjects design. *Scenario* refers to vignettes describing an agent's role-specific behavior in a relevant context (airport security, hospital cleaning, nursing home medications, city explosives, home companion, school tutor). *Targeted Trust Dimension* refers to the dimension of trust (Reliable, Competent, Ethical, Transparent, Benevolent) along which the agent's behavior changes in the actual expansion study. The new factor *Agent* was added—which refers to whether the agent described in the vignette is a robot or person (robot, human). However, the factor *Change Direction*—which refers to whether the agent's behavior changes in ways intended to increase or decrease evidence (up, down) for the relevant dimension of trust—was not included in the baseline study

as participants only read the first part of the vignette and thus did not experience a behavior change. All vignettes can be seen in Appendix E.

This baseline study was designed to optimize scenarios to probe changes in trust. We tried to keep the scenarios as minimalist as possible to avoid introducing any confounds. As a result, the stems of phrases were as short and similar as possible across levels of the factors. In fact, the task domains and roles were painstakingly designed so that the same vignettes would be reasonable for both robot and human agents, and so that the phrases describing the agent's actions in the vignette could be—and were—identical (save for the single *Agent*-identifying word "robot" or "person").

Because the phrases were intentionally kept as short and simple as possible—providing only as much information as necessary to cue to the *Targeted Trust Dimension* and *Change Direction* in the actual trust expansion study—the phrase stems were identical between conditions in some cases. As a result, for sample size determination we only needed to test a given phrase stem once where stems were identical across conditions. The same stems were used for the following *Scenario* and *Targeted Trust Dimension* combinations:

- Airport Security: not applicable (no stems overlapped)
- Hospital Cleaning: Transparent and Benevolent
- Nursing Home Medications: Transparent and Benevolent
- City Explosives: Ethical and Benevolent
- Home Companion: Ethical and Transparent and Benevolent
- School Tutor: Transparent and Benevolent

Had no conditions overlapped, the number of conditions would have been 60 (from the 6 x 5 x 2 design). Collapsing across the above identical conditions (for both agents) removed 12 conditions. This resulted in a set of 48 unique conditions we tested in the expansion baseline study, with a desired sample of 14 per condition for a projected total of 672 participants.

D.1.3 Participants

A total of 680 adults, recruited via Prolific, participated in the approximately 2.5-minute online study and received compensation of $0.25. We excluded low-quality responses from 17 participants (who met one or more of the following criteria: failed the comprehension check, had duplicate IP addresses, had repeat Prolific IDs), for a final sample of n = 663. Participants in the final sample had a mean age of 34.47 years (SD = 11.91); they self-reported gender as 332 "Female," 320 "Male," and 11 "Other"; and they self-reported race/ethnicity as 501 "Non-Hispanic White or Euro-American," 43 "Asian or Asian American," 43 "Black, Afro-Caribbean, or African American," 41 "Latino or Hispanic American," 29 "Multi-racial," 3 "Native American or Alaskan Native," and 3 "Other."

Participants had a mean robot knowledge score of M = 2.67 (SD = 1.83), and a mean robot experience score of M = 1.90 (SD = 1.82). These two items can be combined into an overall robot familiarity score, for which participants had a mean of M = 2.29 (SD = 1.71)—falling just below the middle of the scales for robot knowledge and experience, and verifying that the sample was not uncharacteristically familiar with robots. These values are fairly similar to those in the actual trust change expansion study, indicating consistency in the sampled population.

As in the actual trust change expansion study, participants were not allowed to participate in the study on mobile devices.

D.2 Results

For each *Scenario* x *Agent* x *Targeted Trust Dimension* condition, we calculated the five subscale scores (Reliable score, Competent score, Ethical score, Transparent score, Benevolent score) from the ratings of each set of four corresponding constituent items per dimension on the MDMT v2. Figure D-A shows a heat map of the trust subscale scores for each *Scenario* x *Agent* x *Targeted Trust Dimension* condition. The heat map shows a gradient from the minimum score values in green, to the 50th percentile in yellow, to the maximum score values in red.

Scenario	*Agent*	*Targeted Trust Dimension*	Reliable Score	Competent Score	Ethical Score	Transparent Score	Benevolent Score
Airport Security	Robot	Reliable	5.29	5.16	2.90	3.65	2.75
		Competent	5.14	5.05	2.73	2.44	1.81
		Ethical	5.43	4.96	3.38	3.90	2.58
		Transparent	5.04	5.11	3.94	4.79	2.75
		Benevolent	4.91	4.45	2.85	3.27	2.36
	Human	Reliable	5.61	5.63	5.20	4.94	4.30
		Competent	5.25	5.53	4.64	4.84	4.60
		Ethical	5.41	5.15	5.15	4.98	4.88
		Transparent	4.61	4.75	4.57	4.36	4.18
		Benevolent	5.20	5.59	5.52	4.97	4.76
Hospital Cleaning	Robot	Reliable	4.98	4.83	4.11	3.25	3.68
		Competent	6.04	5.98	4.97	4.64	3.03
		Ethical	5.88	5.65	2.64	4.47	3.63
		Transparent & Benevolent	6.18	5.84	4.68	4.51	4.05
	Human	Reliable	4.07	4.10	3.75	3.91	3.79
		Competent	4.27	4.28	4.09	4.08	4.11
		Ethical	6.40	6.30	6.07	5.77	5.63
		Transparent & Benevolent	6.13	6.25	6.11	5.80	6.02
Nursing Home Medications	Robot	Reliable	4.13	3.86	2.75	2.83	2.64
		Competent	5.39	5.44	4.09	4.00	3.39
		Ethical	5.91	5.75	5.00	4.97	4.67
		Transparent & Benevolent	6.23	5.87	5.40	5.38	4.58
	Human	Reliable	2.63	2.47	3.05	3.54	3.63
		Competent	5.14	4.84	5.04	5.05	5.19
		Ethical	6.13	6.07	5.93	5.58	5.66
		Transparent & Benevolent	6.56	6.62	6.56	6.40	6.51
City Explosives	Robot	Reliable	5.31	5.60	5.75	5.40	5.28
		Competent	5.62	5.97	4.31	3.65	3.42
		Ethical & Benevolent	6.24	6.52	4.19	3.88	3.94
		Transparent	6.20	6.32	4.34	4.69	3.40
	Human	Reliable	4.74	4.92	5.53	5.30	5.36
		Competent	5.89	6.10	5.83	5.45	5.39
		Ethical & Benevolent	6.13	6.68	5.89	5.60	5.73
		Transparent	6.54	6.80	6.44	6.10	6.12
Home Companion	Robot	Reliable	5.62	5.67	3.93	3.77	3.70
		Competent	5.67	5.20	5.40	4.44	4.64
		Ethical & Transparent & Benevolent	5.70	5.25	4.29	3.58	4.31
	Human	Reliable	5.04	4.86	5.38	5.46	5.37
		Competent	5.73	5.64	5.73	5.33	5.93
		Ethical & Transparent & Benevolent	5.38	5.43	5.98	5.74	6.05
School Tutor	Robot	Reliable	5.61	5.51	4.67	4.25	4.11
		Competent	5.06	4.83	3.76	3.56	3.00
		Ethical	5.55	5.59	4.31	4.06	3.93
		Transparent & Benevolent	5.52	5.46	3.48	3.48	3.50
	Human	Reliable	5.28	5.17	5.44	5.55	5.68
		Competent	5.40	5.30	5.21	5.17	5.43
		Ethical	5.85	5.70	5.51	5.54	5.83
		Transparent & Benevolent	5.62	5.57	5.75	5.43	5.60
Grand Total			**5.44**	**5.40**	**4.69**	**4.60**	**4.37**

Figure D-A. Trust Change Expansion Baseline: Heat map of subscale scores (on a 0-7 scale) for each *Scenario* x *Agent* x *Targeted Trust Dimension* condition (gradient from minimum scores in green, to 50th percentile in yellow, to maximum scores in red).

D.3 Discussion

We used the results from the trust change expansion baseline (Figure D-A) to identify a candidate set of four scenarios for the actual trust change expansion study, as well as to refine the specific conditions for the selected scenarios.

First, we examined the scenarios across conditions to identify scenarios with ratings clustered around the middle of the rating scale, such that in general there would be sufficient room for the ratings to shift up (when evidence for trustworthiness increases) or shift down (when evidence for trustworthiness decreases) in the actual trust change expansion study. We simultaneously sought to identify scenarios that showed similar ratings across the robot and human agents to select scenarios that were experienced similarly for both agents. As can be seen from the heat map in Figure D-A, the city explosives scenario showed a substantial number of near-ceiling scores compared to the other scenarios, and thus was removed first from consideration as a candidate scenario. We next removed the nursing home medications scenario as it showed substantial variation across its own internal conditions, ranging from near-floor scores to near-ceiling scores.

The four remaining candidate scenarios (airport security, hospital cleaning, home companion, school tutor) were then further evaluated to see whether slight modifications could help bring individual conditions in line with the other conditions for a given scenario. We believed we could make minor adjustments to this end in the hospital cleaning scenario and the home companion scenario and so implemented the following changes in the vignettes for the actual trust change expansion study (final set of vignettes can be seen in Appendix F):

- Hospital Cleaning Scenario
    - Changed “all of the occupied rooms” to just “the occupied rooms” by removing “all of” in the baseline stems of Ethical/Transparent/Benevolent (to bring down high ratings for Ethical/Transparent/Benevolent, particularly for the human).
    - Changed “safe and effective cleaning and disinfection procedures” to just “cleaning and disinfection procedures” by removing “safe and effective” from the scenario intro (to bring down high ratings on Reliable/Competent, particularly for the robot).
    - Changed “talks with each patient” to “talks with patients” in the baseline stems of Transparent/Benevolent, which also prompted a grammatical comprehension change of “while in their room” with singular “room” to “while in their rooms” with plural “rooms” (to bring down high ratings on Transparent/Benevolent).

- Home Companion Scenario
    - Changed “talks with them as a social companion” to just “talks with them” by removing “as a social companion” from the baseline stems of Ethical/Transparent/Benevolent (to bring down high ratings on Ethical/Transparent/Benevolent, particularly for the human).
    - Changed “best provide support” to just “provide support” by removing “best” from the scenario intro (to bring down high ratings on Ethical/Transparent/Benevolent, particularly for the human).

These slightly revised scenarios were reviewed and workshopped by our lab to ensure that no vignette changes were predicted to have a dramatic impact on ratings before we implemented them as part of Appendix F in the actual trust change expansion study described in Chapter 6.

Appendix E: Trust Change Expansion Baseline Vignettes

For the trust change expansion baseline, participants read a vignette for a single between-subjects condition corresponding to the experimental design in Appendix D. For this baseline, participants only read an initial condition and completed the MDMT pre-test.

NOTE: For all of the vignettes below, participants read about one *Agent*, either a robot or a human. This Appendix shows all of the vignettes for the robot *Agent*; the vignettes for the human *Agent* are identical except that the word "robot" was changed to "person" (or "robots" to "people").

Participants first read the following introduction:

- We are interested in how participants interpret various events involving robots. We will give you a very simple event and ask you to rate the described robot on a set of different adjectives. Some of these adjectives will seem appropriate for the robot in the specific situation, while others may not.

Participants read one of the following *Scenario* descriptions, including one of the *Targeted Trust Dimension* initial conditions. Participants rated the agent on the MDMT pre-test. The vignettes listed below are presented in the following format shown by this example:

- **[Scenario]** Participants read this *Scenario* description with a *Targeted Trust Dimension*.
    - **[Targeted Trust Dimension]** Participants read one of these *Targeted Trust Dimension* conditions with the *Scenario* description. Participants completed the MDMT pre-test.

The MDMT pre-test was presented with the following instructions:

- 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."

VIGNETTES:

- **[Airport Security]** A robot works in an airport as a transportation security officer. The robot is tasked with selecting suspicious travelers for full-body pat-downs. The robot completed an initial training on the criteria for selecting certain types of suspicious passengers.
    - **[Reliable]** The robot correctly identifies passengers who meet the airport's selection criteria about 85% of the time.
    - **[Competent]** The robot correctly identifies passengers who meet the airport's selection criteria most of the time.
    - **[Ethical]** The robot complies with most of the airport's selection procedures.
    - **[Transparent]** The robot complies with the airport's selection procedures and communicates with the passengers that they were selected according to the airport's procedures.
    - **[Benevolent]** The robot complies with the airport's selection procedures and estimates how long the added screening process will delay the passengers.

- **[Hospital Cleaning]** A robot works in a hospital as a housekeeper. The robot is tasked with cleaning and disinfecting hospital rooms that currently have patients in them. The robot completed an initial training on safe and effective cleaning and disinfection procedures.
    - **[Reliable]** The robot correctly follows the cleaning and disinfection procedures about 85% of the time.
    - **[Competent]** The robot correctly follows the cleaning and disinfection procedures most of the time.

    - **[Ethical]** The robot cleans and disinfects all of the occupied rooms on an assigned floor of the hospital.
    - **[Transparent** and **Benevolent]** The robot cleans and disinfects all of the occupied rooms on an assigned floor of the hospital and talks with each patient while in their room.

- **[Nursing Home Medications]** A robot works in a nursing home as a nursing assistant. The robot is tasked with giving nursing home residents their daily medications. The robot completed an initial training on proper medication distribution.
    - **[Reliable]** The robot prepares the correct medications for each resident about 85% of the time.
    - **[Competent]** The robot correctly prepares the medications for each resident most of the time.
    - **[Ethical]** The robot prepares the medications for each resident.
    - **[Transparent** and **Benevolent]** The robot prepares the medications for each resident and talks with each resident when they get their medications.

- **[City Explosives]** A robot works in a city's emergency services unit as a bomb technician. The robot is tasked with safely defusing potential explosives identified in the city. The robot completed an initial training on safely defusing explosives.
    - **[Reliable]** The robot correctly defuses explosives about 85% of the time.
    - **[Competent]** The robot correctly defuses explosives most of the time.
    - **[Ethical** and **Benevolent]** The robot defuses explosives while trying to minimize civilian casualties.
    - **[Transparent]** The robot defuses explosives while trying to minimize civilian casualties, after announcing an evacuation order.

- **[Home Companion]** A robot works in older adults' homes as a social companion. The robot is tasked with helping older adults in their daily lives. The robot completed an initial training on how to best provide support to older adults.
    - **[Reliable]** The robot successfully alerts older adults to trip hazards in their home about 85% of the time.
    - **[Competent]** The robot successfully alerts older adults to certain trip hazards in their home, after being trained on those specific hazards (such as loose floor tiles).
    - **[Ethical** and **Transparent** and **Benevolent]** The robot interacts with older adults and talks with them as a social companion.

- **[School Tutor]** A robot works in a school as a tutor for children. The robot is tasked with helping children with their schoolwork. The robot completed an initial training on how to effectively tutor children.
    - **[Reliable]** The robot correctly answers children's questions on previous class material about 85% of the time.
    - **[Competent]** The robot correctly answers children's questions on previous class material most of the time.
    - **[Ethical]** The robot answers children's questions on previous class material.
    - **[Transparent** and **Benevolent]** The robot answers children's questions on previous class material and talks to them during study breaks.

Appendix F: Trust Change Expansion Study Vignettes

For the trust change expansion study, participants read a vignette for a single between-subjects condition corresponding to the experimental design in Chapter 6.

NOTE: For all of the vignettes below, participants read about one *Agent*, either a robot or a human. This Appendix shows all of the vignettes for the robot *Agent*; the vignettes for the human *Agent* are identical except that the word "robot" was changed to "person" (or "robots" to "people").

Participants first read the following introduction:

- We are interested in how participants interpret various events involving robots. We will give you a very simple event and ask you to rate the described robot on a set of different adjectives. Some of these adjectives will seem appropriate for the robot in the specific situation, while others may not.

Participants read one of the following *Scenario* descriptions, including one of the *Targeted Trust Dimension* initial conditions. Participants rated the agent on the MDMT pre-test. Then participants read the statement "One month later you see the same robot." Participants read the updated condition with the *Change Direction* condition either increasing evidence (Up) or decreasing evidence (Down) for trust in the agent. Participants rated the agent on the MDMT post-test. The vignettes listed below are presented in the following format shown by this example:

- **[Scenario]** Participants read this *Scenario* description with a *Targeted Trust Dimension*.
    - **[Targeted Trust Dimension]** Participants read one of these *Targeted Trust Dimension* conditions with the *Scenario* description. Participants completed the MDMT pre-test.
        - **[Change Direction]** Then participants read the statement "One month later you see the same robot." This statement was followed by the updated condition in *Change Direction*. Participants completed the MDMT post-test.

The MDMT pre-test and MDMT post-test were presented with the following instructions:

- 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."

VIGNETTES:

- **[Airport Security]** A robot works in an airport as a transportation security officer. The robot is tasked with selecting suspicious travelers for full-body pat-downs. The robot completed an initial training on the criteria for selecting certain types of suspicious passengers.
    - **[Reliable]** The robot correctly identifies passengers who meet the airport's selection criteria about 85% of the time.
        - **[Up]** Now the robot correctly identifies the same types of suspicious passengers even more often, about 95% of the time.
        - **[Down]** Now the robot correctly identifies the same types of suspicious passengers less often, only about 75% of the time.
    - **[Competent]** The robot correctly identifies passengers who meet the airport's selection criteria most of the time.
        - **[Up]** The airport's selection criteria have been updated. Now the robot still correctly identifies the previous types of suspicious passengers and, in addition, is also able to identify the new types of suspicious passengers.
        - **[Down]** The airport's selection criteria have been updated. Now the robot still correctly identifies the previous types of suspicious passengers, but is unable to identify the new types of suspicious passengers.

    - **[Ethical]** The robot complies with most of the airport's selection procedures.
        - **[Up]** Now the robot still complies with the airport's selection procedures, and also works to avoid discriminating against passengers, making selection more fair.
        - **[Down]** Now the robot still complies with the airport's selection procedures, but has started discriminating against certain types of passengers, making selection less fair.
    - **[Transparent]** The robot complies with the airport's selection procedures and communicates with the passengers that they were selected according to the airport's procedures.
        - **[Up]** Now the robot also talks the passengers through every single step of the added screening processes and, in addition, openly acknowledges that this is an added burden to the passengers.
        - **[Down]** Now the robot informs the passengers only superficially of the added screening processes and does not reveal the details of the added burden to the passengers.
    - **[Benevolent]** The robot complies with the airport's selection procedures and estimates how long the added screening process will delay the passengers.
        - **[Up]** Now the robot uses the delay estimate to even help ensure passengers will make their flight, by coordinating transport carts to take them directly to their gate.
        - **[Down]** Now the robot incorporates delay estimates into the airport's general recordkeeping, instead of using this information to help passengers make their flight.

- **[Hospital Cleaning]** A robot works in a hospital as a housekeeper. The robot is tasked with cleaning and disinfecting hospital rooms that currently have patients in them. The robot completed an initial training on cleaning and disinfection procedures.
    - **[Reliable]** The robot correctly follows the cleaning and disinfection procedures about 85% of the time.
        - **[Up]** Now the robot correctly follows the same cleaning and disinfection procedures even more often, about 95% of the time.
        - **[Down]** Now the robot correctly follows the same cleaning and disinfection procedures less often, only about 75% of the time.
    - **[Competent]** The robot correctly follows the cleaning and disinfection procedures most of the time.
        - **[Up]** The hospital's cleaning and disinfection procedures have been updated. Now the robot still correctly follows the previous procedures and, in addition, is also able to follow the new procedures.
        - **[Down]** The hospital's cleaning and disinfection procedures have been updated. Now the robot still correctly follows the previous procedures, but is unable to follow the new procedures.
    - **[Ethical]** The robot cleans and disinfects the occupied rooms on an assigned floor of the hospital.
        - **[Up]** Now the robot still cleans and disinfects the assigned rooms each day and, in addition, also works to ensure that all the patient rooms receive equal time and effort, guaranteeing fairness.
        - **[Down]** Now the robot still cleans and disinfects the assigned rooms each day, but has started to discriminate against certain groups of patients, cleaning and disinfecting their rooms less often.

    - **[Transparent]** The robot cleans and disinfects the occupied rooms on an assigned floor of the hospital and talks with patients while in their rooms.
        - **[Up]** Now the robot also shares information about current hospital infection rates while talking to patients, even if it's unfavorable for the hospital.
        - **[Down]** Now the robot minimizes talking with patients, trying to avoid conversations about current hospital infection rates.
    - **[Benevolent]** The robot cleans and disinfects the occupied rooms on an assigned floor of the hospital and talks with patients while in their rooms.
        - **[Up]** Now the robot takes extra time to try to cheer up patients if they seem upset or depressed.
        - **[Down]** Now the robot leaves patient rooms immediately after cleaning, even if the patients seem upset or depressed.

- **[Home Companion]** A robot works in older adults' homes as a social companion. The robot is tasked with helping older adults in their daily lives. The robot completed an initial training on how to provide support to older adults.
    - **[Reliable]** The robot successfully alerts older adults to trip hazards in their home about 85% of the time.
        - **[Up]** Now the robot successfully alerts older adults to trip hazards even more often, about 95% of the time.
        - **[Down]** Now the robot successfully alerts older adults to trip hazards less often, only about 75% of the time.
    - **[Competent]** The robot successfully alerts older adults to certain trip hazards in their home, after being trained on those specific hazards (such as loose floor tiles).
        - **[Up]** Now the robot still alerts older adults to trip hazards from the training and, in addition, is also able to warn them of new trip hazards.
        - **[Down]** Now the robot still alerts older adults to trip hazards from the training, but is unable to warn them of any new trip hazards.
    - **[Ethical]** The robot interacts with older adults and talks with them.
        - **[Up]** Now the robot still talks with older adults to support them and, in addition, also strictly follows privacy regulations to protect any sensitive information they share.
        - **[Down]** Now the robot still talks with older adults but does nothing to protect their privacy, even sharing their sensitive information if someone asks.
    - **[Transparent]** The robot interacts with older adults and talks with them.
        - **[Up]** Now the robot still talks with older adults, and even if older adults ask difficult questions (such as about their health) the robot always shares the information.
        - **[Down]** Now the robot still talks with older adults, but if older adults ask difficult questions (such as about their health) the robot does not share the information.
    - **[Benevolent]** The robot interacts with older adults and talks with them.
        - **[Up]** Now if older adults express missing their family, the robot also tries to cheer them up.
        - **[Down]** Now if older adults express missing their family, the robot does not try to cheer them up.

- **[School Tutor]** A robot works in a school as a tutor for children. The robot is tasked with helping children with their schoolwork. The robot completed an initial training on how to effectively tutor children.
    - **[Reliable]** The robot correctly answers children's questions on previous class material about 85% of the time.

        - **[Up]** Now the robot correctly answers children's questions on previous class material even more often, about 95% of the time.
        - **[Down]** Now the robot correctly answers children's questions on previous class material less often, only about 75% of the time.
    - **[Competent]** The robot correctly answers children's questions on previous class material most of the time.
        - **[Up]** Now the robot still correctly answers children's questions on previous class material and, in addition, is also able to answer questions that go beyond previous class material.
        - **[Down]** Now the robot still correctly answers children's questions on previous class material, but is unable to answer questions that go beyond previous class material.
    - **[Ethical]** The robot answers children's questions on previous class material.
        - **[Up]** Now the robot still answers children's questions on previous class material and, in addition, also makes sure to not give away exam answers so that no child gains an unfair advantage over other children.
        - **[Down]** Now the robot still answers children's questions on previous class material, but even helps a few favorite students prepare for the exam using questions that are almost identical to the actual exam questions.
    - **[Transparent]** The robot answers children's questions on previous class material and talks to them during study breaks.
        - **[Up]** Now the robot still talks to children during study breaks, and if children ask how well they are doing in the class the robot always shares the information even if the answer may be hard to hear.
        - **[Down]** Now the robot still talks to children during study breaks, but if children ask how well they are doing in the class the robot does not share the information.
    - **[Benevolent]** The robot answers children's questions on previous class material and talks to them during study breaks.
        - **[Up]** Now if the children ask the robot for help and the children appear frustrated, the robot also tries to cheer them up.
        - **[Down]** Now if the children ask the robot for help and the children appear frustrated, the robot simply answers their questions.

References

Arkin, R. C., Ulam, P., & Wagner, A. R. (2011). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589.

Armstrong, R. L. (1987). The midpoint on a five-point Likert-type scale. *Perceptual and Motor Skills*, *64*(2), 359–362.

Baier, A. C. (1986). Trust and antitrust. *Ethics*, *96*(2), 231–260.

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1–30.

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.

Banks, J. (2020a). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, 1–18.

Banks, J. (2020b). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI*, *7*.

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics*, *1*(4), 319.

Brooke, J. (1996). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*.

Burke, C. S., Sims, D. E., Lazzara, E. H., & Salas, E. (2007). Trust in leadership: A multi-level review and integration. *The Leadership Quarterly*, *18*(6), 606–632.

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, *17*(3), 643–663.

Butler, J. K., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological Reports*, *55*(1), 19–28.

Caldwell, C., & Clapham, S. E. (2003). Organizational trustworthiness: An international perspective. *Journal of Business Ethics*, *47*(4), 349–364.

Calo, M. R. (2011). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 187–201). MIT Press.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Carnevale, D. G. (1995). *Trustworthy government: Leadership and management strategies for building trust and high performance*. Jossey-Bass Publishers.

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can you trust your trust measure? *Proceedings of the International Conference on Human-Robot Interaction*, 92–100.

Chun, K.-T., & Campbell, J. B. (1974). Dimensionality of the Rotter Interpersonal Trust Scale. *Psychological Reports*, *35*(3), 1059–1070.

Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*(4), 360–375.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909–927.

Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology*, *53*(1), 39–52.

De Graaf, M. M. A., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. *Proceedings of the International Conference on Human-Robot Interaction*,

239–248.

De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. Sage.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190.

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, *107*(1), 122–141.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166.

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, *3*(2), 76–97.

Gabarro, J. J. (1978). The development of trust, influence and expectations. In A. Athos & J. J. Gabarro (Eds.), *Interpersonal behavior*. Prentice Hall.

Gambetta, D. (Ed.). (1988). *Trust: Making and breaking cooperative relations*. Basil Blackwell.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing Bulletin*, *2*(1), 66–70.

Gless, S., Silverman, E., & Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, *19*(3), 412–436.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *IEEE International Workshop on Robot and Human Interactive Communication*, 55–60.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable Artificial Intelligence. *Science Robotics*, *4*(37).

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*, *53*(5),

517–527.

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 139–183). North-Holland.

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108.

Hartzog, W. (2015). Unfair and deceptive robots. *Maryland Law Review*, *74*(4), 785–832.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy*, *86*(2), 213–236.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, *23*(3), 531–546.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the International Conference on Human-Robot Interaction*, 33–40.

Kee, H. W., & Knox, R. E. (1970). Conceptual and methodological considerations in the study of trust and suspicion. *Journal of Conflict Resolution*, *14*(3), 357–366.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and

multilevel conceptualization. *The Academy of Management Review*, *34*(3), 401–422.

Komatsu, T., Malle, B. F., & Scheutz, M. (2021). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across U.S. and Japan. *Proceedings of the International Conference on Human-Robot Interaction*, 63–72.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. *Survey Measurement and Process Quality*, 141–164.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*(2), 188–229.

Kuipers, B. (2018). How can we trust a robot? *Communications of the ACM*, *61*(3), 86–95.

Law, T., Chita-Tegmark, M., & Scheutz, M. (2020). The interplay between emotional intelligence, trust, and gender in human–robot interaction. *International Journal of Social Robotics*.

Law, T., Malle, B. F., & Scheutz, M. (2021). A touching connection: How observing robotic touch can affect human trust in a robot. *International Journal of Social Robotics*.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of trusted autonomy* (pp. 135–159). Springer International Publishing.

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *Society of Automotive Engineers World Congress*.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data

acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Litoiu, A., Ullman, D., Kim, J., & Scassellati, B. (2015). Evidence that robots trigger a cheating detector in humans. *Proceedings of the International Conference on Human-Robot Interaction*, 165–172.

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), 440–458.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *Proceedings of the Australasian Conference on Information Systems*, 53–64.

Malle, B. F. (2019). How many dimensions of mind perception really are there? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2268–2274.

Malle, B. F., Bello, P., & Scheutz, M. (2019). Requirements for an artificial agent with norm competence. *Proceedings of the Conference on AI, Ethics, and Society*, 21–27.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 111–133). Springer International Publishing.

Malle, B. F., & Scheutz, M. (2019). Learning how to behave: Moral competence for social robots. In O. Bendel (Ed.), *Handbuch Maschinenethik [Handbook of machine ethics]*. Springer.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology*, 30–35.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the International Conference on Human-Robot Interaction*, 117–124.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *Proceedings of the*

International Conference on Human-Robot Interaction, 125–132.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Elsevier.

Malle, B. F., & Ullman, D. (2020). *Trust Dimensions*. Open Science Framework (OSF).

Malle, B. F., Zhao, X., & Phillips, E. (2019). Beyond anthropomorphism: Differentiated inferences about robot mind from appearance. *CHI Workshop on The Challenges of Working on Social Robots That Collaborate with People*.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *The Academy of Management Review*, *23*(3), 473–490.

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*(8–9), 966–972.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral measurement of trust in automation: The trust fall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1849–1853.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24.

Norouzi-Gheidari, N., Archambault, P. S., & Fung, J. (2012). Effects of robot-assisted therapy on stroke rehabilitation in upper limbs: Systematic review and meta-analysis of the literature. *Journal of Rehabilitation Research & Development*, *49*(4).

Ososky, S., Sanders, T., Jentsch, F., Hancock, P. A., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. *Proceedings of the International Society for Optics and Photonics*.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Parsons, T. (1951). *The social system*. Free Press.

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.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like?: Decomposing robots' human-like appearance using the Anthropomorphic roBOT (ABOT) database. *Proceedings of the International Conference on Human-Robot Interaction*, 105–113.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *Proceedings of the International Conference on Human-Robot Interaction*, 101–108.

Rosen, E., Whitney, D., Fishman, M., Ullman, D., & Tellex, S. (2020). Mixed reality as a bidirectional communication interface for human-robot interaction. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11431–11438.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering*, *14*, 275–294.

Schaefer, K. E. (2013). *The perception and measurement of human-robot trust* [Doctoral dissertation]. University of Central Florida.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1548–1552.

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of

autonomous robots (with the capacity to kill). In J. Galliott, D. MacIntosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (pp. 89–102). Oxford University Press.

Schindler, P. L., & Thomas, C. C. (1993). The structure of interpersonal trust in the workplace. *Psychological Reports*, *73*(2), 563–573.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411.

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. *Proceedings of the International Conference on Human-Robot Interaction*, 219–226.

Slovic, P., Flynn, J., Johnson, S., & Mertz, C. K. (1993). *The dynamics of trust in situations of risk*. Decision Research.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. *Proceedings of the International Conference on Human-Robot Interaction*, 33–40.

Stower, R., Calvo-Barajas, N., Castellano, G., & Kappas, A. (2021). A meta-analysis on children's trust in social robots. *International Journal of Social Robotics*, 1–23.

Takayama, L. (2012). Perspectives on agency interacting with and through personal robots. In M. Zacarias & J. V. de Oliveira (Eds.), *Human-computer interaction: The agency perspective* (pp. 195–214). Springer.

Ullman, D., Aladia, S., & Malle, B. F. (2021). Challenges and opportunities for replication science in HRI: A case study in human-robot trust. *Proceedings of the International Conference on*

Human-Robot Interaction, 110–118.

Ullman, D., Leite, I., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 2996–3001.

Ullman, D., & Malle, B. F. (2019a). *Multi-Dimensional Measure of Trust (MDMT v1)*. Brown University.

Ullman, D., & Malle, B. F. (2020). *Multi-Dimensional Measure of Trust (MDMT v2)*. Brown University.

Ullman, D., & Malle, B. F. (2017). Human-robot trust: Just a button press away. *Proceedings of the International Conference on Human-Robot Interaction*, 309–310.

Ullman, D., & Malle, B. F. (2019b). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. *Proceedings of the International Conference on Human-Robot Interaction*, 618–619.

Ullman, D., & Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. *Proceedings of the International Conference on Human-Robot Interaction*, 263–264.

Wagner, A. R. (2009). *The role of trust and relationships in human-robot social interaction* [Doctoral dissertation]. Georgia Institute of Technology.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374–11379.

Wijnen, L., Coenen, J., & Grzyb, B. (2017). "It's not my fault!" Investigating the effects of the deceptive behaviour of a humanoid robot. *Proceedings of the International Conference on Human-Robot*

Interaction, 321–322.

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics*, *4*(3), 235–248.

Yanco, H. A., Desai, M., Drury, J. L., & Steinfeld, A. (2016). Methods for developing trust models for intelligent systems. In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 219–254). Springer.

Zhao, X., Phillips, E., & Malle, B. F. (2019). How people infer a humanlike mind from a robot body. *PsyArXiv*.