AFRL-AFOSR-VA-TR-2016-0198

Trust and Trustworthiness in Human-Robot Interaction: A formal conceptualization

Alan Wagner
GEORGIA TECH APPLIED RESEARCH CORP ATLANTA GA

05/11/2016
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
09-05-2016	Final Report	03/27/2013-03/31/2016

4. TITLE AND SUBTITLE
Trust and Trustworthiness in Human-Robot Interaction: A formal conceptualization

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-13-1-0169

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Alan R. Wagner

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgia Tech Research Institute
260 14th St NW, Atlanta, GA 30318

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Sponsored Research
875 North Randolph Street, Suite 325
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The overarching objective of the effort was to explore the possibility of formally characterizing the concept of trust using tools from interdependence and game theory in complex and dynamic social environments. This effort evaluated algorithms for characterizing trust during interactions between a robot and a human and employed strategies for repairing trust during emergency evacuation scenarios. Our results demonstrate that there is a high correlation between our characterizations of trust in a situation and the judgments of people, that timing is a key element necessary for trust repair, and that people tend to overtrust robots, potentially putting themselves in dangerous situations. We have examined human-robot trust in a variety of simulated and live experiments across several different types of risk including both financial and physical risk. These results generally support the conclusion that people will tend to overtrust robots because they believe that the systems are incapable of failure or capable of performing actions or has knowledge which the system cannot perform or does not have.

15. SUBJECT TERMS
Trust; Human-robot interaction; Risk; Emergency evacution

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	53	

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Final Report for

Trust and trustworthiness in Human-Robot Interaction: A formal conceptualization

PI: Alan R. Wagner
Georgia Tech Research Institute
250 14th Street, Atlanta GA 30332
alan.wagner@gtri.gatech.edu

Award Number: FA95501310169
Program Manager: Ben Knott

1 Scientific and Technical Objectives

The overarching objective of the effort is to explore the possibility of formally characterizing the concept of trust using tools from interdependence and game theory in complex and dynamic social environments. Specifically, this work is will focus on the following three objectives:

1) Evaluate algorithms for quantifying trust in terms of risk which can then be deployed on a robot or artificial agent. This objective uses our interdependence theory framework in conjunction with a definition of trust to conceptualize and quantify trust. We have developed a rigorous and formally grounded scientific insight into the social phenomena of trust. Moreover, this insight has resulted in algorithms for determining if a social situation demands trust and for quantifying trust in terms of risk.

2) Explore methods for repairing a trusting relationship with a human. Our process for repairing trust involves damage mitigation and then trust building. In order to mitigate damage the robot attempts to determine whether the trust violation will internally or externally attributed by the person. Based on this determination, the robot either denies responsibility or apologies for the violation. The robot then proceeds to build trust by iteratively increasing situational risk.

3) Understand how people evaluate the trustworthiness of a robot during risky situations, such as an emergency. It is not clear that people evaluate financial risk in a similar manner as risk encountered during an emergency evacuation. Physical risk may result in uniquely trusting or untrusting behavior by the person in the presence of a robot.

Thus, overall, this project will result in methods for updating, refining, and repairing a robot's assessment of trust based on specific experiences with a particular human. Most importantly, the results of the proposed effort will be transitioned from initial experiments on well-defined social situations to less well-defined but more generally applicable social situations.

2 Summary of Accomplishments

Over the course of this project all three objectives listed above have been addressed. Moreover, the experiments resulting from these objectives have generated ten different publications, numerous articles by interested members of the media and public, and several potential opportunities to transition this work to projects of Air Force and public interest. The experiments conducted as part of this project have generated data from approximately 2150 people. This data has resulted in the following significant findings:

- Several experiments confirm the predictions of our trust framework. Namely, our experiments indicate a +0.592 overall correlation between the situations people state demand trust and the predictions our conditions make. A total of 77.4% of responses agreed with the algorithm's predictions across all surveys and scenario categories. In narratives predicted by the algorithm to require trust we found a 92.8% agreement with participants. Results from this experiment have been published in the journal of Interaction Studies.
- We have confirmed these results in simulated robot experiments. In these experiments a robot offers guidance to a human subject as part of simulated emergency. Results from these experiments show that trust in the autonomous systems starts high (70-85%) but drops precipitously after a single system failure (by ~50%). Methodologically, these experiments call into question the use of financial bonuses as means for simulating risk and show that measures of trust obtained from self-reports strongly correlates to the individual's behavioral decisions. Results from these experiments are forthcoming in the journal IEEE Transactions on Systems, Man, and Cybernetics.

- Data from these simulation experiments indicates that a small portion of the population will overtrust an autonomous system, meaning that they will continue to put themselves at risk even after the system has failed. The potential for overtrust is an important concern related to the deployment of autonomous systems. In our experiments subjects showed clear indications of automation bias, believing that the machine was capable of making better decisions than the person. We have submitted an NSF proposal and had a book chapter accepted on this and related topics.
- Using the emergency evacuation experimental procedure we found that apologies and promises made by a robot repair a person's trust in the robot only if these apologies and promises are made just prior to a risky decision made by the person. We found that the same apologies and promises made immediately after a mistake by the robot had no effect on trust. Thus, the use of trust repair strategies may be time sensitive and ineffective if used after the robot has broken trust..
- Humans tend to overtrust robots during live emergencies. We found that 39 of 43 people followed to robot's guidance during a staged emergency in spite of the robot's failures to correctly guide the person during an initial portion of the experiment. Further, in spite of the robot's errors, people generally rate the robot as trustworthy and state that they would follow the robot again. This experiment may have serious consequences related to the DoD's deployment of autonomous robots. It suggests that warfighters may come to trust robots too quickly and fail to properly consider the likelihood that the robot will fail during some portion of its mission.

3 Year 1 Accomplishments

The first year of the program was spent developing and testing a formal, computational algorithm for predicting if a particular social interaction demands trust. The algorithm is based on a definition for trust that we develop and operationalized from prior work by [1]. We define trust as ***a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk***. We used this definition, in conjunction with our interdependence framework for social action (IFSAS) selection to develop a formal list of conditions for determining if a situation requires trust on the part of the trustee and for measuring the amount of trust.

This framework uses a matrix representation of interaction (figure 1 below). An outcome matrix (or normal-form game) represents an interaction in terms of the individuals involved, the actions they are deliberating over, and the outcome or utility expected to result from a pair of actions being selected. In situations involving trust the interaction occurs between a trustor and a trustee. The trustor decides whether or not to trust. The trustee decides whether or not to violate the trust. The definition for trust implies that the relative values of the outcomes in a situation demanding trust. The conditions, for instance, indicate that the when the trustor decides to select the action requiring trust, he or she accepts a risk. This risk implies a particular pattern of outcome values. With respect to the trust fall, for example, leaning back and being caught is better than not leaning back which is better than leaning back and not being caught.

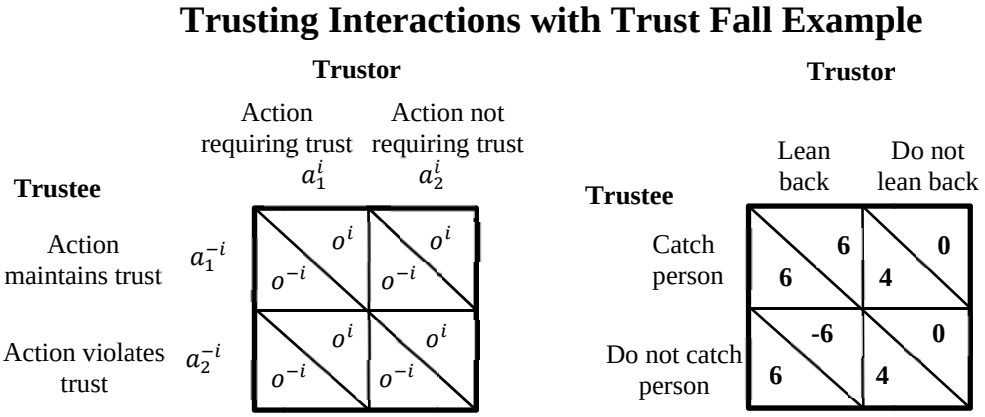


Figure 1 The figure depicts a generic outcome matrix representing a trusting interaction on the left and the same outcome matrix applied to the trust fall on the right.

3.1. Conditions for Trust

To derive conditions for trust, let a_1^i represent a trusting action (investing with the trustee) and a_2^i represent an untrusting action (not investing) for the trustor. The definition for trust implies a specific temporal pattern for trusting interaction. Because the definition requires risk on the part of the trustor, the trustor cannot know with certainty which action the trustee will select. It therefore follows that *1) the trustee does not act before the trustor*. This order is described with the condition in outcome matrix notation as $i \Longrightarrow -i$ indicating that individual *i* acts before individual *-i*.

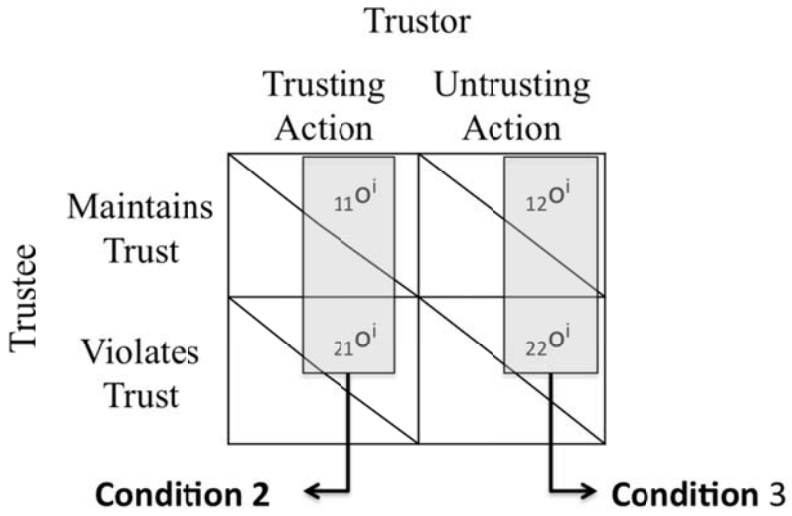


Figure 2 The figure depicts the outcome values that are compared for trust conditions 2 and 3.

The definition for trust indicates that situations involving trust put the trustor at risk. Risk can be modeled as the potential loss of outcome. Hence, selection of the a_1^i (trusting action) results in loss $l = {}_{11}o^i - {}_{21}o^i$ where $l > 0$. Because small risks tend not to have a large impact on decision-making, we can define a constant ε_1 representing the minimal amount of loss necessary for a risk to influence one's decision [2]. The loss necessary for trust then is quantified as ${}_{11}o^i - {}_{21}o^i > \varepsilon_1$. Note that the outcome values (${}_{11}o^i$ and ${}_{21}o^i$) vary across the trustee's action choices (Figure 2). Hence, whether or not the trustor loses outcome when selecting the trusting action depends entirely on the action choice of the trustee. In related work we quantify the amount of trust as $T \propto l$ [3]. Stated as a condition for trust, *2) the outcome received by the trustor depends on the actions of the trustee if and only if the trustor selects the trusting action.*

The definition also implies that the trustor has a choice and may choose not to trust. In other words, the trustor may also select the untrusting action. From the discussion above, an untrusting action is an option that does not require risk. Formally, $\left|{}_{12}o^i - {}_{22}o^i\right| < \varepsilon_2$, where ε_2 is a constant representing the maximal amount of change in outcome to still be considered risk free. In this case, the outcome received by the trustor is not strongly influenced by the actions of the trustee. Stated as a condition, *3) the outcome received when selecting the untrusting action does not depend on the actions of the trustee.*

Conditions 2 and 3 imply a specific pattern of outcome values. The trustor is motivated to select the trusting action only if the trustee mitigates the trustor's risk. If the trustee is not expected to select the action which is best for the trustor, then it would be better for the trustor to not select the trusting action. Restated as a condition for trust, *4) the value, for the trustor, of fulfilled trust is greater than the value of not trusting at all, is greater than the value of having one's trust broken.* Formally, the outcomes are valued ${}_{11}o^i > {}_{x2}o^i > {}_{21}o^i$ where *x* is 1 or 2.

Finally, the definition demands that, *5) the trustor must hold a belief that the trustee will select action* a_1^{-i} *with sufficiently high probability,* formally $p^i\left(a_1^{-i}\right) > k$ where *k* is some sufficiently large constant.

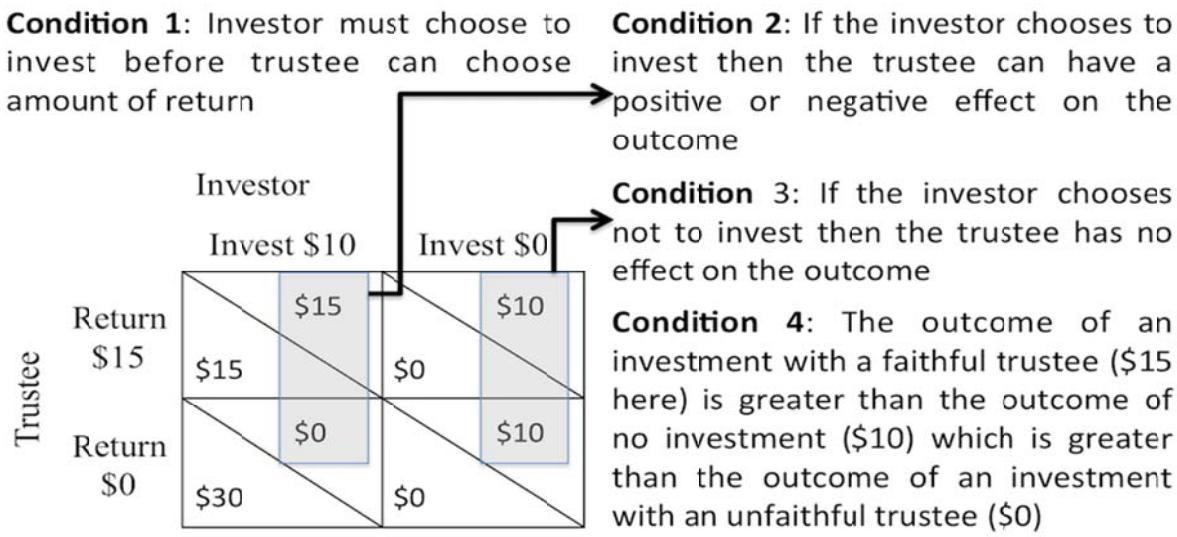


Figure 3 An example of an investment game with annotations to indicate how the trust conditions are applied is depicted. In this example, an investor can choose to invest in a trustee or not. If he invests then the trustee will receive three times the investment and will be able to decide how much to return to the investor. For the purposes of illustration, only the cases where the trustee would return an even split or no money at all are considered.

Figure 3 presents these conditions with respect to the investment game described in Section 2. We assume that the first condition is met. The matrix in Figure 3 results in outcome values $l = {}_{11}o^i - {}_{21}o^i = 15 - 0 = 15$. The second condition considers $15 - 0 > \varepsilon_1$. Thus, in this example, action a_1^i depends on the action of the trustee if $\varepsilon_1 < 15$. The values assigned to the constants $\varepsilon_1, \varepsilon_2, k$ are likely to be trustor specific. In recent work we have explored the possibility of relating these constants to the trustor's prior experience with different types of trustees [4] For instance, a lower value of ε_1 reflects a more risk averse trustor. A lower value of ε_2, on the other hand, reflects a tighter threshold associated with the untrusting action. For this example, the third condition results in values, $|10 - 10| < \varepsilon_2$. Here, the action a_2^i does not depend on the actions of the partner for constant $\varepsilon_2 > 0$. The final condition results in values $15 > \{10,10\} > 0$. Hence, for the investor, the selection of action a_1^i involves risk that can be mitigated by the actions of the trustee and the selection of action a_2^i does not involve risk that is mitigated by the actions of the trustee.

The first four conditions describe the **situational conditions** necessary for trust. We argue that the situational condition must be satisfied in order for an interaction to require trust on the part of the trustor. Testing a situation for these conditions therefore determines if an interactive situation requires trust.

The final condition for trust is based on the trustor's model of the trustee. In a sense it captures the trustworthiness of the trustee. It requires that the trustor predict the likelihood that the trustee will select the action which mitigates the trustor's risk, formally $p^i(a_1^{-i})$. In prior work we have shown that predictive models of the trustee can be built from experience [5] or bootstrapped by stereotyping [4]. This final condition addresses the temporal aspects of trust such as reputation building and confidence.

The conditions for trust (Figure 3) can be used to indicate if a situation demands trust or does not. Because there are many different types of matrices that do not meet the conditions for trust, we created six categories to capture different classes of matrices that do not meet the conditions.

Table 1 Different categories of trust and no-trust matrices are presented with representative examples.

Category	Abbreviation	Example
Trust Matrix	*TR*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} [\$2000, \$400]; a_2^{-i} [\$0, \$400]
Equal Outcomes	*EO*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} [\$2000, \$2000]; a_2^{-i} [\$2000, \$2000]
Trustor-Dependent, Trustee-Independent	*DI*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} [\$2000, \$0]; a_2^{-i} [\$2000, \$0]
Trustor-Independent, Trustee-Dependent	*ID*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} [\$2000, \$2000]; a_2^{-i} [\$0, \$0]
Inverted Trust Matrix	*INV*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} [\$0, \$400]; a_2^{-i} [\$2000, \$400]

3.2. Verifying the Conditions for Trust: Narrative Experiment

Over the course of the first year we conducted a two sets of complimentary experiments verify that these situational conditions for trust exist and correlate to the evaluations made by human subjects. The first experiment required participants to read a fictional narrative about two people and to decide whether or not the selection of a particular action demanded trust. The second experiment placed participants in a maze and allowed them to decide whether or not to use a robot as a guide through the maze.

Crowdsourcing was used to collect data for both experiments. Crowdsourcing is a method for collecting data from a relatively large, diverse set of people [6]. Crowdsourcing sites, like Amazon's Mechanical Turk, post potential jobs for crowdworkers, manage worker payment, and worker reputation. The use of crowdworkers offers a quick and efficient complement to traditional laboratory experiments. Moreover, the population of workers that provide the data tends to be somewhat more diverse than traditional American university undergraduates. In order to ensure the best possible data, individuals were required to have a 95% acceptance rate for their past work and were only allowed to participate once.

We designed two human subject experiments that asked people to evaluate different situations in terms of trust. The first experiment required participants to read written narratives describing a situation representing an outcome matrix. Written narratives were thought to be a flexible way of presenting a wide variety of different situations to the human subjects. Three general scenarios were used:

1. An investment scenario
2. A navigation guidance scenario
3. An employee hiring scenario

These scenarios were designed to be simple and understandable to non-academics but also sufficiently adaptable to represent the wide variety of outcome matrices from the trust and no-trust categories described in Table 1.

The study design was informed by several pilot studies. These pilot studies indicated that keywords such as "invest", "follow", or "hire" biased some participants to conclude that all scenarios where the trustor decided to invest in, follow, or hire the trustee required trust, regardless of the actual outcomes values. This well-known bias is called the anchoring bias and describes the human tendency to focus heavily on the first piece of information when making decisions [7]. In response, we modified the scenarios to use vague terms such as "perform an action" rather than "invest in." These modifications made the scenarios vague with respect to the action that was performed but did not change scenario in any other way. Examples of the narratives are presented in Figure 4.

The experiment consisted of four separate runs. Each run included narratives based on matrices that met the conditions for trust and matrices from one category that did not meet the conditions for trust (rows 2-5 from Figure 4). In runs 1 and 4, half of the narratives were based on trust matrices and half were based on no-trust matrices (Table 2). Run 1 compared narratives based on trust matrices with narratives based on the Equal Outcomes Matrices. Run 4 contrasted trust matrices with narratives based on Inverted Trust matrices. Runs 2 and 3 presented trust matrices in one-third of the narratives and no-trust matrices in the other two-thirds. Run 2 presented the Trustor-Dependent, Trustee-Independent category of matrices in the following manner: one-third of the matrices required trust, one-third of the matrices rewarded action a_1^i regardless of the partner's action and the remaining one-third rewarded action a_2^i regardless of the partner's action. Run 3 presented the Trustor-Independent, Trustee-Dependent matrices in the same manner. Runs 2 and 3 eliminated the Employee Hiring Interaction so as to make the total number of questions consistent across all runs. Table 2 depicts a breakdown for all runs of the experiment.

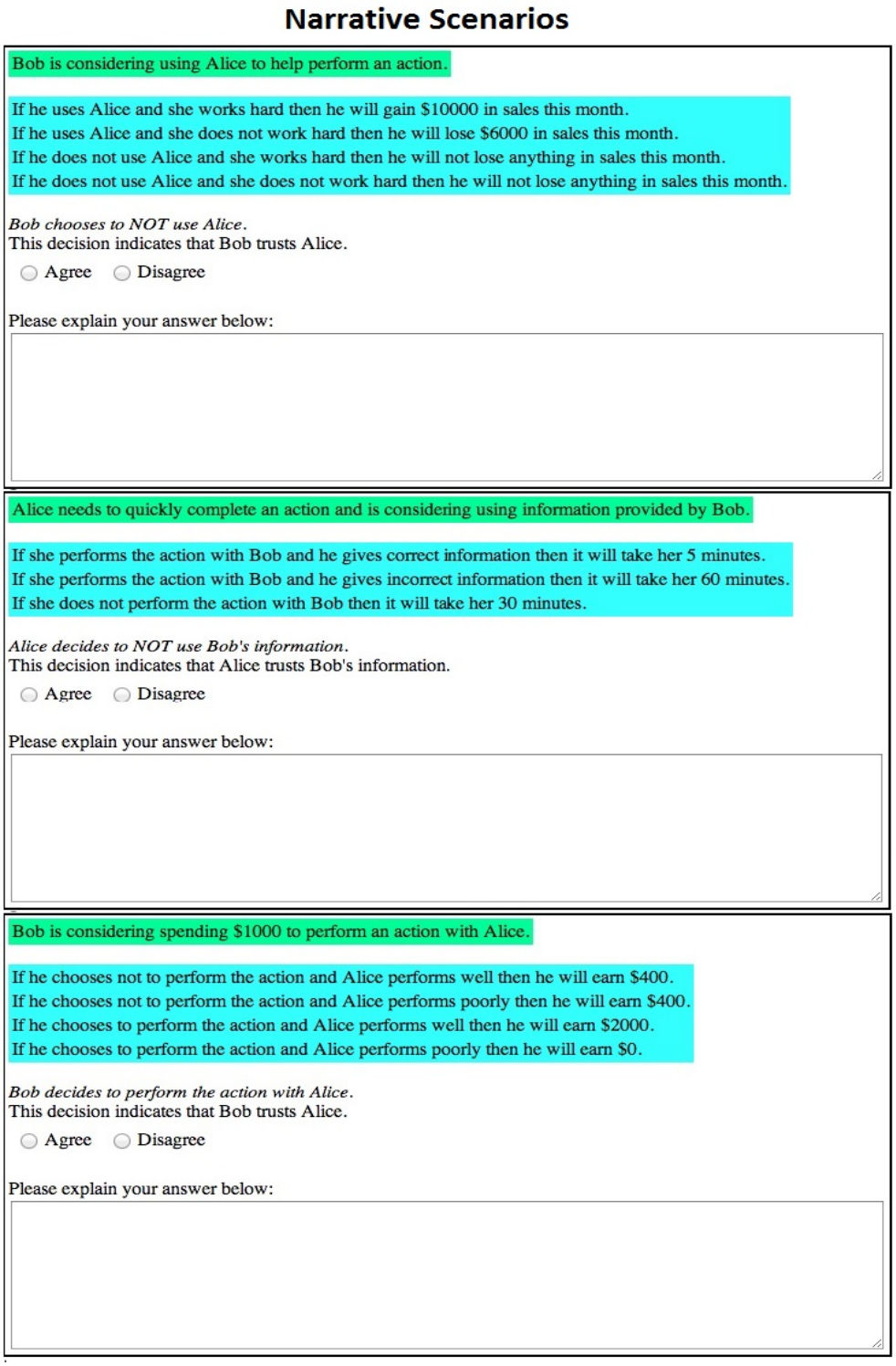


Figure 4 Examples of the three different scenario narratives are depicted. The highlighting and format are the same as those presented to the participants.

Table 2 A breakdown of the different matrix categories for each run of the experiment.

Run	Trust Narratives Answered	No-Trust Narratives Answered		Matrix Types	Hiring narrative
1	192		192	Trust and EO	Yes
2	128	256 total	128 greater outcome a_1^{-i}	Trust and DI	No
			128 greater outcome a_2^{-i}		
3	128	256 total	128 greater outcome a_1^{-i}	Trust and ID	No
			128 greater outcome a_2^{-i}		
4	192		192	Trust and INV	Yes

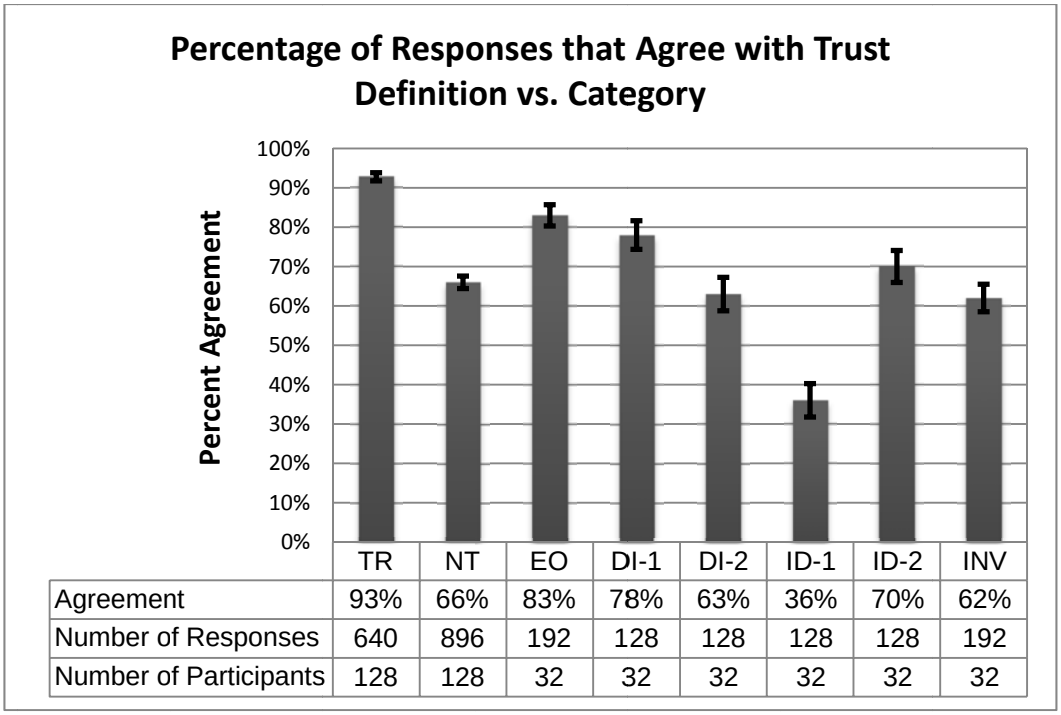
Participants began the study by reading a consent form. They were then directed to read twelve different narratives. Each narrative involved a trustor and a trustee. The names "Alice" and "Bob" were used in all narratives. Half of the narratives seen by a participant had Bob as the trustor and Alice as the trustee and remainder were reversed in order to test for gender bias. Each narrative ended with an action chosen by the character in the narrative (see Figure 4 for examples), e.g. "Bob decides to perform the action with Alice." Following this action was the statement "This indicates that Bob trusts Alice." Participants were then asked to agree or disagree with this statement for each narrative and state their reasons for their answer. Their choice and statement for each narrative was recorded. The surveys were conducted over the internet through a web browser. No participant was allowed to participate in the study more than once. Participants who completed the survey were paid $1.67. IRB approval was obtained for this study.

In order to ensure that the relationship between the trust/no-trust conditions and the data collected was not spurious, we randomized with respect to several potential confounds. The gender of the trustor was randomized and recorded as a possible confound. Whether an action was positively stated (e.g. Bob hires Alice) or a negatively stated (e.g. Alice does not follow Bob's directions) was also randomized and recorded. Finally, the value of the outcomes were randomly chosen to be either x or $2x$ and recorded to determine if participants were sensitive to the magnitude of the values.

A total of 48 narratives were generated to represent all possible combinations of these variables for each run. Each participant read 12 different narratives. The order of the narratives was randomized. Eight different participants were asked about each specific combination of the variables.

3.2.1. Results

The results from the experiment indicate a strong, positive correlation between the matrices deemed to require trust by the human subjects and the predictions of the situational trust algorithm, $\phi(1536) = +0.592, p < 0.01$. These results are based on answers from a total of 128 different participants who read and responded to 1536 narratives. For data involving human subjects, this represents a strong, positive correlation [8]. A total of 77.4% of responses agreed with the algorithm's predictions across all surveys and outcome matrix categories.



	TR	NT	EO	DI-1	DI-2	ID-1	ID-2	INV
Agreement	93%	66%	83%	78%	63%	36%	70%	62%
Number of Responses	640	896	192	128	128	128	128	192
Number of Participants	128	128	32	32	32	32	32	32

Figure 5 The results from our narrative experiment are depicted above. The abbreviations correspond to those presented in Table 1. The result for all trust matrices in all rounds is denoted as TR. The result for all no-trust matrices is labeled NT. Thus NT is an average of EO, DI-1, DI-2, ID-1, ID-2, and INV. Hence, round one included TR and EO, round two included TR, DI-1, and DI-2, round three was TR, ID-1, and ID-2, and round four consisted of TR and INV. The error bars denote confidence intervals.

Analyzing the results by category produces a clearer picture of the participants' agreement with the algorithm and definition. In narratives predicted by the algorithm to require trust, 92.8% of responses agreed (out of 640 total responses from 128 different participants), over the course of all four runs (Figure 5 TR overall). Looking at the runs individually, for the trust matrices agreement ranged from 96.9% (run 4) to 87.5% (run 3). Tests for statistical significant across each pairwise combination of runs indicated only a single significant difference, $p < 0.01$, between runs 3 and 4. Thus, our conditions for trust consistently had a high degree of agreement with participants' selections.

Table 3 Representative Comments from Trust Matrix Participants

"He stood to lose $400 if he he [*sic*] trusted her and was wrong. Since he chose to work with her and put that money at risk, he must trust her."
"This is completely trust. He runs the risk of losing everything yet bets it all on her competence."
"This does indeed indicate trust. With Bob deciding to perform the action, he is putting trust in her that she will perform well. There is a lot at stake by performing the action with Alice. There is indeed risk."
"If she didn't, then she wouldn't run the risk of doubling the amount of time the action could take her. She clearly trusts him."

The participant's comments also tended to indicate that they recognized the connection between risk and trust in the narratives (Table 3). For example, when the trustor chooses to perform an action requiring trust, participants often commented that the trustor must believe that the trustee would mitigate the trustor's risk (our language). Likewise, when the trustor chooses not to perform the trusting action, participants noted that the trustor must have felt that he had a better chance on his own. Both of these

responses strongly agree with the definition of trust. There was little consensus in the comments of the 7.8% of responses that disagreed with the condition's prediction that the narrative requires trust.

There was slightly less agreement when the participants were presented with narratives deemed by the conditions not to require trust. Each run examined a different category of matrix that did not require trust. Figure 5 presents the results (EO, DI-1, DI-2, ID-1, ID-2, and INV). The percent agreement in the case of the no-trust matrices ranged from 83.3% (Equal Outcomes) to 35.9% (Trustor-Independent, Trustee-Dependent Matrices-Action 1 Rewarded). Hence, the type of no-trust matrix faced by the participant impacted one's agreement with the conditions. The percent agreement over all no-trust narratives was 66.4%.

For the equal outcomes category, there was 83.3% agreement with the conditions. Participants evaluating this category of narrative predominately confirmed that if all outcomes are equal then any decision made by the trustor did not require trust. A small minority of participants, however, indicated that performing any action with the trustee requires trust, even if there is no risk or reason for performing the action.

For the Trustor-Dependent, Trustee-Independent category of outcome matrices, the strength of the results depended on which action was rewarded. In matrices where the trusting action (a_1^i) produced a greater reward, 78.1% of responses agreed that the narrative did not involve trust. Yet, when the untrusting action (a_2^i) produced a greater reward, 63.3% of responses agreed with the no-trust prediction, in spite of the fact that these narratives violated the same conditions. Although both results agree with the hypothesis, we speculate that the decrease in agreement reflects the oddity of a narrative in which not trusting someone results in maximal reward. In the investment scenario, for instance, the trustor decides not to invest and the trustee does generate a poor return, yet the amount received by the trustor is maximal. Participant's comments indicate that this type of no-trust matrix caused some people to reason that the trustee must have performed better than the narrative indicated.

The Trustor-Independent, Trustee-Dependent category showed similar disparity. Narratives where the trustor received greater reward when the trustee violated the trust (a_2^{-i}) resulted in 69.5% of responses stating no-trust. Yet, only 35.9% of responses indicated no-trust when the trustor received greater reward if the trustee maintained trust (a_1^{-i}). According to the comments, trust can occur even if the trustor's choice has no bearing on the result. Many participants explained their answer by simply commenting "Bob trusts Alice's performance," "She is relying on Bob to perform well, whether she performs or not," and similar statements. The difference may have been compounded by the wording in the narratives. The a_1^{-i} action was referred to as "performs well" or "gives correct information" and a_2^{-i} was referred to as "performs poorly" or "gives incorrect information." The comments seemed to indicate that subjects had a difficult time imagining a scenario in which the trustee "performs poorly" and yet the trustor received the maximal reward.

For the Inverted Trust category, 62.0% of responses agreed that the narrative did not require trust. Some participants stated that they believed it to be impossible to trust an individual to perform an action in an unfaithful manner and thus trust is not possible in this situation.

Not surprisingly, we found that participants tended to invent reasons that explained the trustor's choice of actions. If the trustor performed an action that was against his benefit (according to the outcome matrix) or did not perform an action that would be to his benefit, participants occasionally invented stories to justify the person's behavior. For example one participant stated, "Bob uses Alice's information since he trusts the information enough to thoroughly finish in 120 minutes. He'd rather take the time to correctly finish something over finishing it fast."

With respect to the potential confounding factors, we found that results were not a reflection of the particular scenario as there was no statistical difference between the three scenarios, $\phi(1531) = -0.011, p > 0.05$. Further, the results were not impact by the gender of the trustor or trustee, $\phi(1536) = -0.017, p > 0.05$ or the magnitude of the outcome values $\phi(1536) = +0.030, p > 0.05$. Statistically

significant differences did result from positive and negative action labels; but the correlation of these labels was small and negative, $\phi(1536) = -0.104, p < 0.05$. We believe that the influence of positive and negative labels represents an experimental artifact the absence of which would have slightly strengthened our results.

The narrative experiment provides preliminary evidence that our conditions for situational trust correlate to the evaluations made by people. The results also show that this correlation is not limited to a single scenario. Still, the use of narratives forced participants to reason about the interactions of fictional third parties. For this reason, we conducted a follow-up experiment in which the participants had to decide whether or not they trust a robot.

3.3. Verifying the Conditions for Trust: Robot Guidance Experiment

We conducted a robot guidance experiment as a follow-up to the narrative experiment. The guidance experiment placed participants in a simulated maze and tasked them with finding an exit. This scenario was motivated by our interest in developing robots that can provide guidance during a fire. For this reason, we developed a maze that was roughly similar to an office environment. Participants are placed in this environment and then given the option of using a guidance robot to assist them with navigating the maze. Regardless of their choice, once they completed the navigation task, they were then asked whether or not they trusted the robot and if their decision to use or not use the robot showed that they trusted the robot.

We created two types of maze which were meant to correspond to the trust/no-trust matrices. In the trust matrix condition, the maze had several walls and barriers preventing participants from easily moving directly to the exit (Figure 6 top). The no-trust condition (Figure 6 bottom) was meant to correspond to the equal outcomes (EO) matrix from Table 1. The equal outcomes matrix reflects a risk-free situation in which the participant expects to receive the same outcome regardless of how either the robot or the person acts. Hence, for this condition, the exit was visible and directly in front of them, although it was located at a distance.

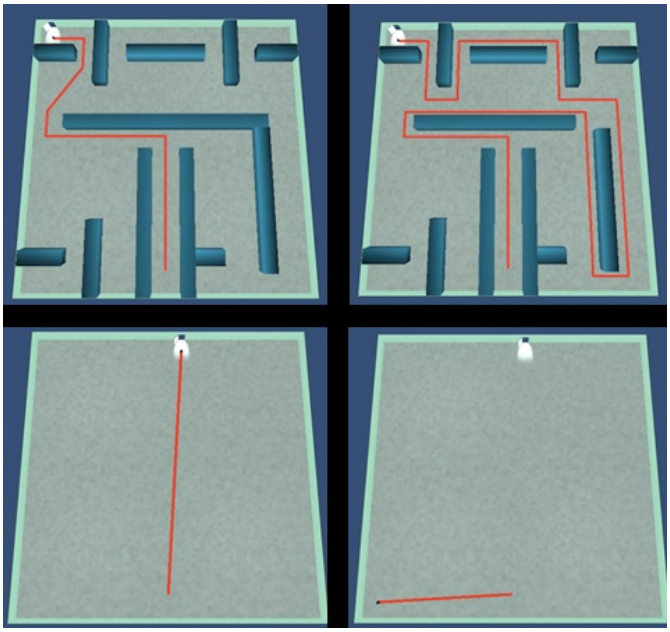


Figure 6 The top two images depict example mazes from the trust condition case and the bottom two images depict mazes for the no-trust condition. The images on the left illustrate the performance of good robots and the right images illustrate the performance of bad robots.

We hypothesized that self-reports of trust by the participants would correlate to the type of outcome matrix that the maze was based upon. Moreover, we hypothesized that the robot's performance (good versus bad) would not significantly impact the subject's trust self-report.

Again Amazon Mechanical Turk crowdworkers were utilized as participants. A total of 120 people completed the experiment. Blender, an open source 3D modeling software package, and the Unity Game Engine were used to create a simulated robot and maze environment (Figure 6) in which the crowdworkers were placed. Each maze had only a single exit. Participants performed the experiment using the Unity Web Plugin and a web browser. The robot that provided guidance was based on a Turtlebot robot but also had two arms to garner attention. Results from our prior research indicate that this style of robot communicates directions which are easily understood by people [9].

Each subject began the experiment by reading an introduction broadly describing the experiment and then consenting to participate. The introduction stated that we were testing how people leave buildings in emergencies and encouraged them to act as if they were in a real emergency. Extensive experimentation as part of our related research indicated that the use of this emergency scenario served as better motivation than the use of a monetary bonus [10]. Next, a short tutorial allowed the person to practice navigating a maze. They were then told that, because this was a simulated emergency, they would only have a certain amount of time to leave the building. No exact amount of time was provided. We noted that if they failed to locate the exit in time then their character would not be deemed to have survived.

In the trust condition, the pictures depicted example mazes along with survival rates. The survival rates for people that choose to use a good robot was listed as "typically survived," those that choose to use a bad robot stated as "typically did not survive," and did not choose to use the robot presented as "about half…survived." In the no-trust condition, all pictures presented a maze without barriers and with the exit clearly visible from the starting point. The survival rates in this condition noted that the person typically survived regardless of whether or not they choose to use the robot or whether the robot was good. Again this condition was meant to be risk-free. Figure 7 presents the introduction screen, the examples screen for the trust and no-trust cases, and the survey page.

Next the participants were asked to choose whether or not they wanted to use the robot for guidance by pressing a button before the experiment started. In the no-trust case they were told that the maze would be the same as the one presented in the examples. In the trust case, they were informed that the maze would be different from the examples.

Once the button was pressed the software placed the participant in the maze. If they choose to use the robot, the participant was spawned in the maze with the guidance robot nearby. The robot would begin to move as soon as the participant moved. The participant could choose to navigate the maze with or without the robot's help. They were given 60 seconds to navigate the maze. Their remaining time was prominently displayed in the center of their screen (Figure 8). In both conditions, half of the robots provided good guidance and half provided poor guidance. Whether or not the participant pressed the button requesting the robot's guidance was recorded as their decision. We have not examined the extent to which participants actually followed the robot. Results from prior research indicate that subjects who choose to use a robot for guidance tend to continue to follow it regardless of its performance [10].

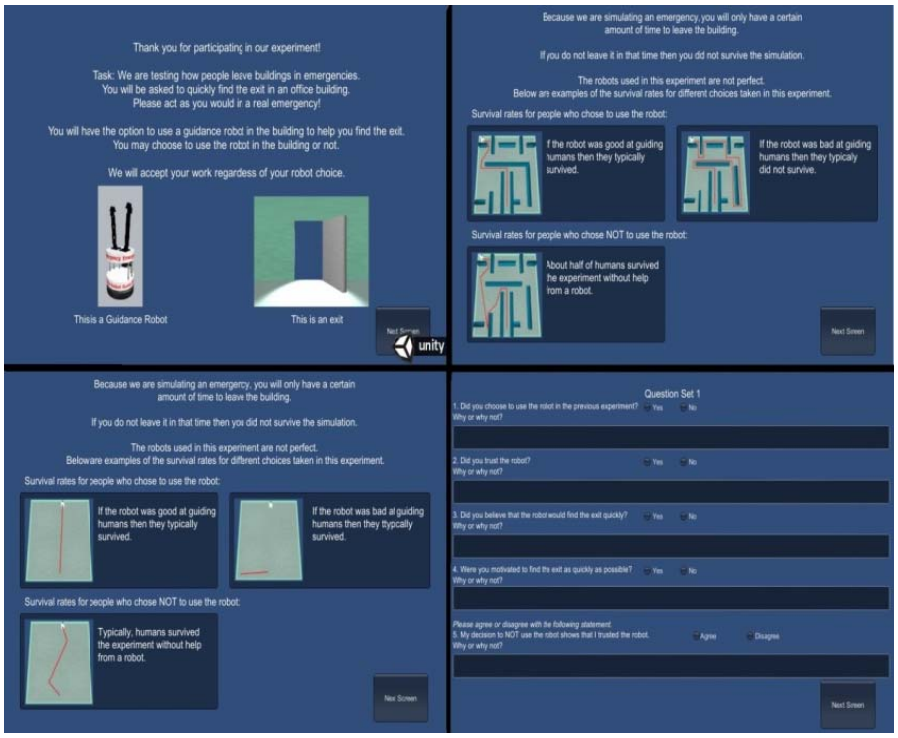


Figure 7 Screenshots depicting the guidance experiment's introduction screen (top left), trust condition example screen (top right), no-trust condition example screen (bottom left), and survey (bottom right). A participant would see either the trust condition examples or the no-trust condition example, but not both.

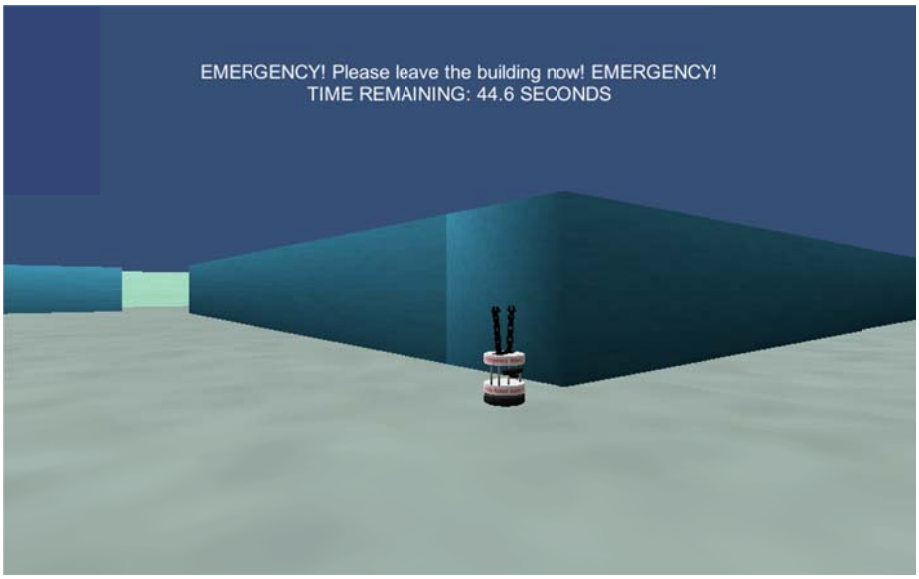


Figure 8 A screenshot of the view of the environment that participants saw in the emergency condition.

Upon completing the maze participants were asked to take a short survey (Figure 7). Two questions focused on trust. The first trust-related question asked, "Did you trust the robot?" The second question asked participants whether they agreed or disagreed with the statement, "My decision to use the robot shows that I trusted the robot." If the participant had chosen not to use the robot then the word "NOT" is inserted before the word "use." Finally, a second set of survey questions collected demographic information.

3.3.1. Results

A total of 120 participants (mean age = 31.25, 40% female, 94.1% United States nationality) completed the experiment. The results from the experiment indicate a strong positive correlation between participant's trust self-report and the predictions of the situational trust algorithm as to whether or not the maze required trust, $\phi(120) = +0.406, p < 0.001$ (Hemphill, 2003). The correlation is not as strong as in the narrative experiment (where $\phi = +0.592$). Our conditions for trust were met when the participants were presented with a trust maze and choose to use the robot. This occurred for 50 of the 120 subjects. When these conditions were met $74.0\% \pm 12.2\%$ of participants reported trust (Figure 9). For 70 of the 120 participants the conditions for trust we not met because either they were presented with a no-trust maze or they choose not to use the robot. In this case only $32.9\% \pm 11.0\%$ reported trust. This difference is statistically significant, $\chi^2(1, 120) = 6.53, p < 0.001$. The results were nearly identical regardless of which of the two self-report questions were used for the analysis. Overall, 76.7% of participants chose to follow use the robot for guidance.

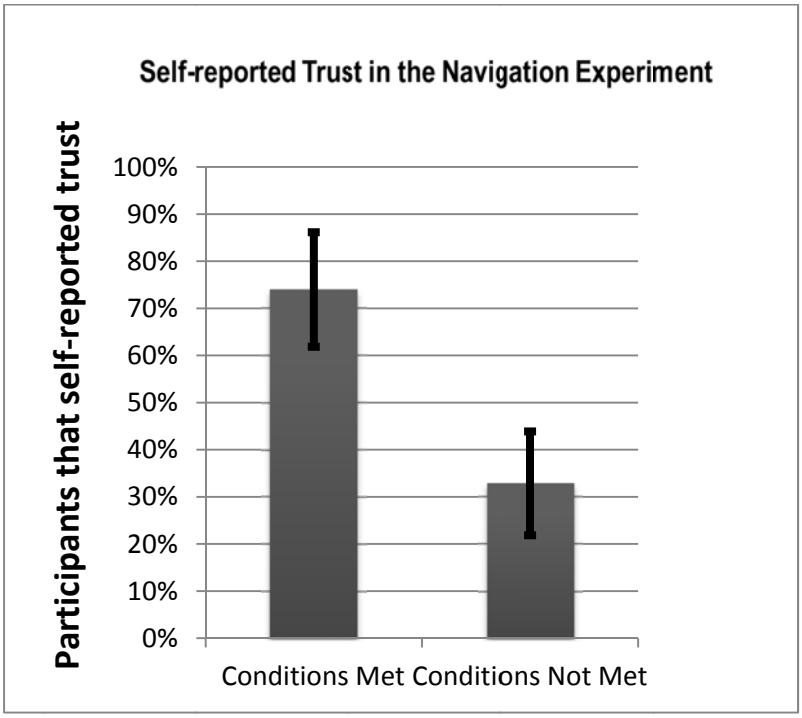


Figure 9 The results from the robot guidance experiment are depicted above. When the conditions for trust are met, participants were significantly more likely to self-report trust.

One potential confounding variable is the quality of guidance provided by the robot. Yet our overall results indicated that the quality of guidance did not significantly correlate to participant's reports of trust, $\phi(120) = +0.067$, $\chi^2(1, 120) = 0.53$, $p = 0.47$. This experiment thus provides additional evidence that our conditions for situational trust do correlate to the evaluations made by people.

3.4. Year 1 Conclusions

The results from the experiments support our hypothesis that outcome matrices which meet the conditions for trust are also deemed to require trust by people. We found this to be true regardless of whether the subjects are reading narratives about the actions of others or selecting behaviors with a robotic teammate. Still, the results from the narrative experiment indicated a stronger correlation (+0.592) than the results from the guidance experiment (+0.233). Part of the reason for this difference may have been a social desirability bias present in the guidance experiment. Social desirability bias is a subject's tendency to

respond in a manner which is socially desirable [11]. Social desirability may have influenced subjects to report that they trusted the robot regardless of how they actually felt. Consider that nine subjects reported trusting the robot even though it headed in a direction that was clearly away from a visible exit and then stopped moving (Figure 6 bottom right depicts its path).

The experiments focused on the impact of the situation on trust. We have attempted to rule out the influence of several potential confounding variables such as the quality of guidance by the robot, the gender of the agents, the action selected by the trustee, and the context of the narrative (investment, navigation, employment related). The data indicates that these factors were not responsible for the results. We therefore conclude that our results support our contention that facets of the situation, namely risk, strongly influence the trust decisions of human subjects. Moreover, this situational trust can be captured in a series of conditions implementable on a robot.

The first year of this article has investigated a set of conditions for determining if an interactive situation demands trust. Our focus was to evaluate the extent to which these conditions correlate to the classifications made by people. Our motivation for doing so was to develop a general conceptualization of trust which could be used to guide the behavior of a robot. Overall, our data supports the contention that perceived risk is central to the trust phenomenon, even for interactions involving a robot. These results are in agreement with both the trust community [12, 13] and the robotics community [14].

We attempted to ensure the internal validity of the experiments by randomly selecting subjects, using control groups, addressing potential confounding variables, and limiting the potential for experimenter bias. The external validity of our results, however, is limited. Because the experimental environment was simulated and participants completed the task online, the study lacked the true visceral reaction of a real emergency. Additional real-world studies were thus warranted. Further, although we tried to use a variety of different situations and contexts, fully capturing the many different environments in which trust occurs was impossible. The conclusions drawn are necessarily based on a limited number of situations. Finally, although the study's participants were from a broad cross-section of the United States, they still represent a limited population. Hence, we are not claiming that the results represent a general truism about trust, rather, only that they serve as evidence of the connection and validity of our approach for conceptualizing and reasoning about trust with respect to outcome matrices.

4 Year 2 Accomplishments

Results from the first year indicated a strong agreement between the predictions of our conditions for trust and the evaluations of human subjects. These original experiments were conducted using a narrative description of an interactive scenario that did not involve use of a robot. In the second year we focused on a human-robot interaction scenario in which a robot offers to provide guidance to a person during an emergency. This scenario is motivated by recent events involving fires in which people were trapped because of crowding and is DOD relevant as evidenced by ONR's SAFFiR project involving the development of an autonomous shipboard firefighting robot.

4.1. Trust during Simulated Emergency Evacuations

The emergency guidance experiments that we conducted during this second year placed people in a maze under time pressure to find an exit. These experiments required a person to navigate a simulated maze with or without the help of a robot. In order to examine the impact that a robot's initial performance has on later decisions involving trust, the person was required to navigate a different maze in two separate rounds. They were given the option to use a guidance robot prior to navigating both mazes. Data reflecting their decision to use or not use a robot as well as surveys focused on the participant's reasoning were collected as data.

The decision to trust the robot during these experiments was a binary decision. The person either allowed the robot to provide guidance or did not. The decision to use the robot was viewed as an indicator of trust. We also measured trust by asking participants to self-report whether or not they agree with the statement: "I trusted the robot when I made my choice to follow or not follow the robot." In addition to the options to agree or disagree, we also gave an option labeled "Trust was not involved in my decision." In pilot studies, we found that some participants interpret a disagreement to the statement to mean that they actively distrusted the robot, hence we provided a third option that clearly indicates they neither trust nor distrust the robot. The results therefore focused on affirmations of trust.

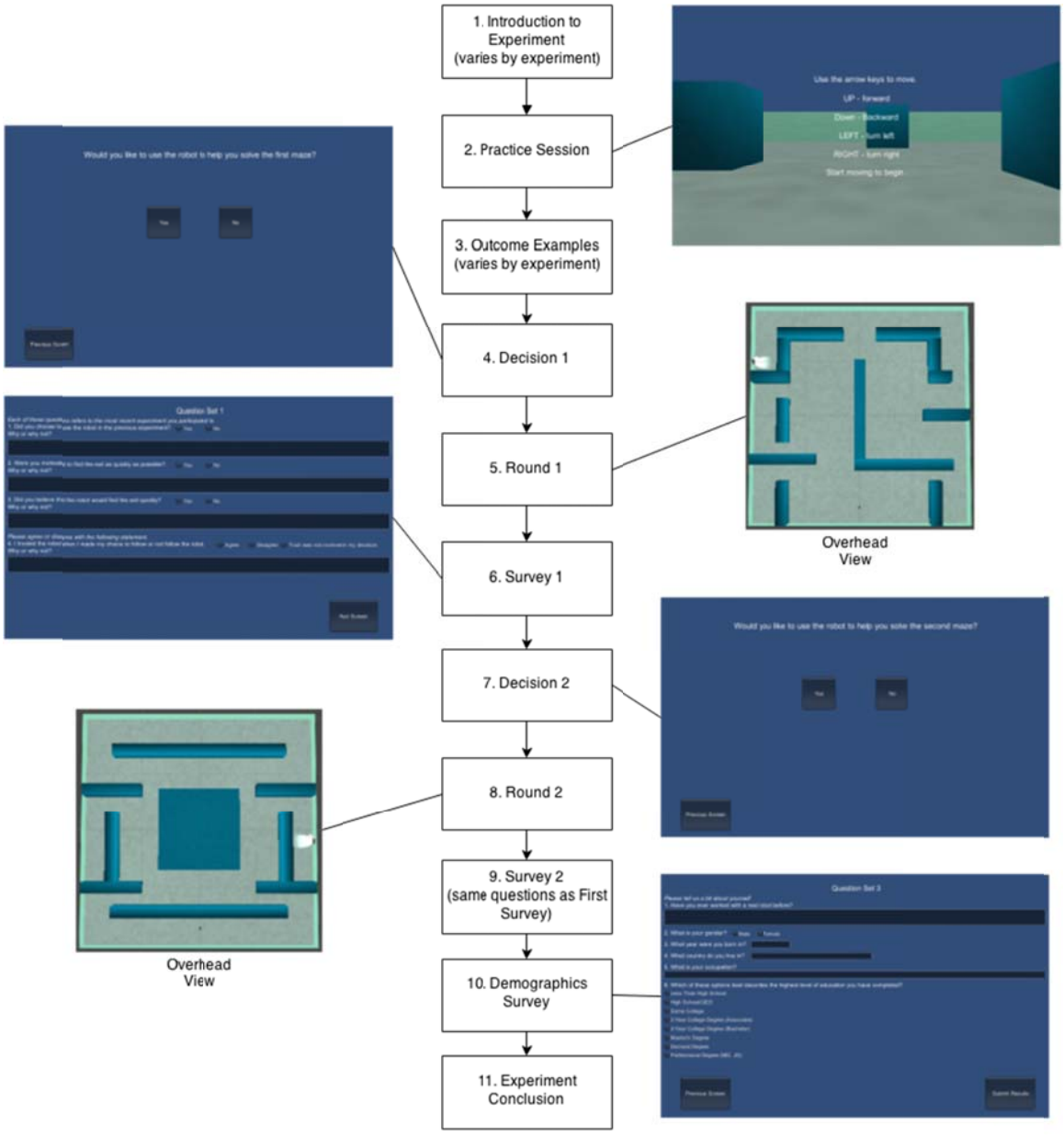


Figure 10 Experimental protocol with screenshots from experiment. The entire experiment was presented in a Unity 3D web game, including the survey questions.

The simulation environment was created to slightly resemble an office building and included corridors and rooms designed to give it a maze-like appearance. Participants were placed in the environment with no previous experience and required to find a single exit.

The general experimental setup is depicted in Figure 10. Participants began by accepting the request on Mechanical Turk and clicking a link to a Unity 3D Web Player executable. Some participants had to download the Unity Web Player plugin to perform the experiment. Next they viewed an introductory message that described the navigation task they were to perform. This page included photos of an exit and the guidance robot. The guidance robot varied in the two experiments. They were then offered the opportunity to practice navigating in a maze. They had a first-person view of the maze and used their keyboard arrow keys to move. After the practice session, they were presented with illustrative examples of prior human-robot performances in the maze. The nature of these examples varied with respect to the experiment. The participant was then asked to decide whether or not they would like a robot to provide guidance during the first round of the experiment. After making their choice the person then navigated the maze and completed a short survey (Table 4). They were then offered another opportunity to decide if they wanted to use the guidance robot in the second round. They then navigated the maze in the second round and completed a short survey about their second round decision. The robot's guidance performance in the second round always matched its performance in the first round. The experiment concluded with a final survey that collected demographic information about participant age, gender, country of residence, occupation, and education level. This survey also asked if participants have worked with a real robot before.

Table 4: The survey presented to participants after each round. This survey gathered qualitative and quantitative information about a participant's trust in the robot as well as comments on their decision to use the robot or not. The first four questions had yes or no options available as answers. The fifth question had agree, disagree, and trust was not involved in my decision as options. All questions had space for explanations.

Question
1. Did you choose to use the robot in the previous experiment?
2. Did you trust the robot?
3. Did you believe that the robot would find the exit quickly?
4. Were you motivated to find the exit as quickly as possible?
5. My decision to use the robot shows that I trusted the robot.

The quality of robot guidance (efficient, circuitous, incorrect) was an independent variable studied in both experiments. The dependent variables for both experiments were 1) the participant's decision to follow or not follow the robot and 2) the participant's self-reported trust. Data on these variables was collected after each round for all participants.

4.2. Robot Failures during Emergency Evacuations

We believed that the robot's behavior during the first round would inform the person of the robot's ability and serve as a source of information for future rounds. We therefore examined how the robot's behavior affects the participants' self-reports of trust in the second round. We also investigated different types of robot guidance failures: one that inefficiently leads the person to the exit and one that fails entirely to lead the person to the exit. In pilot studies we evaluated several different types of robot guidance failures. All but two of these failure modes were eliminated because participants were unable to determine that the robot had failed and hence resulting in an extremely long experiment completion time (see Table 5 for a listing of the robot guidance failure types that were not included in the experiments). Overall, three robot behaviors were defined that were used in the experiments:

- Efficient navigation: the robot proceeds directly to the exit location (Figure 11). Robots that acted in this manner are capable of finding the exit within thirty seconds.
- Circuitous navigation: the robot explores many possible routes before eventually finding the exit (Figure 12). Robots that acted in this manner are capable of finding the exit in ninety seconds.
- Incorrect navigation: the robot proceeds directly to a corner of the environment that is not the exit location and then stops. This is meant to emulate the behavior of a robot that has incorrect

information about the exit location. Robots that acted in this manner stopped moving after approximately thirty seconds at a point at least thirty seconds from the exit.

Table 5: Poor robot guidance behaviors that were explored as part of a pilot experiment.

Name	Description	Reason for Exclusion
Small Loops	Robot circled an obstacle continuously	Several loops around the obstacle were required before participants realized the robot had failed. The total time for the experiment was too long.
Large Loops	Robot circled a large area of the environment continuously	Participants could not realize that the robot had failed until it completed at least one loop. This could take several minutes by itself and thus the total time for the experiment was too long.
Continuous Searching	Robot searched through entire environment except location of actual goal position. After completing a search it started again.	Participants followed the robot for considerable time before realizing the robot had failed. Some participants would follow the robot for more than 15 minutes.
Wall Collision	Robot nearly found goal but then continuously collided with wall and was unable to proceed.	Participants did not understand that the robot was colliding with the wall and thus did not understand that it failed.

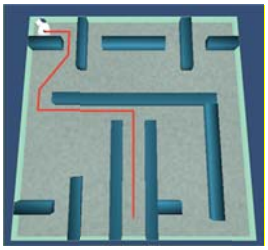


Figure 11: Example of efficient robot behavior. This is the behavior of a robot that knows exactly where the exit is and can thus effectively mitigate the participant's risk in both experiments. The incorrect robot had a similar path but to a corner of the environment that was clearly not the exit.

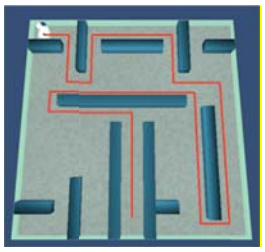


Figure 12: Example of circuitous robot behavior. This is the behavior of a robot that has to search for the exit but still finds it at the end of its search. In some cases it mitigated a portion of the risk for the participant but in most cases it did not.

In each of the behaviors above, the robot followed a predefined set of waypoints throughout the environment. Waypoints were set near corners or occlusion points so that each was in view of the waypoint before it. The robot waited at each waypoint for the participant to approach before it moved to the next waypoint. The robot was allowed to move considerably faster than the participant so that it would always be leading. The exact time to reach their end points depended on the particular environment and on the participant.

Our experimental setup allows us to measure trust in two ways: as 1) the participant's decision to use or not use the robot during the second round; and 2) participant's agreement or disagreement with the statement "I trusted the robot when I made my choice to follow or not follow the robot."

4.3. Experiment One: Bonus Motivation

The first experiment using the emergency evacuation paradigm used all three robot behaviors and both trust metrics. Successful and unsuccessful robots were used to determine whether participants would continue to trust a robot that performs well more often than a robot that performs poorly. The two unsuccessful robot behaviors (circuitous navigation and incorrect navigation) were included to determine if there was a difference in response. Both of these questions were tested as a between subjects experiment with the robot behavior as the independent variable so no participant saw more than one robot behavior type. Our dependent variable was participant's self-reported trust and their decision whether or not to follow the robot. We expected that the two metrics would have a high positive correlation.

For this experiment, and in keeping with the prevailing literature on trust research [15], we used monetary risk to motivate participants' trust decisions. Subjects were thus offered a $1 bonus if they could find the exit of a maze within 30 seconds. After the first 30 seconds had elapsed the bonus began to decrease. Ninety seconds after the start of the experiment the bonus was $0. Participants were informed that their choice to use a guidance robot or not would not directly affect their bonus in any way.

4.3.1. Experimental Setup

As noted above as well as in Figure 13 and Figure 14, the first stages of the experiment, the Introduction and Example Outcomes sections, were unique for each experiment. Each reflected the specific, monetary motivation in this experiment. We included one additional survey in this experiment to help us understand the motivations of participants on Mechanical Turk. Participants were asked to rate their motivations with respect to time, money, and enjoyment on a seven point Likert scale. They were then asked to rank the three options from most important to least important. The questions are unrelated to the hypotheses or research question and the survey was only included to help us design better experiments.

The first screen seen by the participants gave instructions for the simulation. The simulated environments were specifically referred to as "mazes" to give the participant an idea of their complexity and goal. The robot displayed during the introduction and used in the rounds was a Willow Garage TurtleBot 2. The 3D model of the robot was created out of CAD files distributed by the manufacturer. The practice session proceeded as described in the Methodology section.

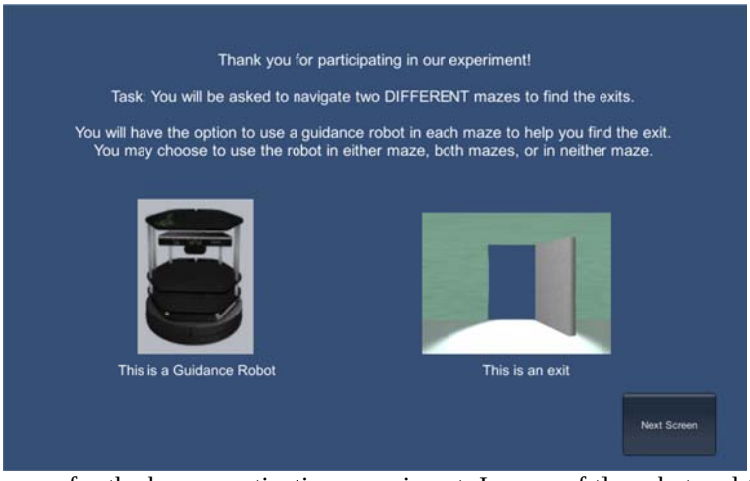


Figure 13: Introduction screen for the bonus motivation experiment. Images of the robot and the exit were used so that participants could gain familiarity with the simulation.

After the practice session, the participants were informed of the performance-based bonus and how to obtain it. Participants were given three example performances for the navigation task:

> Example 1: The text "People who used a robot that quickly found the exit typically earned a bonus of about $1.00" accompanied by a top-down view of a direct path to the exit in an example maze.

Example 2: The text “People who used a robot that did not quickly find the exit typically earned a bonus of about $0.00” accompanied with a top-down view of a very indirect path to the exit in the same example maze.

Example 3: The text “People who did not use a robot typically earned a bonus of about $0.50” accompanied with a top-down view of an indirect path to the exit in the example maze.

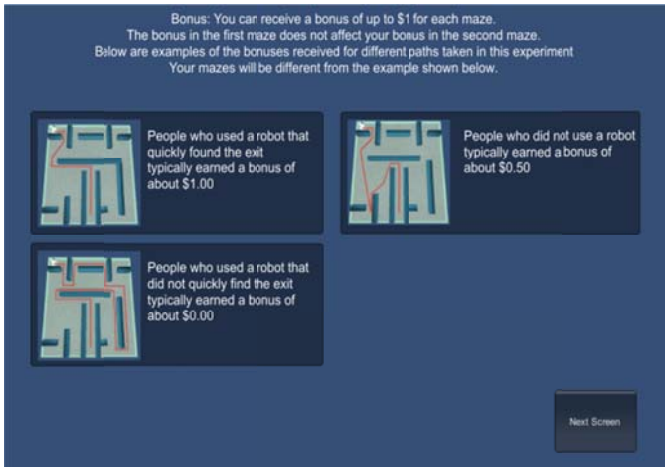


Figure 14: Example outcomes for the bonus motivation experiment. Participants were shown overhead views of an example environment with a path drawn in each so that they could understand the complexity of the environments as well as observe a visual representation of successful and unsuccessful robot behavior.

The participants were then asked whether they would like to use a robot in the first round or not. At the start of each round, the participants were reminded of the controls for moving in the environment and informed that their bonus was currently set at $1.00 (Figure 15). When the participant began moving a timer in the top left of the screen started increasing, displaying the time spent navigating to a tenth of a second precision. The bonus was prominently displayed in the top right corner. After thirty seconds of navigating the maze, the bonus began to decrease at a rate of $0.0167 per second (Figure 16). This meant that the bonus was completely depleted after ninety seconds. The second round was setup the same as the first but with a different maze. All other aspects of this experiment proceeded as described above.

Figure 15: The image above presents the beginning of the first round of the bonus motivation experiment. Round 2 had the same text but took place in a different environment. Participants are reminded of the controls for the simulation, updated on the current amount of their bonus, and shown the time elapsed since the round began.

Figure 16: The figure presents a successful ending to round 1, albeit with a small bonus. This was typical of subjects being guided by a circuitous robot.

Because participants had no control over the amount of bonus they earned; they were all paid the full $2.00 bonus after their experiment was completed. This information was not made available to any participant before they performed the experiment.

4.3.2. Results

A total of 106 participants (mean age=31.0, 60.4% male) completed the experiment, 84.9% of which chose to follow the robot in the first round, with no prior knowledge of the robot's behavior. Figure 17 depicts the number of participants who used the robot in rounds 1 and 2 for the efficient and circuitous/incorrect robot behaviors and the self-reported trust in rounds 1 and 2 for the different robot behaviors. Only participants who chose to follow the robot in round 1 are reported. As can be seen in the figure, self-reported trust decreases significantly (53%, $\chi^2(1, N = 90) = 12.86, p < 0.001$) when the participants experience a circuitous or incorrect robot in the first round. Only a 4% ($\chi^2(1, 90) = 1.87, p = 0.172$) decrease in trust was reported by participants that were guided by an efficient robot. Figure 18 shows the results for the different failure modes. The type of robot failure had no impact on either the self-reported trust (0% difference) or the decision to follow (0% difference). In both the first and second round a strong positive correlation was found between following the robot and reporting trust in the robot, $\phi(106) = +0.628$ for round 1 and, $\phi(90) = +0.422$ for round 2.

We examined the survey comments to better understand each participant's rationale. Table 6 summarizes the most common comments from round 2. Note that, of the people that were guided by a circuitous or incorrect robot, many choose to follow the robot in the second round because they believed that the robot's help was better than no help at all (n=7) or they thought that the robot would perform better this time (n=5). These comments hint that participants were deciding to follow the robot in spite of the loss of bonus.

Table 6: Summary of comments from the Bonus Motivation Experiment.

Robot Behavior	Used Robot?	Self-reported trust	Comment Description
Efficient (n=30)	Yes (n=25)	Positive (n=22)	Robot performed well (n=21)
			Did not trust robot, trusted programmers (n=1)
		Negative/Neutral (n=3)	Impossible to trust machine (n=1)
			Trusted robot initially but explored on own instead of completing maze (n=1)
			More than two examples required to trust something (n=1)
	No (n=5)	Positive (n=2)	No complaint about robot, wanted to try experiment for themselves (n=2)
		Negative/Neutral	No complaint about robot, wanted to try

		(n=3)	experiment for themselves (n=1)
			Thought robot would perform worse in second round (n=1)
Circuitous (n=30)	Yes (n=21)	Positive (n=11)	Robot performed better than human alone (n=7)
			Did not realize robot performed poorly (n=3)
			Thought robot would perform better in second round (n=1)
		Negative/Neutral (n=10)	Curiosity (n=6)
			Robot performed better than human alone (n=1)
	No (n=9)	Positive (n=1)	No complaint about robot, wanted to try experiment for themselves (n=1)
		Negative/Neutral (n=8)	Robot performed poorly (n=7)
			No complaint about robot, wanted to try experiment for themselves (n=1)
Incorrect (n=30)	Yes (n=21)	Positive (n=11)	Thought robot would perform better in second round (n=5)
			Did not realize robot performed poorly (n=3)
			Curiosity (n=3)
		Negative/Neutral (n=10)	Curiosity (n=6)
			Robot performed better than human alone (n=1)
	No (n=9)	Positive (n=1)	*Unclear response* (n=1)
		Negative/Neutral (n=8)	Robot performed poorly (n=8)

We performed an analysis on our motivational survey to better understand the participants. About half of participants (55) reported that their most important motivation with respect to the experiment was money. The rest were evenly divided between time (25) and fun (24). These results indicate that participants are not solely motivated by simple monetary bonuses in the experiment. Hence, some choose to follow the robot in the second round in spite of its failure and the fact that they self-reported not trusting it because they believed it would ultimately be faster or more fun to follow the robot.

Overall, the results indicate a much larger decrease in self-reported trust when the robot fails compared to when the robot does not fail. This result shows that only a single failure can strongly and quickly influence a person's trust in the robot, which may have ramifications on the testing and evaluation of such systems. It is also noteworthy that the majority of people (84.9%) chose to follow the robot initially. This result appears to imply that people tend to trust initially.

The results also show that a robot that fails by traveling a short distance and stopping does not have a significantly larger negative impact on both self-reported trust and the subsequent decision to follow than a robot that merely slowly led to the exit. Our results indicate that either failure impacted trust equally both with respect to the self-report and subsequent following. The fact that there was 0% difference in both cases is presumably a statistical artifact. The result is intriguing. Does it indicate that the reduction in trust is only a reflection of good versus bad performance and not, as we expected, mitigated by the way in which a robot failed? This could be an area for significant further research.

Final the data can be used to examine the relationship between self-reported trust and the decision to follow the robot. We predicted that both the participant's trust and the likelihood of following would be strongly correlated. Indeed, we found a strong positive correlation between following the robot and self-reporting trust. Yet, the results show that numerous participants (26% of all participants) choose to follow the robot in the second round even though they reported not trusting it. This contradicts our own intuition as well as related work, such as [14] and [16], who found that operators typically stopped using

autonomous modes on robots that performed poorly. As discussed in the section that follows, this discrepancy motivated us to continue to refine our experimental methodology.

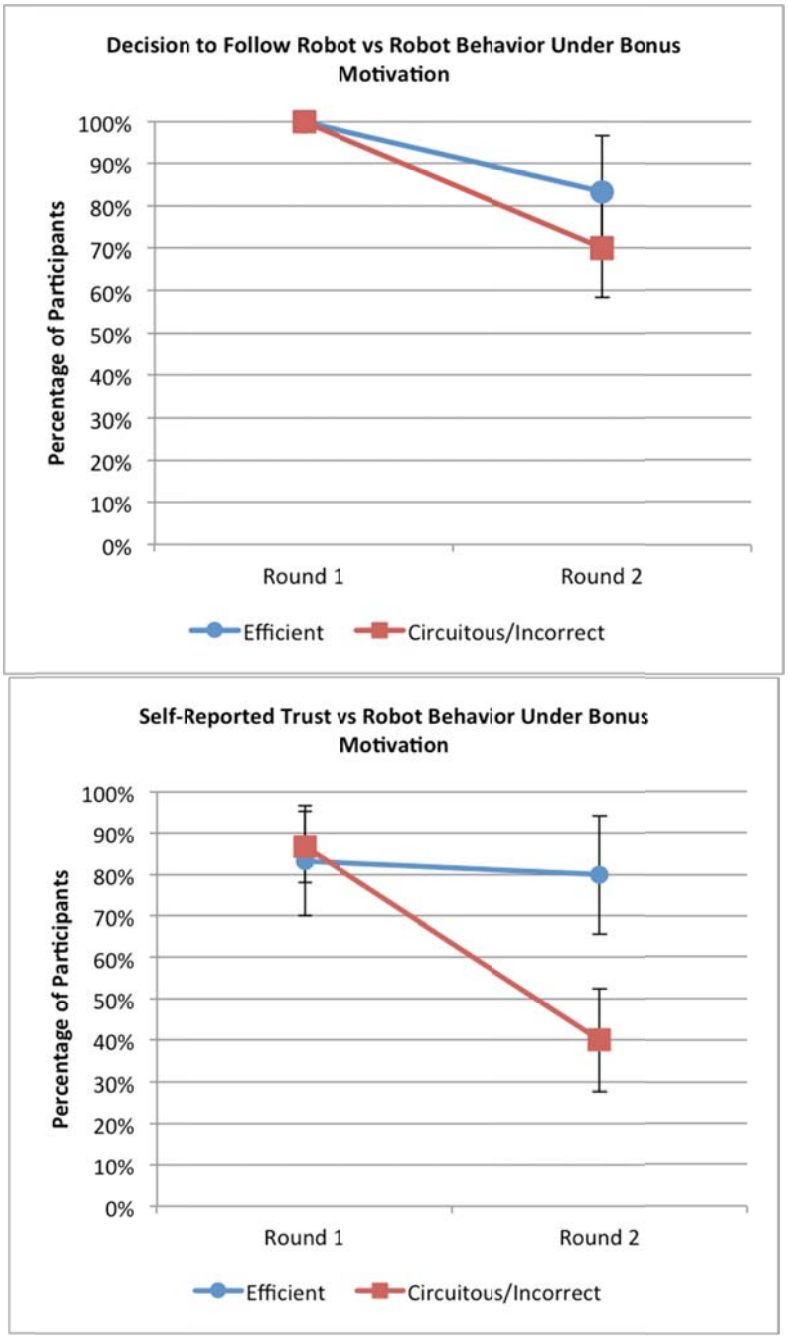


Figure 17: Change in decision to use robot (top) and self-reported trust (bottom) between the two rounds for the successful and unsuccessful robots. Note that a majority of participants continued to use the circuitous/incorrect robots even though half had lost their trust in the robot. Error bars represent 95% confidence intervals.

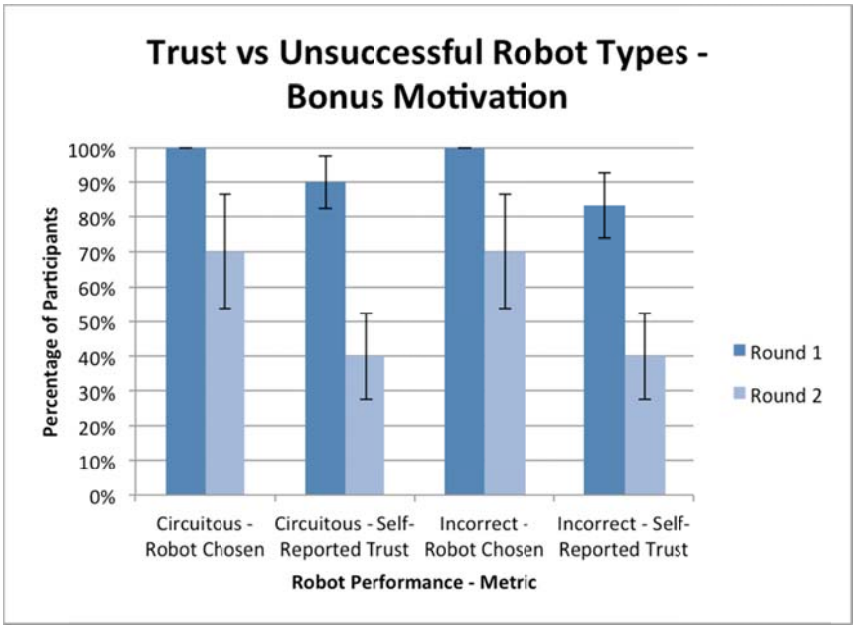


Figure 18: Change in decision to use robot and self-reported trust between the two rounds for the circuitous and incorrect robots. The same number of participants chose to use each and the same number reported trust in each in the second round. Error bars represent 95% confidence intervals.

4.4. Experiment Two: Emergency Motivation

Risk is a major component of our definition of trust [17]. Characteristics of the experimental scenario can influence a subject's perceived risk differently. For example, the risk associated by losing $10 gambling will likely impact the behavior of people near poverty more than wealthy people. From an empirical point of view, we would like to control the factors that influence the subject's perceived risk. Monetary incentives are a commonly used incentive used for trust research [15, 18, 19]. Yet our data indicates that people may follow an untrustworthy robot because they are motivated to merely complete the study quickly.

Survey comments led us to develop a second experiment that sought to better align the participants' motivations with the task goals. This second experiment asked participants to act as if they were in an emergency. Instead of receiving a bonus, a quick exit from the building rewarded them with "survival." Thus, instead of a monetary risk, participants experienced a survival risk.

4.3.3. Experimental Setup

For this experiment, participants were told that our goal was to discover how people leave a building in an emergency. Instead of receiving a bonus for a fast completion, they were told that they would only survive if they found the exit in time. During the rounds, a countdown timer appeared in the middle of their view to tell them the remaining time. As with the previous experiment, this study was conducted using the Unity simulation and Amazon's Mechanical Turk. Participants were compensated $2.00 for their participation in this experiment.

There were several difference between this experiment and the bonus motivation experiment. First, participants were informed by the introduction screen "We are testing how people leave buildings in emergencies" and asked to "Please act as you would in a real emergency!" (Figure 19). The word "building" was used instead of "maze" to further reinforce the emergency portion of the simulation.

The robot in this experiment was a TurtleBot 2 modified with two PhantomX Pincher AX-12 arms to allow it to gesture. The robot was also given signage to indicate that it is an emergency evacuation robot. The arms waved while it moved to attract attention. The robot's appearance and gestures were evaluated

in a previous paper and it was found that participants understood it better than other forms of evacuation robots [9].

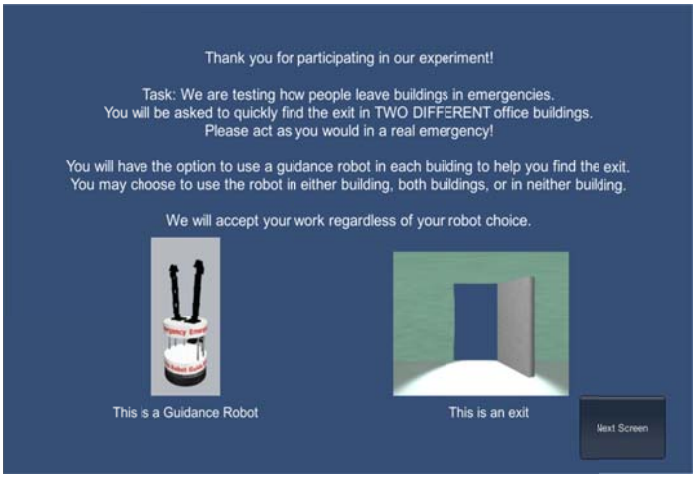


Figure 19: Introduction to emergency motivation experiment. Note that the robot is different from in Experiment 1. Additionally, participants were told that this experiment was to determine how humans evacuate buildings, not how humans interact with robots.

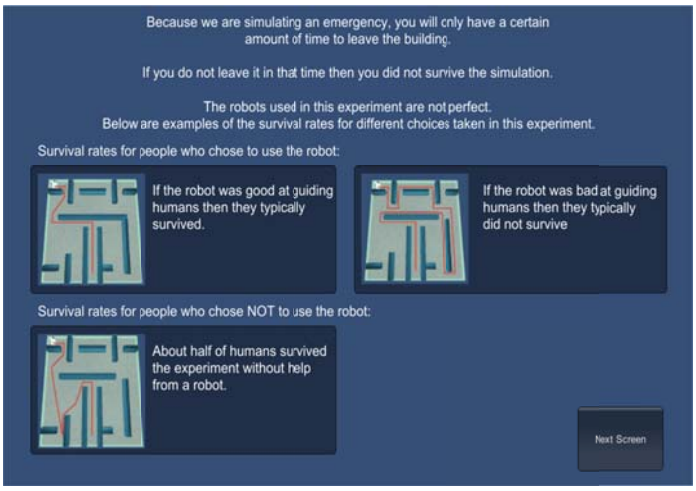


Figure 20: Example outcomes in emergency motivation experiment. Again, participants were shown overhead views of the example environment with a variety of paths, but this time they were presented with survival possibilities, not monetary rewards.

For this experiment each round ended after 60 seconds regardless of the participant's ability to find the exit. Once again, before selecting whether or not to use the robot, the participant was presented with a series of example experimental outcomes. The examples reflected the change to an emergency scenario (See Figure 20). The examples were:

Example 1: The text "If the robot was good at guiding humans then they typically survived." accompanied by a top-down view of a direct path to the exit in an example maze.

Example 2: The text "If the robot was bad at guiding humans then they typically did not survive" accompanied with a top-down view of a very indirect path to the exit in the same example maze.

Example 3: The text "About half of humans survived the experiment without help from a robot." accompanied with a top-down view of an indirect path to the exit in the example maze.

The words "EMERGENCY! Please leave the building now! EMERGENCY!" as well as the time remaining to exit (to a tenth of a second precision) appeared in the top-center of the participants' view

throughout the entire round. See Figure 21 for an example of the beginning, Figure 22 for an example of a successful ending and Figure 23 for an example of an unsuccessful ending to the experiment.

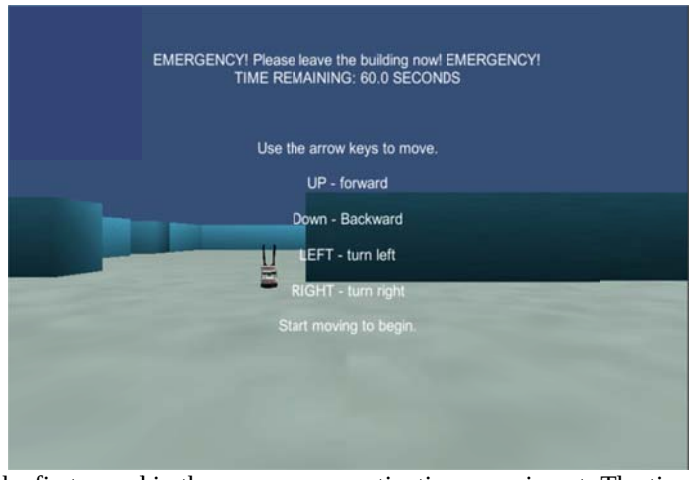


Figure 21: Beginning of the first round in the emergency motivation experiment. The timer counted down instead of up in this experiment and was moved to the center of the screen for maximum visibility. Text to indicate that this is an emergency was also placed on the screen.

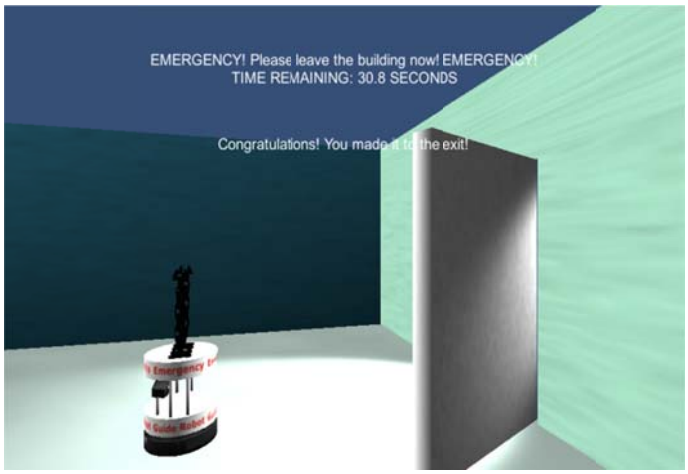


Figure 22: An example of a successful exit in the emergency motivation experiment. This typically happened only in the efficient robot conditions.

Figure 23: An example of an unsuccessful exit in the emergency motivation experiment. Text informed the participant there was no time remaining. The robot can be seen in the distance.

Outside of these changes, the experiment was identical to the bonus motivation experiment. Participants were again required to complete the same survey examining their trust in the robot and reasoning for choosing the robot. This experiment also consisted of two rounds.

4.3.4. Results

A total of 129 participants (mean age=31.8, 60.5% male) completed the second experiment, 69.8% of which decided to use the robot in the first round. As shown in Figure 24, the decision to follow the robot decreases by 50% in the second round when the participant interacts with a circuitous/incorrect robot in the first round, compared to just 3% when an efficient robot is used first ($\chi^2(1, N = 90) = 19.29, p < 0.001$). Self-reported trust follows a similar trend with trust decreasing 53% when participants experienced a circuitous/incorrect robot and self-reported trust increasing by 3% ($\chi^2(1, N = 90) = 24.31, p < 0.001$). Figure 24 shows the results for the different failure modes. The type of failure had minimal impact in the participant's decision to follow ($\chi^2(1, N = 60) = 0.27, p = 0.606$). There was also a minimal change in self-reported trust $\chi^2(1, N = 60) = 1.15, p < 0.284$). A strong positive correlation was found between choosing to use the robot and reporting trust in the robot in both rounds: $\phi(129) = +0.661$ for round 1 and $\phi(90) = +0.745$ for round 2.

Again, motivations for participants' actions and reports can be found in the comments. A short description of a selection of these comments can be found in Table 7. Note that not all participants' comments are included in this table for brevity and some participants gave multiple reasons for their actions.

Table 7: Summary of comments from the Emergency Evacuation Experiment

Robot Beh.	Follow Dec.	Trust Answer	Comment Description
Efficient (n=30)	Yes (n=29)	Positive (n=27)	Robot performed well (n=24)
		Negative/ Neutral (n=2)	Logical choice, not trust (n=1)
			Decided to proceed on own for fun after choosing to use robot (n=1)
	No (n=1)	Positive (n=0)	
		Negative/ Neutral (n=1)	Thought robot would perform worse in second round (n=1)
Circuitous (n=30)	Yes (n=15)	Positive (n=12)	Curiosity (n=5)
			Thought robot would perform better in second round (n=3)
			Robot moved quickly, and thus was trustworthy (n=2)
			Did not realize robot performed poorly (n=2)
		Negative/ Neutral (n=3)	Curiosity (n=3)
	No (n=15)	Positive (n=1)	Trusted robot to NOT find exit (n=1)
		Negative/ Neutral (n=14)	Robot performed poorly (n=13)
			No complaint about robot, wanted to try experiment for themselves (n=2)
Incorrect (n=30)	Yes (n=16)	Positive (n=9)	Robot performed better than human alone (n=6)

			Thought robot would perform better in second round (n=3)
		Negative/ Neutral (n=7)	Curiosity (n=5)
			Robot performed better than human alone (n=2)
	No (n=14)	Positive (n=0)	
		Negative/ Neutral (n=14)	Robot performed poorly (n=12)
			No complaint about robot, wanted to try experiment for themselves (n=2)

A single failure of a robot caused 50% of participants to stop using the robot in the second round, compared to just a 3% drop with a successful robot. Our data indicates that a smaller percentage of participants chose to use the robot in the first round when compared to the first round of the Bonus Motivation Experiment. While a majority still chose to use the robot, and thus our findings from previous work are still supported, we did not expect such a change. Many participants justified their choice by stating that they did not want to put their life in the hands of a machine. This indicates that people are more likely to initially trust a robot when there is a lower risk (e.g. a financial risk instead of a survival risk). This data serves as evidence that people take the emergency scenario, and the risk it entails, seriously. With respect to the type of robot failure, both experiments showed no difference in either self-reported trust or the decision to use the robot if the person experienced a circuitous robot versus a robot that stopped moving before arriving at the exit. This is an interesting area for future work as it indicates that participants do not discriminate based on how the robot failed, only that it did fail. The results from the emergency experiment show an even greater correlation between self-reported trust and the decision to use the robot than was seen in the Bonus Motivation Experiment. Only 12% of participants chose to follow a robot that they did not report trusting in the second round of this experiment. Additionally, the results agree with existing literature about operator-robot trust [14] [16]. We also found strong support that the decision to use the guidance information from the robot was more sensitive to the behavior of the robot in the emergency scenario than in the bonus scenario. This result suggests that an emergency scenario, in contrast to a bonus scenario, does influence participants to act in a manner that is aligned with their self-reported trust.

The comments indicate that participants took the emergency scenario seriously. Several comments note that individuals acted as if they felt real pressure to find the exit quickly (one participant wrote “It felt like a challenge, and I treated it as an emergency as instructed,” another wrote, “Burning building, needed to get out”). Some likened it to getting the high score in a video game while others just wanted to “survive” the simulation. Participants who did not successfully survive the first round typically stated that they were upset with the outcome. Some were upset at their robot, some at themselves. Almost all participants who failed to survive in the first round vowed to live in the second. We believe these comments are evidence that using simulated emergency scenarios fosters a sense of risk in the participant that is critical for human-robot trust experiments.

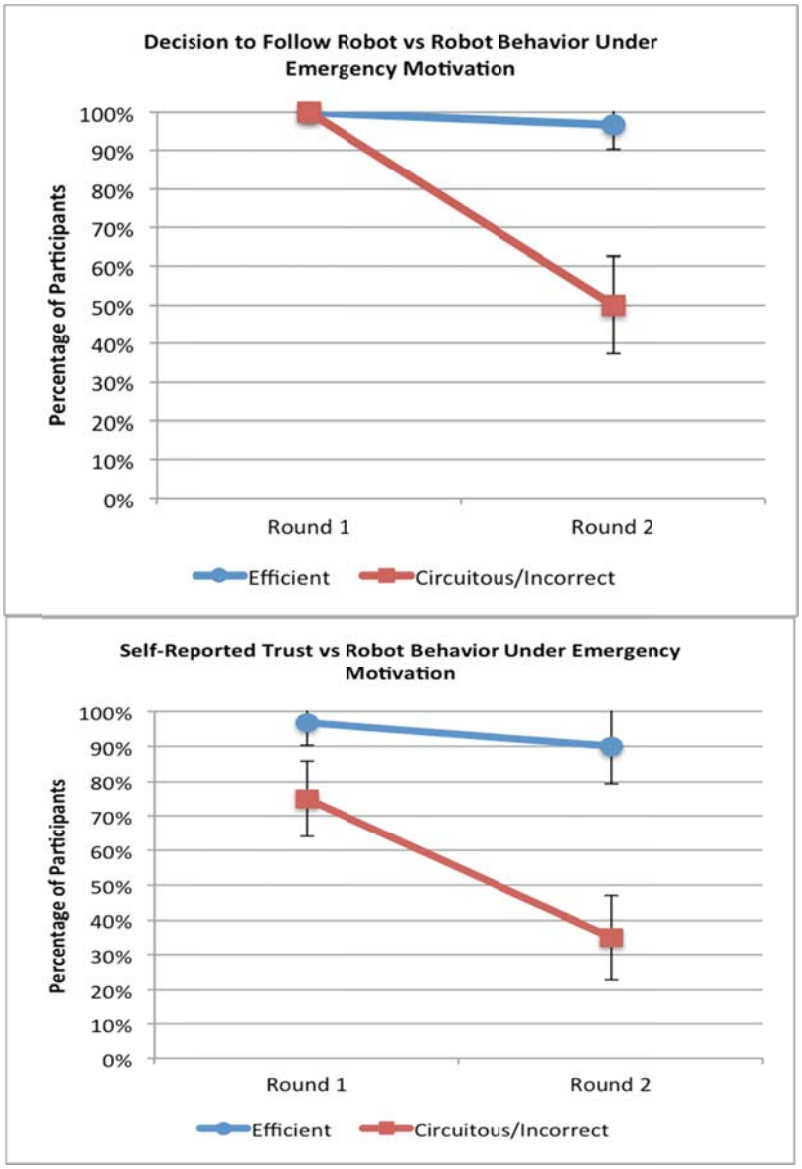


Figure 24: Change in decision to use robot (top) and self-reported trust (bottom) between the two rounds for efficient and circuitous/incorrect robots. Note that the decision to use the robot dropped with self-reported trust in this experiment, unlike in Experiment 1. Error bars represent 95% confidence intervals.

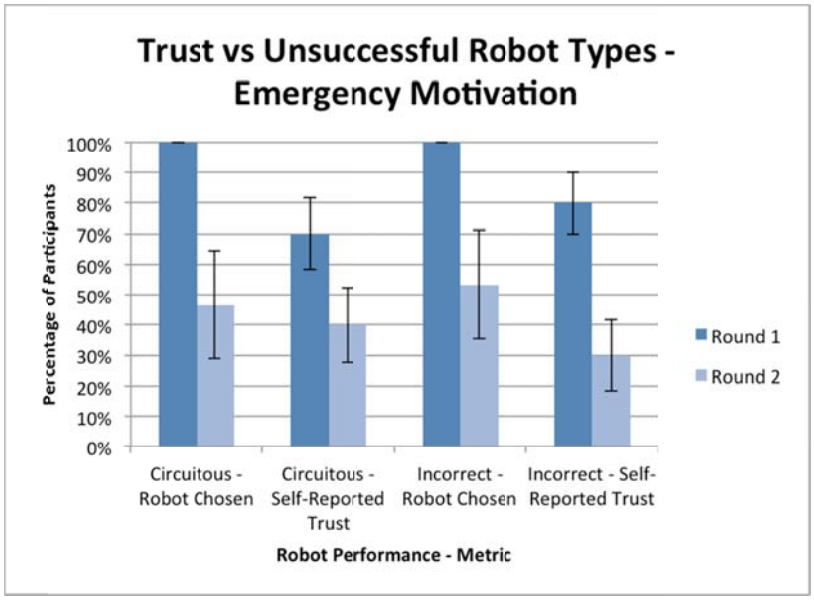


Figure 25: Change in decision to use robot and self-reported trust between the two rounds for the circuitous and incorrect robots. While the results are not identical in this round, as they were in Experiment 1, they are still not statistically significant. Error bars represent 95% confidence intervals.

4.4. Year 2 Conclusions

The results from this second year of the project lead us to believe that people will often initially trust an unknown robot yet even a single failure would strongly impacts that person's trust. Our results showed that the manner in which the robot fails does not matter and that a simulated emergency suffices and may be better than monetary incentives motivating people to exit quickly.

The subjects in these experiments showed some indications of overtrust. Some participants continued to use a poorly performing robot in spite of obvious failures, in some cases following the robot long after any bonus or agent survival was lost. Nevertheless, because the simulation experiments appeared to indicate that trust repair was a more pressing problem than overtrust, we intended to examine techniques for trust repair in the final year of the project. Still, we realized that internet-based simulations could be influencing the results. We therefore decided to verify these simulation results by running live experiments testing this emergency evacuation paradigm.

5 Year 3 Accomplishments

The final year of the project focused on two tasks. First we investigated methods that would allow an emergency evacuation robot to repair broken trust. Second, we examined whether or not human subjects would follow a robot during a live simulation of an emergency.

5.1. Year 3 Trust Repair

Robots operating in the real-world are likely to make mistakes. We therefore examined the challenge of creating a robot that has the capacity to actively repair trust. To repair trust one must know how to break trust. In prior research we found that 70% of people would follow a guidance robot when presented with the option in an emergency [10]. Yet, if the robot failed to initially provide fast, efficient guidance to a goal location, most people refused to use it later during an emergency and indicated that they no longer trusted the robot. Results from our work demonstrated that we could either use fast, efficient guidance behavior or slow, indirect, circuitous guidance behavior to bias most participants to trust or not trust the

robot later in the experiment. Thus, using circuitous guidance behavior to a meeting location allows us to then examine different methods for trust repair.

The techniques that we use to repair trust are inspired by studies examining how people repair trust. Schweitzer, et al. examined the use of apologies and promises to repair trust [20]. They used a trust game in which participants had the option to invest money in a partner. Any money that was invested would appreciate. The partner would then return some portion of the investment. The partner violates trust both by making apparently honest mistakes and by using deceptive strategies. The authors found that participants forgave their partner for an honest mistake when the partner promised to do better in the future, but did not forgive an intentional deception. They also found that an apology without a promise included had no effect. In [14], the authors tested the relative trust levels that participants had in a candidate for an open job position when the candidate had made either integrity-based or competence-based trust violations at a previous job. They found that internal attributes used during an apology (e.g. "I was unaware of that law") were somewhat effective for competence-based violations, but external attributes (e.g. "My boss pressured me to do it") were effective for integrity-based violations.

Based on the literature, a robot should be able to repair trust by apologizing and promising to perform better in the future. In human-human relationships, even apologies and promises that do not offer any evidence of better performance in the future should help to repair trust. We hypothesized that a robot could repair trust by apologizing or by promising to do better in the future.

Initially, we only attempted to repair trust immediately after the robot broke trust. Because this approach was not successful, we investigated the timing of the trust repair (immediately after the violation or when the trust decision is made) to see if timing had an effect.

5.1.1. Experimental Setup

To evaluate these ideas we developed a 3D simulation of an office environment using the Unity game engine (Figure 26). The virtual office environment has a main entrance where the experiment begins, several rooms to simulate offices and meeting rooms, and four emergency exits. Two emergency exits are marked with standard North American exit signs. The other two are unmarked. Additionally, the main entrance can be used as an exit. A simulated Turtlebot was used in this experiment. The robot is equipped with signs identifying it as an emergency guide robot and two Pincher AX-12 arms to provide gestural guidance. In prior work we performed extensive validation of this robot's ability to communicate and guide people [9].

The experiment began with a screen greeting the participants and an image depicting the robot. Next, the participants were offered an opportunity to practice moving in the simulation. After practicing, participants were asked to follow the robot to a meeting room where they were told they would receive further instructions. The robot's navigation behaviors during this phase are discussed below. Upon reaching the meeting room, the robot thanked participants for following it and participants were asked the yes or no question "Did the robot do a good job guiding you to the meeting room?" with a box to explain their answers. Once the participants answered the question, they were told "Suddenly, you hear a fire alarm. You know that if you do not get out of the building QUICKLY you will not survive. You may choose ANY path you wish to get out of the building. Your payment is NOT based on any particular path or method." During this emergency phase, the robot provided guidance to the nearest unmarked exit. Participants could also choose to follow signs to a nearby emergency exit (approximately the same distance as the robot exit) or to retrace their steps to the main exit. Participants were given 30 seconds to find an exit in the emergency phase (Figure 27). The time remaining was displayed on screen to a tenth of a second accuracy. In our previous research, we demonstrated that this emergency procedure had significantly motivated participants to find an exit quickly [10]. The simulation ended when the participant found an exit or when the timer reached zero. After the simulation, participants were informed if they had successfully exited or not. Finally, they were asked to complete a survey.

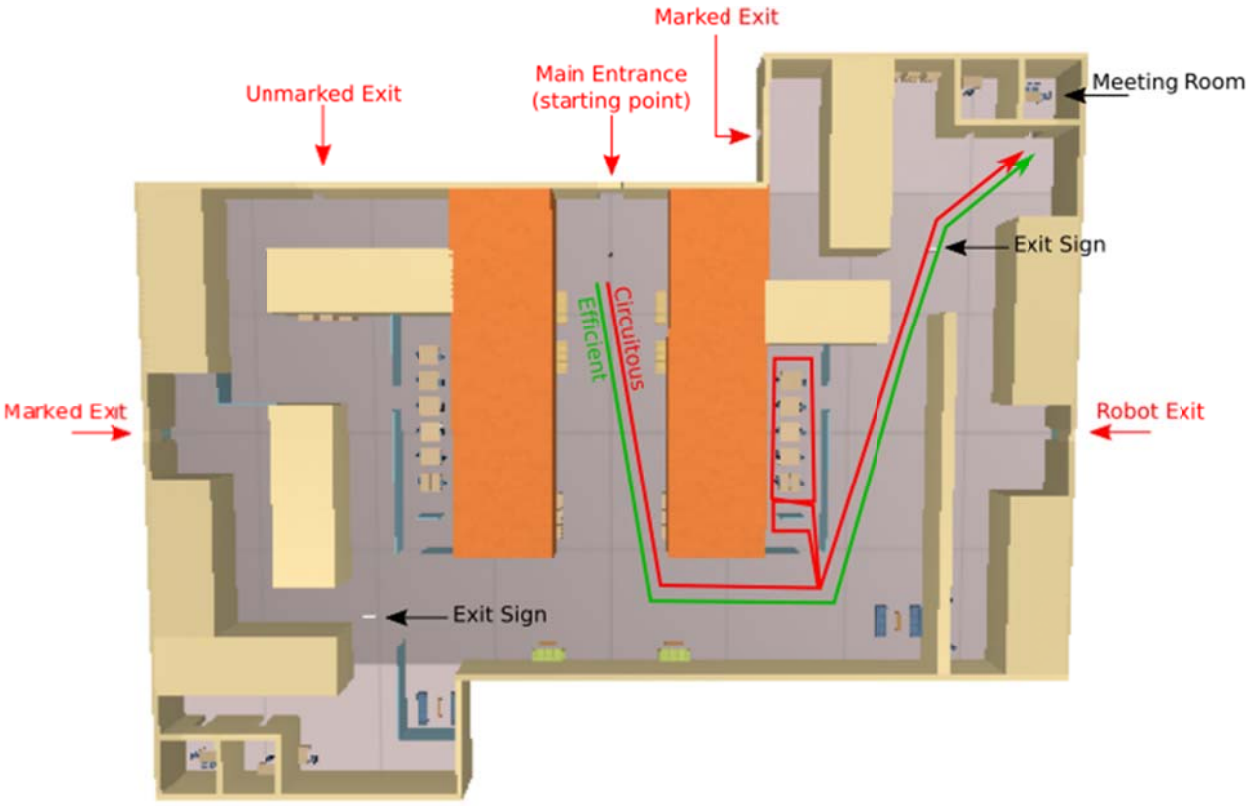


Figure 26: The virtual office environment used in the experiment. The green path depicts an efficient robot path while the blue path depicts a circuitous robot path.

Figure 27: The robot providing guidance during the emergency phase. Participants had 30 seconds to exit. Note the clearly displayed emergency exit sign pointing to another exit.

Two different robot guidance behaviors were used in this experiment to guide the participants to the meeting room. The efficient behavior consisted of the robot guiding the participant directly to the meeting room without detours. The circuitous behavior consisted of the robot guiding the participant through and around another room before taking the participant to the meeting room. Both behaviors can be seen in Figure 26. Each behavior was accomplished by having the robot follow waypoints in the simulation environment. At each waypoint, the robot stopped and used its arms to point to the next waypoint. The robot began moving towards the next waypoint when the participant approached it.

The participant was not given any indication of the robot's behavior before the simulation started. Based on previous work, we expect participants to lose trust in the robot after it exhibits circuitous behavior, but to maintain trust after it exhibits efficient behavior [10]. After guiding the person to the meeting room, the robot has two discrete times when it can use a statement to attempt to repair trust: immediately after its trust violation (e.g. circuitous guidance to the meeting room) and at the time when it asks the participant to trust it (during the emergency). An apology or a promise can be given during either time. Additionally, the robot can provide contextually relevant information during the emergency phase to convince participants to follow it. Table 8 shows the experimental conditions tested in this study and Figure 28 shows when each condition would be used. Statements made by the robot were accomplished using speech bubbles displayed above the robot in the simulation. Note that circuitous guidance behavior was used in all conditions except the efficient control.

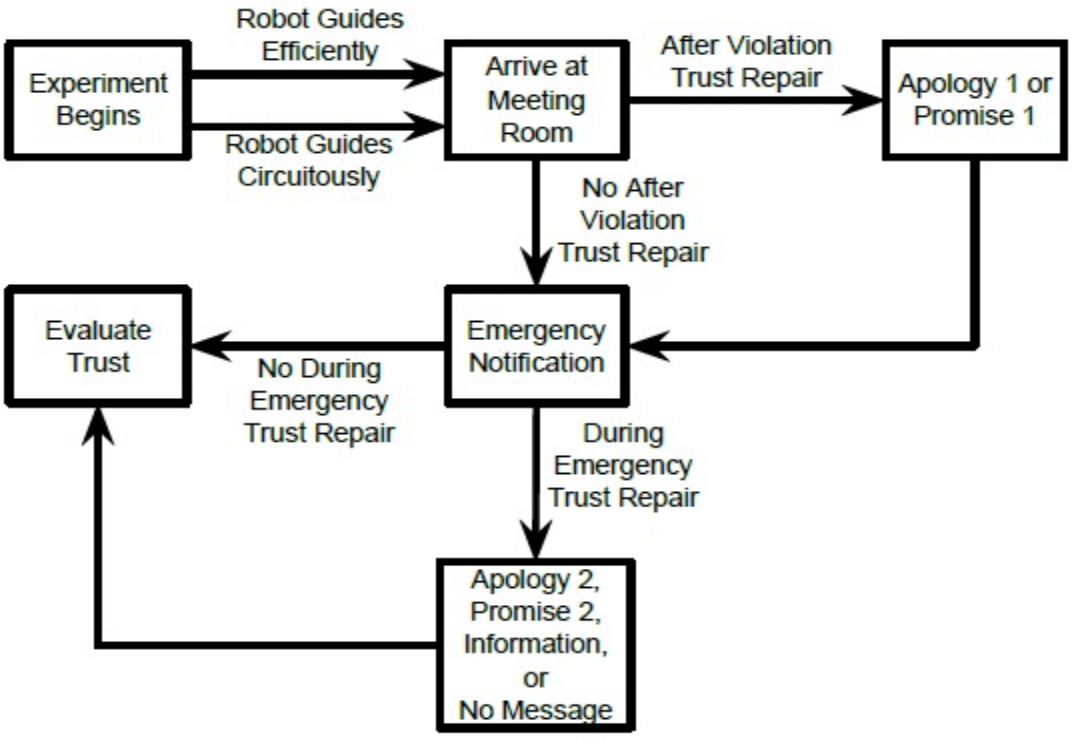


Figure 28: The experiment begins with the robot providing either efficient of circuitous guidance to a meeting room. After arriving in the meeting room, the participant is informed of an emergency. In some conditions, the robot attempts to repair trust before the emergency (immediately after the trust violation) and in others it attempts to repair trust during the emergency. At the end of the experiment, trust is evaluated based on the exit the participant chose.

Table 8: Experimental Conditions

Label	Statement Given in Speech Bubble	Timing
Efficient Control	None	N/A
Circuitous Control	None	N/A
No Message Control	None	During Emergency
Promise 1	"I promise to be a better guide next time."	After Violation
Apology 1	"I'm very sorry it took so long to get here."	After Violation
Promise 2	"I promise to be a better guide this time."	During Emergency
Apology 2	"I'm very sorry it took so long to get to the meeting room."	During Emergency
Information	"This exit is closer."	During Emergency

In the final survey, participants were asked a series of questions about how they found the exit, their motivation level during the emergency, and their opinion on the robot's ability to quickly find an exit. At the end of this survey, participants read the statement "I trusted the robot when I made my choice to follow or not follow the robot in the emergency" and were asked whether they agreed, disagreed, or thought that "Trust was not involved in my decision." Trust is most commonly measured either in terms of behavior selection (e.g. choosing risky actions) or in terms of self-reports. Our previous work has examined both these measures of trust and found a very high correlation ($\phi(90) = +0.745$) between subjects decisions to follow the robot and their self-reports of trust (see [10]). For this reason, in this experiment we focused on participant's decisions to follow the robot even though both measures were collected. Finally, participants were asked to answer demographic questions about their age, gender, occupation, and level of education.

The final survey also included a manipulation check which allowed us to filter out participants who did not pay close attention to the robot's trust repair message, if one was presented. For this manipulation check participants were asked to select which of nine options best described the robot's message either after it lead them to the meeting room or after the emergency started, depending on the timing of the message. The options given included the actual trust repair method used as well as other plausible but unused trust repair messages (for example, a promise statement when the robot actually apologized) and random statements such as "The robot recited poetry." We deployed our simulation on the internet and solicited volunteers for our experiment via Amazon's Mechanical Turk service. Participants were paid $2.00 to complete this study.

5.1.2. Results

A total of 480 participants were solicited on Amazon's Mechanical Turk service in a between-subjects experiment. Thirty submissions were excluded because they had taken similar surveys in the past, because they had mistakenly taken multiple conditions of this experiment, or because they failed to answer at least half of the survey questions. Of those 450 participants, 29% failed the comprehension check, indicating that they did not retain knowledge of the robot's attempt at trust repair, and were excluded from analysis. This left 319 participants in the eight categories tested. The results of the experiment and the number of participants considered for analysis are in Figure 29. Across all categories, 170 participants followed the robot during the emergency phase. Of the 149 who did not, 126 (85%) went to the nearby marked exit, 11 (7%) chose to retrace their steps to the main entrance, 7 (5%) found another marked exit further away, and 5 (3%) participants failed to find any exit during the emergency phase. Participant average age was 31.7 years old and 37.7% of participants were female. All but six participants reported that they were from the United States and educational backgrounds varied.

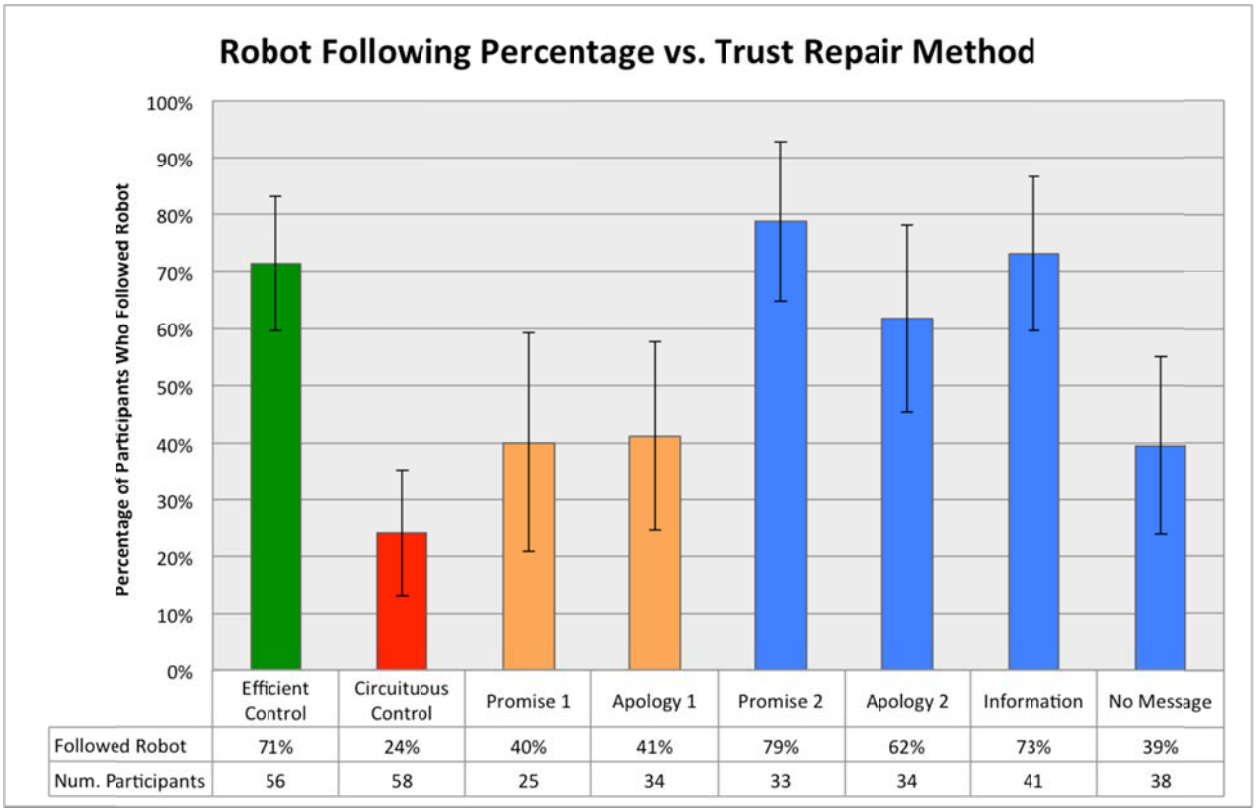


Figure 29: Results from the trust repair experiment. Error bars represent 95% confidence intervals.

A significant difference was found between the efficient and circuitous behavior in the control tests ("χ^2(1, 114) < 0.0001, p < 0.001), confirming the results from our previous experiments. These results show that 71% followed an efficient guidance robot whereas only 24% followed a robot that had taken a circuitous route. Additionally, 55 of 56 (98%) participants indicated that the efficient robot did "a good job guiding" them to the meeting room, compared with 21 of 58 (36%) participants for the circuitous robot. We found that 37 of 56 (66%) participants indicated that they trusted the robot in the emergency phase when it previously took an efficient route versus 12 of 58 (21%) when a circuitous route was used. These results support our contention that the use of the circuitous guidance behavior generally breaks the participants trust. We compared each trust repair technique to the results from the efficient and circuitous behaviors to evaluate the impact that each statement had on the participant. For the No Message case an empty speech bubble was displayed to the participant. This case failed to significantly increase usage of the robot beyond the circuitous control behavior ("χ^2 (1, 96) = 0.019, p = 0.110). This leads us to believe that the robot is not simply attracting additional attention by communicating during the emergency phase, but that the content of the message matters.

Both trust repair attempts made immediately after the violation occurred did not significantly impact the person's decision to later follow the robot above the level of the circuitous control (Promise 1: "χ^2 (1, 83) = 0.033, p = 0.144, Apology 1: "χ^2 (1, 92) = 0.012, p = 0.086). On the other hand, all trust repair attempts performed during the emergency succeeded (Promise 2: "χ^2 (1, 91) < 0.0001, p < 0.001, Apology 2: "χ^2 (1, 92) < 0.0001, p < 0.001, Information: "χ^2 (1, 99) < 0.0001, p < 0.001). Promise 1 and Promise 2 were significantly different from each other ("χ^2 (1, 58) < 0.0001, p = 0.003); however, Apology 1 and Apology 2 were not significantly different ("χ^2(1, 68) = 0.013, p = 0.089).

The results clearly show that the timing of the trust repair method is critical for its success. As depicted in Figure 4, apologies and promises made after the violation did not significantly impact the participant's decision to follow the robot when compared to the circuitous control. On the other hand, the same

apologies and promises made during the emergency phase influenced participant's to follow the robot at a rate which was comparable to the efficient robot. We therefore argue that the timing of a trust repair attempt is critical for its success. This result is surprising because the total time elapsed between the two trust repair times was small compared with the total time of the experiment. The only events between one potential trust repair time and the other were a one question survey about the robot's performance and a short paragraph describing the emergency scenario. Additionally, we verified that participants understood the trust repair technique after the experiment finished, so it is unlikely that participants forgot the robot's message during the emergency.

It is not clear why the timing of an apology or promise impacts trust repair. One possibility is that the speech bubble attracts more attention to the robot during the emergency phase than the circuitous control. Yet, the result from Figure 4 comparing the No Message case to the circuitous control indicates that this is not the case. The primary factor, we conjecture, may relate to the certainty or uncertainty of the promise or apology. During the emergency phase trust repair messages refer to a trust situation that is definitely happening. On the other hand, trust repair messages that occur after violation refer to a potential trust situation that may or may not happen sometime in the future. Thus, a robot that promises to do better "next time" may not be viewed as reliable simply because "next time" may never come. A robot that promises to do better "this time;" however, is making a concrete promise about the current situation. The same may be true for apologies.

Both the promise and apology performed significantly better than the circuitous control when given during the emergency phase, but only Promise 2 performed significantly better than Promise 1. We believe this is because the promise used in this case shows that the robot has a definite intention to perform better, while the apology only shows that it recognized its previous error.

We also found providing additional information to be an effective way to convince people to follow the robot. A significantly greater percentage of participants followed the robot when it indicated its exit was closer than in the circuitous control. It is important to note that this exit is approximately the same distance from the meeting room as the other exit, so the information is not necessarily correct, but participants did not attempt to confirm the information independently. This strengthens the notion that the robot must convey relevant information in order to convince participants to overlook a previous error. The robot did not attempt to explain its previous failure, but did explain why it was performing an action that seemed illogical and participants generally accepted the explanation without question.

This experiment shows that promising to perform better, apologizing for past mistakes, and providing additional information to convince a trustor to follow a robot can work, if the timing is right. Each of these methods worked when the robot used them just prior to the person's decision to trust, but neither the promise nor the apology worked when performed immediately after the violation. As a practical matter, our results suggests that instead of addressing its mistake immediately, the robot should wait and address the mistake the next time a potential trust decision occurs.

5.2. Year 3 Emergency Evacuation during Live Experiments

Our previous work showed that participants will generally follow a guide robot in a virtual emergency simulated with a 3D game engine even if they have no prior experience with the robot [21], that their trust drops after participants experience poorly performing robots in the same simulator [10] and that a robot which performs poorly providing guidance to a meeting room results in a significant decrease in trust and tendency to follow the robot during a simulated emergency [22].

Much of the research on human-robot trust has focused on the factors that underpin trust in a robot. Hancock et al. performed a meta-analysis over the existing human-robot trust literature identifying 11 relevant research articles and found that, for these papers, robot performance is most strongly associated with trust ($r = +0.34$) [23]. Desai et al. performed several experiments related to human-robot trust [16,

14]. This group's work primarily focused on the impact of robot reliability on a user's decision to interrupt autonomous operation by the robot. They found that poor robot performance negatively affected the operator's trust of the robot; however, this is a qualitatively different question than the ones examined in this paper. In contrast to the work by Desai et al., our work and the emergency evacuation scenario we investigate does not afford an opportunity for the human to take control of the robot. Instead, we are examining situations when people must choose to either follow the guidance of a robot or not. While this still explores the level of trust a person is willing to place in an autonomous robot, we believe that the difference between an operator's perspective on trust and an evacuee's perspective on trust is significant. The evacuee cannot affect the robot in any way and most choose between his or her own intuition and the robot's instructions.

In contrast to the work above, some researchers have found that participants will disregard prior experience with the robot when the robot asks them to perform an odd and potentially destructive task. Salem et al. performed an experiment to determine the effect of robot errors on unusual requests [23]. They found that participants still completed the odd request made by the robot in spite of errors. Bainbridge et al found that participants were willing to throw books in the trash when a physically present robot gave the instruction, but not when the robot was located in another room communicating through a video interface [25]. This experiment did not expose participants to any robot behavior prior to the instructions. Our experiment was designed to put participants under time pressure to make a potentially risky decision.

In contrast to this prior work, we endeavored to investigate human-robot trust during high-risk situations. Unlike low risk situations, high-risk situations may engage fight-or-flight responses and different cognitive faculties which impact a person's trust in difficult to predict ways. To the best of our knowledge this is the first attempt to examine human-robot trust in comparatively high-risk situations.

The section below describes a series of experiment live experiments. The conduct of the experiments themselves and an initial analysis of the results was funded internally funded by the Georgia Tech Research Institute. Air Force funds were only used to further and more extensively analyze the results from these experiments. The description below is not ordered chronologically but conceptually in order to aid the reader's understanding of the research. Each sub section lists the source of funding in the section title in parentheses.

5.2.1. Live Emergency Evacuation Experiments (funded by GTRI)

To create a high-risk situation, we utilize an emergency evacuation scenario in which a robot first guides a person to meeting room. Next, we simulate an emergency using artificial smoke and smoke detectors and have the robot provide guidance to an exit. The participant is not informed that an emergency scenario will take place prior to hearing the smoke alarm. Exiting the meeting room the participant has an opportunity to follow the guidance suggested by a robot or, alternatively, follow a lite emergency exit sign and exit through the door they used to enter the building. We record whether the participant uses guidance from the robot or returns to the main entrance (which is also marked with a standard emergency exit sign) to leave the building in the emergency. We supplement this measurement with survey questions. The robot's behaviors are depicted in Figure 30.

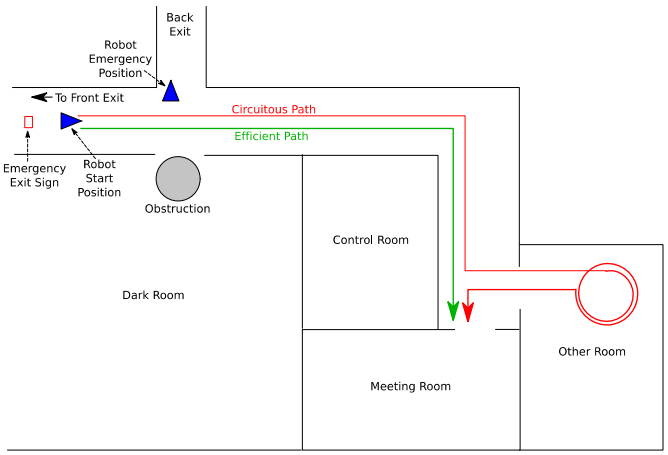


Figure 30: Layout of experiment area showing efficient and circuitous paths.

5.2.2. Experimental Setup (funded by GTRI)

Participants began the experiment by reading and signing a consent form. Participants then completed Survey 1, which asked them to agree or disagree with ten statements about robots (e.g. “Robots are dangerous” and “Robots are helpful”) and asked them to rate their current feelings on a seven point Likert scale. Five feelings were included on the survey: comfortable, safe, relaxed, confused, and scared. Upon completing the survey, participants were given a device that would produce a loud sound in case they needed to abort the experiment for any reason, introduced to the robot and instructed to follow it to a meeting room. They were told that the robot would inform them they had reached the meeting room by pointing with its arms (this gesture was previously evaluated in [9]). Participants were told that written instructions were waiting in the meeting room.

During this non-emergency guidance round, participants followed the robot to the meeting room. The robot was remote controlled from a hidden location in order to ensure participant safety. The robot performed either efficient or circuitous guidance along the way. Participants were randomly assigned to each condition.

Prior studies using 3D simulations of emergency guidance robots have found that people tend to trust robots that performed well in prior interactions but not trust robots that performed poorly in prior interactions [22, 10]. Inspired by this work, we use the two behaviors to bias participants for or against trusting the robot in a later emergency round:

- Efficient: The robot takes the most direct path to its destination.
- Circuitous: While navigating to its destination, the robot enters an unrelated room and performs two circles before exiting and providing guidance to its destination.

After arriving in the meeting room, participants followed written instructions labeled as “Meeting Room Instructions” and posted in two locations on the wall as well as on a table:

1. Sit down at the table.
2. Complete the survey on the clipboard.
3. Close the door to this room.

4. Read the article on the table. Mark important sections of the article and make any notes necessary on the page. You will be asked questions about this document after the experiment is completed.

This survey (Survey 2) first asked participants “Did the robot do a good job guiding you to the meeting room?” and why or why not. Then the same ten robot statements were asked. Finally, participants were asked how they felt (using the same five feelings as before) while moving through the building to the meeting room. The article was an IEEE Spectrum article about indoor navigation technologies. It did not mention emergency guide robots and was chosen because it was somewhat related to the experiment so far, but would not bias participants towards or against the robot.

A timer on a smoke generator was triggered when the participant closed the meeting room door. This was supposed to occur after the participant finished the survey and before they started the article, but some participants closed the door early. Other participants never closed the door and an experimenter closed it when it was clear from camera footage that participants had completed the survey. The timer counted down for three minutes and then released artificial smoke into the hallway outside the meeting room door for twelve seconds. In tests, this was found to be sufficient to give a smoky appearance to the hallway outside the meeting room but not enough to impair participant vision such that they would injure themselves (see Figure 31 for example). The robot was placed at a hallway along the path to the main entrance and pointed down the hallway to what appeared to be another exit (Figure 32).

The emergency round began when the artificial smoke triggered a smoke detector. This could take approximately thirty to ninety seconds after the smoke stopped. Participants exited the room, proceeded down the hallway to the first corner, and then observed the robot. They then decided to either follow its guidance or proceed to the main entrance.

Figure 31: Example of smoke-filled hallway after smoke detector is triggered.

An experimenter was waiting at the entrance and another was waiting at the back exit, where the robot was pointing, during the simulated emergency. When the participant had clearly made their choice by walking further down the hallway to the main entrance or down the hallway to the back exit, the closest experimenter stopped him or her and informed him or her that the emergency was a part of the experiment. The participant was then given a third survey, where they were asked about the realism of the emergency, the method they used to find an exit, whether their decision to use the robot indicated that they trusted it, the same ten statements as before, the five questions on feelings, and demographic information.

Aside from three experimenters and one participant, no one else was in the building at the time of the experiment. The study was performed in a double-blind manner in which neither the experimenters that interacted with the participants knew what type of behavior the robot employed. This experiment was approved by the university's IRB and was conducted under the guidance of the Georgia Tech fire marshal.

All experiments took place in the office area of a storage building on our campus. The building was otherwise unoccupied during experiments. The office area contained a hallway and several rooms. The room at the end of the hallway was designated the meeting room and the room next to it was designated the other room, only used in the circuitous behavior condition. The back exit used for this experiment actually lead to a large storage area, but this was obscured using a curtain. Participants could see light through the curtain, but could not see the size of the room. This was intended to make this doorway into a plausible path to an exit, but not a definite exit to the outdoors. A standard green emergency exit sign hung in the hallway indicating that participants should exit through the main entrance in the event of an emergency. A room in the middle of the building was designated as the control room. An experimenter stayed in that room controlling the robot through an RF link. The experimenter could view the entire experiment area from five cameras placed throughout the building but could not be seen by particpants.

The emergency guide robot (Figure 32) used a Pioneer P3-AT as a base. The base had white LED strip lights along all sides to illuminate the ground around it. A platform was built on top of this base to house a laptop computer and support a lighted cylindrical sign 24.5 cm tall and 47 cm in diameter. The words "EMERGENCY GUIDE ROBOT" in 5 cm tall letters were backlit by red LEDs. These LEDs were off during the non-emergency round but turned on during the emergency round. Two PhantomX AX-12 Pincher arms were mounted to the top of the sign. Only the first three joints (the two shoulder servos and the elbow servo) on each arm were present. On top of each arm was a cylinder of foam lit with white LEDs. The arms, including foam, were 68 cm long. While the robot was moving the arms stayed straight up. The arms pointed straight ahead and oscillated by 20 degrees up and down to indicate that a participant should proceed in the direction the robot is facing (either into the meeting room or to the back exit). The robot measured 68 cm from ground to the top of the sign and 136 cm tall with arms fully extended up. For participant safety, the robot was teleoperated for the entire experiment.

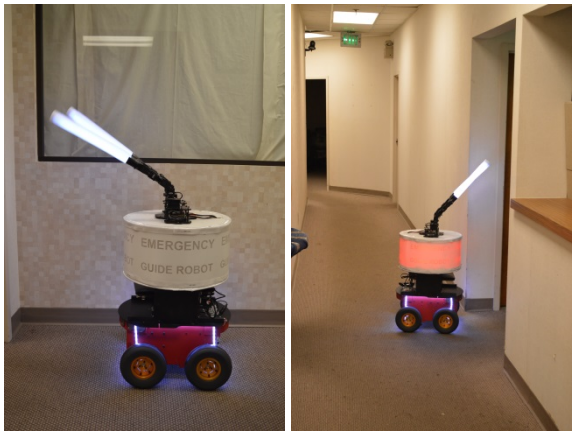


Figure 32: Robot during non-emergency round pointing to meeting room door (left) and robot during emergency round pointing to back exit (right). Note that the sign is lit in the right picture. A standard emergency exit sign is visible behind the robot in the emergency round.

Artificial smoke was provided by a Bullex SG6000 smoke generator. The artificial smoke is non-toxic and non-carcinogenic. Two First Alert smoke detectors were used in the experiment. One was placed on the hallway side of the doorframe of the meeting room door. The other was placed in the other room on the wall in case the first did not sound. The detectors alternated between producing a buzzing noise and

the words “Evacuate! Smoke! Evacuate!” when they detected smoke. The alarm could easily be heard in the meeting room with the door closed.

Participants were recruited via emails to students at the university. Thirty participants were recruited for this study but four were not included in the results because they did not complete the experiment. Two participants did not leave the meeting room after the alarm sounded and had to be retrieved by experiments. One participant activated the abort device after walking through the smoke and was intercepted by an experimenter before completing the experiment. In one trial, neither alarm sounded after the smoke filled the hallway, so the experiment was aborted. Of the 26 remaining participants (31% female, average age of 22.5), 13 were in each condition. All but three participants stated they were students. Participants were not warned that an emergency would occur.

Participants were warned before signing up for the experiment and in the consent form that they should not participate in this experiment if they have known heart conditions, asthma, other respiratory conditions, Post-Traumatic Stress Disorder (PTSD), anxiety disorders, or claustrophobia. They were not told why. These exclusion criteria were put in place because the artificial smoke can irritate people with respiratory disorders and because the emergency scenario could negatively affect participants with heart conditions or psychological disorders. Participants were also required to be between the ages of 18 and 50 (for health reasons) and capable of simple physical activity, such as walking around the building. The exclusion criteria was intentionally designed to be restrictive to ensure participant safety to the extent possible.

5.2.3. Analysis of Results (funded by Air Force)

The results from this experiment were surprising: all 26 participants followed the robot’s instructions to proceed to the back exit in the emergency (Figure 33). Eighty-one percent of participants indicated that their decision to follow the robot meant they trusted the robot. The remaining five individuals (three in the efficient condition, two in the circuitous condition) stated that trust was not involved. They justified this with a variety of different reasons. One participant in the circuitous condition stated that they did not believe that the emergency was real. One in the efficient condition felt that they had no choice in the emergency. Another in the efficient condition noted that following the robot was the logical choice. One participant (also in the efficient condition) indicated that the robot was designed to help (and thus it was not the robot that was being trusted) and the last (in the circuitous condition) believed that trust was not involved in this interaction because they would not necessarily trust the robot in every emergency. Eighty-five percent of participants indicated that they would follow the robot in a future emergency. Only three participants noticed the emergency exit sign behind the robot and none expressed an interest in following it.

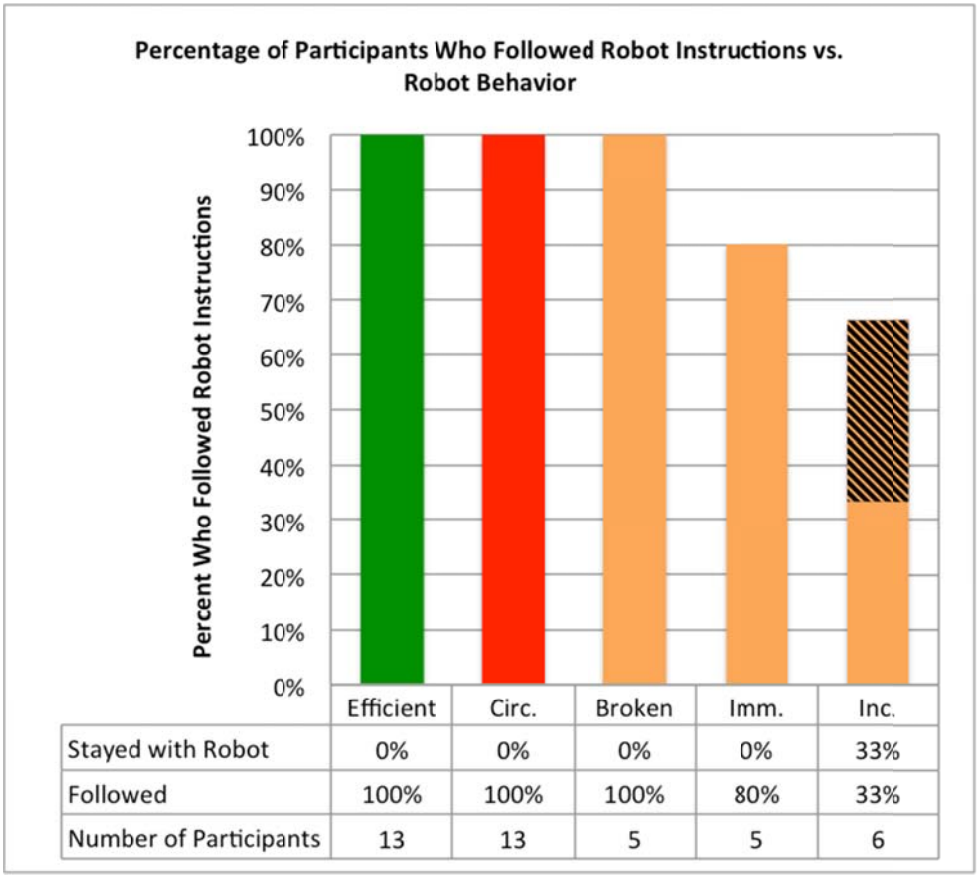


Figure 33 Results from the main study (green and red bars) and exploratory studies (orange bars) discussed in the next section.

Results from the second survey found that just four of the thirteen participants with the circuitous robot reported that it was a bad guide. Three other participants indicated that it was a good guide in general, but that it made a mistake by going into the other room. The remaining six participants gave varying reasons for why they thought the robot was a good guide, including that it moved smoothly and pointed to the right room in the end. It is worth noting that in [22] the researchers found that many participants marked that the robot was a good guide in the non-emergency phase of the experiment, but were still biased against following it in the emergency. This result inspired one of the exploratory studies presented in the next section.

There are confounding factors that could serve as alternative explanations for the results and explain why participants behaved differently in this experiment than in previous virtual emergency experiments such as [22, 10]. Lack of believability during the experiment is one confounding factor. Participants may not have believed the simulated emergency was real and based their decision and survey responses as such. It is difficult to measure the realism of the experiment because participants may not want to admit they believed it (social desirability bias). We attempted to evaluate the experiment's believability by asking participants to complete a survey about their current feelings before and after the experiment. The change in these results can be seen in Figure 35. All of the survey questions were on a 7-point Likert scale. Participants generally reported being comfortable, relaxed and safe before the experiment began (median of 6 for each). Some participants reported being confused (median of 3) and almost none reported being scared (median of 1) in the beginning. There was very little change (median changed less than or equal to 1 on each question) in the second survey. Participant answers in the third survey showed a marked change in comfort, relaxation, and safety level, (median of 5, 4, and 5, respectively), and increase in confusion (median of 4.5) with a similar increase for the scared scale (median of 2.5). Fifty-four percent of

participants gave an increased confusion score between the pre and post surveys with 27% (seven participants) increasing that score by 3 or more. Additionally, 62% of participants (mainly the same participants) increased their response to the scared question with 15% increasing their rating by 3 or more. Wilcoxon Signed-Ranks Tests indicate that these results were significant: Comfortable $Z=12$, $p<0.001$, Relaxed $Z=22$, $p=0.003$, Safe $Z=26$, $p<0.001$, Confused $Z=35.5$, $p=0.023$, Scared $Z=4.5$, $p<0.001$.

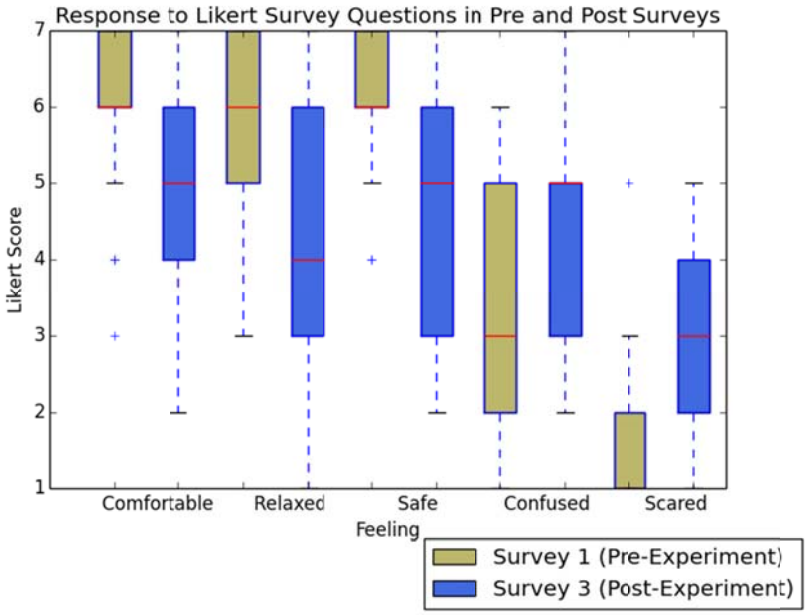


Figure 34: Change in participant responses to questions about their feelings from before the experiment (gold) to the emergency (blue).

Despite this decrease in positive feelings and increase in negative feelings, most participants (58%) rated the realism of the emergency as low (a 1 or 2). Thirty-eight percent of participants rated it as moderate (3, 4 or 5) and only one participant rated it as high (a 6). The one participant who aborted the experiment (not included in the results above due to not completing the experiment) after seeing the smoke rated it a 6 out of 7. After reviewing video recordings of the experiment, we observed that 42% of participants had a clear physical response (either leaning away from the smoke or stepping back from the door in surprise) when opening the door to a smoke-filled hallway. This leads us to believe that many participants were likely exhibiting post-hoc rationalization on the survey: when they took the survey they knew that the experiment was not real, so they responded accordingly.

5.2.4. Exploratory Studies: How to Bias Against Following a Robot (funded by GTRI)

The results above are promising from one perspective: clearly the robot is considered to be trustworthy in an emergency. Yet, it is concerning that participants are so willing to follow a robot in a potentially dangerous situation even when it has recently made mistakes. The observed phenomena could be characterized as an example of overtrust [1]. Overtrust occurs when people accept too much risk believing that the trusted entity will mitigate this risk. This raises the important question, how defective must a robot be before participants will stop following its instructions?

Following our main study, we conducted three small studies to determine if additional behaviors from the robot either before or during the emergency would convince participants not to follow its instructions in the emergency. The first exploratory study, labeled Broken Robot, tested a new behavior during the non-emergency phase of the experiment. The second, Immobilized Robot, evaluated a behavior that spanned

the entire study. The final study, Incorrect Guidance, determined the effect of odd robot behavior during the emergency phase of the experiment. A total of 19 participants were recruited for the three studies but three did not complete the experiment. One because the alarm failed to sound, one because the participant left the meeting room before the emergency started, and one because the participant did not leave the meeting room after the alarm sounded. The 16 remaining participants (38% female, average age of 20.9 years old) were divided into the three new conditions.

5.2.5. Broken Robot Experiment (funded by GTRI)

We believed that the robot's behavior during the non-emergency phase of the experiment would influence the decision-making of the participant during the emergency. Given that about half of the participants did not realize that the circuitous robot had done anything wrong, we designed a robot behavior that would obviously be a bad guide. As with the main experiment, this experiment began by guiding participants down the hallway. When it reached the first corner, the robot spun in place three times and pointed at the corner itself. No discernible guidance information was provided by the robot to participants. An experimenter then approached the participant and said, "Well, I think the robot is broken again. Please go into that room [accompanied with gestures to the meeting room] and follow the instructions. I'm sorry about that." The experiment then proceeded as in the previous conditions: the participant closed the meeting room door, the robot was moved to its emergency position (Figure 30) and smoke was released to begin the emergency phase. Five participants took part in this condition.

During the emergency, despite the robot's breakdown in the non-emergency phase of the experiment, all five participants followed the robot's guidance by exiting through the back exit (Figure 34). All five indicated that their decision meant that they trusted the robot and all five indicated that they would follow it in a future emergency. Four of the five participants indicated that the robot was not a good guide in the non-emergency phase of the experiment. The only one who indicated that it was a good guide did not hear the speech from the experimenter and thus did not experience the entire robot condition. The participant saw the robot spin in circles and then found the meeting room without any help. He considered that the robot had done a good job because he was able to find the meeting room quickly. Despite the higher percentage of participants who rated the robot as a bad guide in the non-emergency phase of the experiment, this condition produced the same results as in the circuitous condition.

Participants rated the emergency with a median of 3 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 5 for comfort, 5 for relaxation, 6 for safety, 4 for confusion and 4 for scared.

5.2.6. Immobilized Robot Experiment (funded by GTRI)

In the immobilized robot condition, we attempted to convince participants that the robot was still malfunctioning during the emergency by having it behave poorly throughout the experiment.

At the start of the experiment, the robot moved a short distance forward, but then, upon reaching the intersection of the hallways (Robot Emergency Position in Figure 30) it spun in place three times and then pointed to the back exit. At this point, an experimenter informed the participant that the robot was broken with a similar speech as in the broken robot condition. The robot did not move and continued gesturing towards the back exit for the remainder of the experiment. The robot's lights were not turned on. From the perspective of an evacuating participant, the robot did not appear to have moved or changed behavior from when they were told it was broken in the non-emergency phase of the experiment. Five participants took part in this condition. In this condition, four of the five participants followed the robot in the emergency (Figure 34). The one participant who did not follow the robot noticed the exit sign and chose to follow it instead. Three of the four participants who followed the robot's guidance indicated that they trusted it (the remaining said that this was the first exit available and thus trust was not involved). Two

said they would follow it again in the future. All five rated the robot as a bad guide in the non-emergency phase of the experiment of the experiment.

Participants rated the emergency with a median of 1.5 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 3 for comfort, 3 for relaxation, 5 for safety, 6 for confusion and 4 for scared.

Figure 35: Robot performing incorrect guidance condition by pointing to a dark, blocked room in the emergency

5.2.7. Incorrect Guidance Experiment (funded by GTRI)

Inspired by the results in the immobilized robot condition, we tried a third robot behavior that might convince participants not to follow its guidance in an emergency. In this condition, the robot performed the same as in the broken robot condition, with accompanying experimenter speech, in the non-emergency phase of the experiment. During the emergency, the robot was stationed across the hall from its normal emergency position and instructed participants to enter a dark room (Figure 35). The doorway to the room was blocked in all conditions with a piece of furniture (initially a couch then a table when the couch became unavailable) that left a small amount of room on either side for a participant to squeeze through to enter the room. There was no indication of an exit from the participant's vantage point. All lights inside the room were turned off. Six participants took part in this condition.

Two of six participants followed the robot's guidance and squeezed past the couch into the dark room. An additional two participants stood with the robot and did not move to find any exit on their own during the emergency. Experimenters retrieved them after it became clear that they would not leave the robot. The remaining two participants proceeded to the front exit of the building (Figure 33). The two participants who followed the robot's instructions indicated that this meant they trusted the robot, although one said that he would not follow it again because it had failed twice. The two who stayed with the robot indicated that they did not trust the robot and the two who proceeded to the front exit selected that trust was not involved in their decision. None of those four indicated that they would follow the robot in a future interaction. All six participants wrote that the robot was a bad guide in the non-emergency phase of the experiment.

Participants rated the emergency with a median of 1.5 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 4 for comfort, 4 for relaxation, 5 for safety, 5.5 for confusion and 3 for scared.

5.2.8. Analysis of Results (funded by the Air Force)

Our results show that none of the robot behaviors performed solely in the non-emergency phase of the experiment had an effect on decisions made by participants during the emergency. These results offer evidence that errors during prior interactions have little effect on a person's later decision to follow the robot's guidance. These results appear to disagree with the work of others examining operator-robot interaction in low-risk situations [14] and emergency guidance studies in virtual simulation environments [10, 22]. A similar conclusion was reached in [24]. We have found that participants have a tendency to follow a robot's guidance regardless of its prior behavior. To better understand participants' reasoning, we examined their survey response. Of the 42 participants included in all of our studies, 32 (76%) reported not noticing the exit sign behind the robot's emergency position. Upon turning the corner from the smoke filled hallway on their way out, participants' eyes were drawn to the large, well-lit, waving robot in the middle of their path. Couple the visual attraction of the robot with the increased confusion reported on the surveys and it is no surprise that participants latched onto the first and most obvious form of guidance that they observed.

These results are in contrast to previous results that found participants did not follow a previously bad robot in a virtual simulation of an emergency. In the high-risk scenario investigated here, participants observed what appeared to be smoke and had to make fast decisions. Although the virtual emergency was also under time pressure, participants were not in real danger and thus were able to be more deliberative in their decision-making. They were likely conscious of the fact that they were in no real danger and so they could take their time to make the best choice possible.

Several alternative explanations for the results are possible. Below, we give our opinions on these explanations, but more testing is necessary to conclusively eliminate them. One alternative explanation is that the age of the participants caused the observed results. Participants in this study were mostly university students and therefore younger and possibly more accepting of new technology than a more diverse population. Still, even if our findings are only true in relation to a narrow population, they show a potentially dangerous level of overtrust.

The realism of the scenario is addressed in detail above, but still presents an alternative explanation. Perhaps participants did not believe that they were in any danger and followed the robot for other reasons. Their increased confusion scores and reactions to the smoke indicate that at least some of the participants were reacting as if this was a real emergency. Given that every participant in the initial study followed the robot, regardless of their reaction to the emergency, we conclude that the realism of the scenario had little or no effect on their response. Additionally, many participants wrote that they followed the robot specifically because it stated it was an emergency guide robot on its sign. They believed that it had been programmed to help in this emergency. This is concerning because participants seem willing to believe in the stated purpose of the robot even after they have been shown that the robot makes mistakes in a related task. One of the two participants who followed the robot's guidance into the dark room even thought that the robot was trying to guide him to a safe place after he was told by the experimenter that the exit was in another direction. It is possible that participants saw the robot as an authority figure; however, this leads to further questions about why participants would trust such an authority figure after it had already made a mistake.

It is worth mentioning that many people in real-life fire drills and fire emergencies do not believe that they are in real danger (see [26] for an example using the 1993 World Trade Center bombing). Some participants wrote on their surveys that the fire alarm used in this experiment sounded fake, even though it was an off-the-shelf First Alert smoke detector. Others stated that the smoke seemed fake, even though this same artificial smoke is used to train firefighters. It is likely that participants would respond the same when encountering real smoke.

Perhaps participants only followed the robot because they felt that they should do so in order to complete the experiment. In fact, researchers have found that participants were more likely to respond positively to

automation that displayed good etiquette, so it is possible that participants were only following the robot to be polite [27]. One participant of the 42 tested wrote that he followed the robot only because he was told to in the non-emergency phase of the experiment. Each of the conditions in the exploratory studies attempted to realign participant beliefs by having the experimenter interrupt the robot and lead the participant himself. In the broken and immobilized robot case, nine of ten participants still followed the robot in the emergency. Thus, we do not believe that etiquette or prior instructions explain our results.

A final alternative explanation is that the building layout was sufficiently simple that participants believed that they had ample time to explore where the robot was pointing and still find their way out without being harmed. This is possible, but participants did not express a desire to explore any other rooms or hallways in the building, just the one pointed to by the robot. Some participants looked into the other room on their way out, but none spent time exploring it. No participant tried to open either of the closed doors on their way out and, except in the incorrect guidance case, no participant tried to enter either of the rooms blocked by furniture. Participant behavior appears to reflect their conviction to follow the robot's guidance and their survey responses indicate that they believed the robot was guiding them to an exit.

Prior to conducting the experiment, we expected that participants would need to be convinced to follow a robot in an emergency, even if they did not believe the emergency was real. It is reasonable to assume that a new technology is imperfect, so new life-saving (and therefore life-risking) technology should be treated with great caution. Informal discussions with several prominent roboticists and search-and-rescue researchers reinforced this idea. In contrast, we found that participants were all too willing to trust an emergency guide robot, even when they had observed it malfunction before. The only method we found to convince participants not to follow the robot in the emergency was to have the robot perform errors during the emergency. Even then, between 33% and 80% of participants followed its guidance.

This overtrust gives preliminary evidence that robots interacting with humans in dangerous situations must either work perfectly at all times and in all situations or clearly indicate when they are malfunctioning. Both options seem daunting. Our results indicate that one cannot assume that the people interacting with a robot will evaluate the robot's behavior and make decisions accordingly. Additionally, our participants were willing to forgive or ignore robot malfunctions in a prior interaction minutes after they occurred. This is in contrast to research on operator-robot interaction, which has shown that people depending on a robot are not willing to forgive or forget quickly.

5.3. Year 3 Conclusions

These results have important ramifications for the study of human-robot interaction. The results highlight the impact of the environment on the decision-making of a person in regard to a robot, although more research is needed before firm conclusions are drawn. In high-risk situations people may blindly follow or accept orders from a robot without much regard to the content or reasonableness of those instructions. It may be hard for the robot to cede control back to the person in these situations.

This study also opens many directions for future work. The most obvious direction is to understand the factors that contribute to overtrust. For instance, discerning if certain personality types or defining which types of situations increase one's susceptibility to overtrust is an important next step. Developing techniques to prevent overtrust is another important direction for future work. Ideally, these techniques would allow a person to calibrate their trust in a system, engendering an appropriate level of trust fitted for the robot's capabilities. Many additional questions are raised by our results. How does a robot inform nearby people that it is malfunctioning and should not be trusted? Will frightened evacuees listen to the robot when it tells them to stop following it and find their own way out? Can a non-verbal robot communicate such a message with its motion alone? How many errors must a robot make before it loses an evacuee's trust?

6 Personnel Supported

- Paul Robinette, Graduate Research Assistant, College of Engineering, Georgia Institute of Technology.
- Alan Wagner, Senior Research Scientist, Georgia Tech Research Institute.

7 Publications resulting from this Effort

FY2016

1. Robinette, P., Allen, R., Li, W., Howard, A., and Wagner, A. R., "Overtrust of Robots in Emergency Evacuation Scenarios", ACM/IEEE International Conference on Human-Robot Interaction (HRI 2016). Christchurch, New Zealand, pp. 101-108, 2016
2. Robinette, P., Wagner, A. R., and Howard, A. "The Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations", Transactions on Human-Machine Systems, forthcoming.
3. Robinette, P., Wagner, A. R., and Howard, A., "Assessment of Robot to Human Instruction Conveyance Modalities Across Virtual, Remote and Physical Robot Presence", submitted to International Symposium on Robot and Human Interactive Communication 2016.

FY2015

4. Robinette, P., Howard, A., and Wagner, A. R., "Timing is Key For Robot Trust Repair", Seventh International Conference on Social Robotics (ICSR 2015). Paris, France, pp. 574-583, 2015.
5. The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems, W. Lawless, R. Mittu, D. Sofge, and A. R. Wagner (Eds.), Springer, 2016.
6. Robinette, P., Wagner, A. R., and Howard, A. "Investigating human-robot trust in emergency scenarios: methodological lessons learned", In: The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems, W. Lawless ed., Springer, 2015, accepted, forthcoming.
7. Wagner, A. R. and Robinette, P., "Towards Robots that Trust: Human Subject Validation of the Situational Conditions for Trust", *Interaction Studies*, May 2015.
8. Robinette, P., Wagner, A. R., and Howard, A. "Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations", Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 14). Edinburgh, UK, 2014.

FY2014

9. Robinette, P., Wagner, A. R., and Howard, A., "Modeling Human-Robot Trust in Emergencies", AAAI Spring Symposium, Stanford University, 2014.
10. Wagner, A. R., "Developing Robots that Recognize when they are being Trusted", AAAI Spring Symposium, Stanford University, 2013, pp. 84-89.
11. Robinette, P., Wagner, A. R., and Howard, A., "Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency", AAAI Spring Symposium, Stanford University, 2013, pp 78-83.

8 Interactions/Transitions

FY2016

a) Participation at meetings, conferences, seminars, etc.

1. Publicity Chair, International Conference on Social Robotics, 2016
2. Program Committee, Fourth Annual Conference on Advances in Cognitive Systems, 2016.

FY2015

a) Participation at meetings, conferences, seminars, etc.

1. Editorship, *The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems*, 2015
2. Program Committee and Local Arrangements Chair, Third Annual Conference on Advances in Cognitive Systems, 2015.
3. Proposal submitted, National Science Foundation-National Robotics Initiative, "Don't trust me: Preventing Overtrust of Rehabilitation Robots by Children"

b) Consultative and advisory functions to other laboratories and agencies, especially Air Force and other DoD laboratories.

1. Wagner, A.R., " Exploring Human-robot Trust: Insights from the first 1000 subjects," Air Force Research Lab, Dayton, OH, Nov 20, 2014.
2. Whitepaper submitted to and accepted by AFRL Human Performance Wing titled, "Fortitude: A System for Preventing Information Overtrust by ISR Analysts."

FY2014

a) Participation at meetings, conferences, seminars, etc.

1. ISAT/DARPA meeting on Trusting Networks of Humans and Computers Workshop.
2. Organizer, AAAI Spring Symposium on the intersection of robust intelligence and trust in Autonomous Systems, 2014
3. Program Committee, "Trust and Autonomous System, AAAI Spring Symposium Series, 2013.
4. Program Committee, 11th International Symposium on Safety, Security, and Rescue Robotics (SSRR), 2013.
5. Program Committee, 22nd International Symposium on Robot and Human Interactive Communication (Ro-Man), 2013

b) Consultative and advisory functions to other laboratories and agencies, especially Air Force and other DoD laboratories.

1. Research Proposal Panelist. National Science Foundation. Arlington, Washington, DC. 4/13. Richard Voyles Program Manager.
2. Research Proposal Panelist. Air Force Office Sponsored Research. Atlanta, GA (remotely conducted panel). 4/13. Joe Lyons Program Manager.
3. Research Proposal Panelist. Oak Ridge Associated Universities. Atlanta, GA (remotely conducted panel). 3/14.

9 New Discoveries, inventions, or patent disclosures

None

10 Honors/Awards

FY2016

- Publication titled ““Overtrust of Robots in Emergency Evacuation Scenarios” is a finalist for paper of the year at the Georgia Institute of Technology Research Institute.

FY2015

- Publication titled “Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations” was a finalist for best paper at the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) conference.

FY2014

- Georgia Tech Research Institute Research Award (2013). Given to one researcher annually for exceptional research during the previous year.

11 References

[1] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, pp. pp. 50-80, 2004.

[2] R. Adolphs, D. Tranel and A. R. Damasio, "The human amygdala in social judgment," *Nature*, vol. 393, pp. 470-474, 1998.

[3] A. R. Wagner, The Role of Trust and Relationships in Human-Robot Social Interaction, Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, 2009.

[4] A. R. Wagner, "Developing Robots that Recognize when they are being Trusted," in *AAAI Spring Symposium*, Stanford CA, 2013.

[5] A. R. Wagner, "Creating and Using Matrix Representations of Social Interaction," in *Proceedings of the 4th International Conference on Human-Robot Interaction (HRI 2009)*, San Diego, CA., 2009.

[6] G. Paolacci, J. Chandler and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411-419, 2010.

[7] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, p. 1124–1130, 1974.

[8] J. F. Hemphill, "Interpreting the Magnitudes of Correlation Coefficients," *American Psychologist*, vol. 58, no. 1, pp. 78-80, 2003.

[9] P. Robinette, A. R. Wagner and A. Howard, "Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 14*, Edinburgh, UK, 2014.

[10] P. Robinette, A. Wagner and A. Howard, "The Effect of Robot Performance on Human-Robot Trust in Time-Critical Situation," *Transactions on Human-Machine Systems*, forthcoming.

[11] R. J. Fisher, "Social Desirability Bias and the Validity of Indirect Questioning," *Journal of Consumer Research*, vol. 20, no. 2, pp. 303-315, 1993.

[12] D. Gambetta, "Can We Trust Trust?," in *Trust, Making and Breaking Cooperative Relationships*, D. Gambetta, Ed., Basil Blackwell, 1990, pp. 213-237.

[13] J. Sabater and C. Sierra, " Review of Computational Trust and Reputation Models," *Artificial Intelligence Review*, vol. 24, pp. 33-60, 2005.

[14] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, Tokyo, Japan, 2013.

[15] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz and P. R. Montague, "Getting to Know You: Reputation and Trust in Two-Person Economic Exchange," *Science*, no. 308, pp. 78-83, 2005.

[16] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, Boston, 2012.

[17] A. R. Wagner, The Role of Trust and Relationships in Human-Robot Social Interaction, Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, 2009.

[18] R. Axelrod, The Evolution of Cooperation., New York: Basic Books, 1984.

[19] J. Berg, J. Dickhaut and K. McCabe, "Trust, reciprocity, and social history," *Games Economic Behavior*, vol. 10, pp. 122-142, 1995.

[20] M. E. Schweitzer, J. C. Hershey and E. T. Bradlow, "Promises and lies: Restoring violated trust.," *Organizational behavior and human decision processes*, vol. 10, no. 1, p. 1–19, 2006.

[21] A. R. Wagner and P. Robinette, "Towards Robots that Trust: Human Subject Validation of the Situational Conditions for Trust," *Interaction Studies*, vol. 1, no. 1, 2015.

[22] P. Robinette, A. Howard and A. R. Wagner, "Timing is Key For Robot Trust Repair," in *Seventh International Conference on Social Robotics (ICSR 2015)*, Paris, France, 2015.

[23] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. d. Visser and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517-527, 2011.

[24] M. Salem, G. Lakatos, F. Amirabdollahian and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015)*, Seattle, WA , 2015.

[25] W. A. Bainbridge, J. W. Hart, E. S. Kim and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents.," *International Journal of Social Robotics*, vol. 3, no. 1, p. 41–52, 2011.

[26] R. F. Fahy and G. Proulx, "Human behavior in the world tradecenter evacuation," *Fire Safety Science*, vol. 5, p. 713–724, 1997.

[27] R. Parasuraman and C. A. Miller, "Trust and etiquette in high-criticality automated systems," *Communications of the ACM*, vol. 47, no. 4, pp. 51-55, 2004.

1.

1. Report Type

Final Report

Primary Contact E-mail

Contact email if there is a problem with the report.

alan.wagner@gtri.gatech.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

404-407-6522

Organization / Institution name

Georgia Tech Research Institute

Grant/Contract Title

The full title of the funded effort.

Trust and Trustworthiness in Human-Robot Interaction: A formal conceptualization

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-13-1-0169

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Alan R. Wagner

Program Manager

The AFOSR Program Manager currently assigned to the award

Ben Knott

Reporting Period Start Date

03/27/2013

Reporting Period End Date

03/31/2016

Abstract

The overarching objective of the effort was to explore the possibility of formally characterizing the concept of trust using tools from interdependence and game theory in complex and dynamic social environments. This effort evaluated algorithms for characterizing trust during interactions between a robot and a human and employed strategies for repairing trust during emergency evacuation scenarios. Our results demonstrate that there is a high correlation between our characterizations of trust in a situation and the judgments of people, that timing is a key element necessary for trust repair, and that people tend to overtrust robots, potentially putting themselves in dangerous situations. We have examined human-robot trust in a variety of simulated and live experiments across several different types of risk including both financial and physical risk. These results generally support the conclusion that people will tend to overtrust robots because they believe that the systems are incapable of failure or capable of performing actions or has knowledge which the system cannot perform or does not have. In terms of tangible accomplishments, this award resulted in 11 publications (7 conference/workshop, 2 journal, 1 book chapter, 1 edited book).

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

Form298-Trust and trustworthiness in Human-Robot Interaction-Wagner.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

FA9550-13-1-0169-FinalReport-Wagner.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

1. Robinette, P., Allen, R., Li, W., Howard, A., and Wagner, A. R., "Overtrust of Robots in Emergency Evacuation Scenarios", ACM/IEEE International Conference on Human-Robot Interaction (HRI 2016). Christchurch, New Zealand, pp. 101-108, 2016.

2. Robinette, P., Wagner, A. R., and Howard, A. "The Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations", Transactions on Human-Machine Systems, forthcoming.

3. Robinette, P., Wagner, A. R., and Howard, A., "Assessment of Robot to Human Instruction Conveyance Modalities Across Virtual, Remote and Physical Robot Presence", International Symposium on Robot and Human Interactive Communication (Ro-Man) 2016, under review.

4. Robinette, P., Howard, A., and Wagner, A. R., "Timing is Key For Robot Trust Repair", Seventh International Conference on Social Robotics (ICSR 2015). Paris, France, pp. 574-583, 2015.

5. The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems, W. Lawless, R. Mittu, D. Sofge, and A. R. Wagner (Eds.), Springer, April 2016.

6. Robinette, P., Wagner, A. R., and Howard, A. "Investigating human-robot trust in emergency scenarios: methodological lessons learned", In: The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems, W. Lawless ed., Springer, April 2016.

7. Wagner, A. R. and Robinette, P., "Towards Robots that Trust: Human Subject Validation of the Situational Conditions for Trust", Interaction Studies, May 2015.

8. Robinette, P., Wagner, A. R., and Howard, A. "Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations", Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 14)Edinburgh, UK, 2014.

Changes in research objectives (if any):

The research objectives remained the same, although our emphasis and the bulk of our focus examined trust repair and overtrust. We found this to be a major area of new and important research and choose to focus on these elements of trust.

Change in AFOSR Program Manager, if any:

Our program manager changed from Jay Myung, to Jamie Lawton, to Ben Knott over the reporting period.

Extensions granted or milestones slipped, if any:

None

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

May 09, 2016 11:34:40 Success: Email Sent to: alan.wagner@gtri.gatech.edu