Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems

Munjal Desai[1], **Kristen Stubbs**[1], **Aaron Steinfeld**[2] and **Holly Yanco**[1]

Abstract. One of the most significant challenges of human-robot interaction research is designing systems which foster an appropriate level of trust in their users: in order to use a robot effectively and safely, a user must place neither too little nor too much trust in the system. In order to better understand the factors which influence trust in a robot, we present a survey of prior work on trust in automated systems. We also discuss issues specific to robotics which pose challenges not addressed in the automation literature, particularly related to reliability, capability, and adjustable autonomy. We conclude with the results of a preliminary web-based questionnaire which illustrate some of the biases which autonomous robots may need to overcome in order to promote trust in users.

1 Introduction

Effective human-robot interaction depends not just on the design of the interaction but also on the level of trust that the user has in the robotic system. In their survey of trust and automation research, Lee and See define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [20]. The level of trust that people have in an automated system is a key factor that influences their use of that system. An inappropriate level of trust may result in inappropriate use (misuse) or disuse of automation, which may result in poor performance [20]. While considerable work has been done on trust in automation, we feel that the added uncertainty and vulnerability inherently present in robots necessitates dedicated work on trust and robotics.

People who place little trust in automation may disuse it. This lack of usage has potential safety implications, as the automation may be better equipped to handle crises than they are. For example, a crash avoidance system in a car may be disabled or tampered with if the driver considers past behaviors untrustworthy, yet the crash avoidance system might be able to react more quickly than the driver in some situations (e.g., [41]). As we have observed during the course of our previous studies in human-robot interaction (HRI), robot users who do not trust a robot often disengage its autonomous capabilities, such as obstacle avoidance. This avoidance of automation, particularly when reverting to direct control (teleoperation), may lead to damage of both the robot and its environment. For example, Yanco et al. observed and evaluated robotic systems designed for urban search and rescue (USAR) at the Robot Rescue Competition at AAAI 2002 [45]. During one of the system's runs, there was a clear Plexiglas sheet present in the path of the robot. The sensors on the robot detected the Plexiglas, and because the robot was utilizing automation, the robot would not drive forward even when the user tried to force it to drive forward. This frustrated the user, who switched to manual control and drove through the Plexiglas sheet.

Problems can also arise when people place too much trust in an automated system. For example, robot users may incorrectly internalize a certain level of trust (e.g., that the robot will not collide with nearby humans) and inadvertently place themselves or others in harm's way. For example, during the experiments carried out by Desai [6], there was an expert user who had been trained with the system and was aware of the system's capabilities and shortcomings. This user decided to use the maximum level of autonomy supported by the system. The robot performed poorly, but the user kept the system under full automation for more than half the run. One of the statements that the user made before starting the run indicated that the user was very enthusiastic about autonomy and wanted to try it out. This statement indicated a very high level of initial trust in the system. This attitude combined with the fact that the user received little feedback about the performance of the system resulted in a situation in which the user's trust was largely misplaced and the robot made numerous collisions. One of the factors that made judging the robot's performance difficult from the user's perspective was the fact the user and the robot were not collocated, so the user could not easily observe the robot or its environment. This separation is not normally the case with most automation systems.

Advances in social robotics may exacerbate situations in which users mistrust a system. Related work by Parasuraman and Miller [33] described how automation etiquette has an impact on user trust and overall human-automation performance in traditional automation tasks. Their study showed that "good automation etiquette can compensate for low automation reliability," suggesting that people place more trust in a polite robot than is warranted by the robot's actual abilities. In addition, van Mulken [44] found that displaying information in a personified manner did not affect trust. This indicates that designers of social robots may need to consider how their robot presents information (what level of politeness and personification) in order to foster appropriate levels of trust in the system.

Due to potential problems with trusting a robotic system too much or too little, it is important to develop a model that will allow a robotic system to estimate its user's current level of trust. In this paper, we present an overview of previous work in trust in automated systems and discuss specific areas in which robotics poses challenges not previously addressed by this body of work. We also discuss a preliminary web-based questionnaire which we conducted to examine people's attitudes toward robotic automation.

[1] University of Massachusetts Lowell, One University Avenue, Lowell, MA 01854. Email: {mdesai,kstubbs,holly}@cs.uml.edu
[2] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. Email: steinfeld@cmu.edu

2 Research on Trust in Automation

Parasuraman and Riley define automation as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" [34]. Automation has traditionally been employed in systems that are complicated, tedious, or time critical, but it has also been used for economic reasons [34]. When automation was first introduced in the 1930's, its use was limited to large industries; however, at the present, automation can be found in many places, from home appliances to the Mars rovers.

Automation has always had weaknesses: namely, it has only been effective in well-structured and controlled environments and continues to remain so. To avoid catastrophic failures in safety critical systems due to either flaws or limitations of automation, an operator must be present at all times to take control of the system. Situations of this kind in which a human operator is working with an automated system are referred to as "human-in-the-loop control." While utilizing a human operator may be beneficial in certain situations, addressing the inadequacies of automation for the human-in-the-loop control creates a different set of problems. When an operator is added to the system, improving overall system performance requires more than simply optimizing operator performance and, separately, optimizing automation performance. The interaction between the two needs to be considered as well.

For several decades, researchers in the automation field have examined the factors which influence people's trust of automated systems and how this level of trust, in turn, effects the way in which people use, misuse, or disuse automation. Researchers have shown that trust influences operators' use of automation (e.g., [8, 19, 31]): the more operators trust automation, the more they tend to use it. Moreover, if an operator trusts his own abilities more than those of the automated system, he tends to choose manual control; however, if an operator trusts the automation more than his own capabilities, he is more likely to choose automation over manual control.

Not only has it been demonstrated that a user's level of trust affects how much he will rely on an automated system, but numerous studies have also been conducted to examine the factors which influence this level of trust (see [20] for an overview). Specifically, Dzindolet et al. [8] demonstrated the impact of system performance on user trust. The results of this study indicated that while users initially placed trust in a decision aid and agreed with its suggestions, as users observed the system making errors they would distrust even a generally high-performing aid unless provided with reasons as to why the errors had occurred. However, providing this type of information increased trust in the automated aid even when the aid performed poorly. Besides the performance of the system, additional factors contributing to a user's trust of an automated system include the recency of errors made by the system, the user's prior knowledge about the system's performance, the user's knowledge of the capabilities of the system, and the user's expectations of the system's performance [39].

Different models have been hypothesized regarding how these different factors influence each other and ultimately the operator's reliance on automation. One such model was proposed by Riley [37] and is shown in Figure 2. The dashed lines indicate the unproven hypotheses and the solid lines indicate relationships that have been proved by experiments. This model, like most, does not consider some factors that are relevant to robots such as interface usability, proximity to robot (co-located or remote-located), situation awareness, dynamics of operating environment, etc. These factors heavily influence automation in robotics. Since the performance of automation has an effect on users' trust of automation it is important to consider these factors. Of these factors, some work has been done on interface usability [2] and situation awareness [23].

Figure 1 shows a generic trust model which has been augmented with factors more relevant to robotics. Many researchers have proven that the way the user allocates control or uses automation depends on the amount of trust that the user has in the automation and the amount of trust that user has in his own capabilities (e.g., [5, 8, 19, 21, 37]). A user's trust of his own capabilities is most often referred to as "self-confidence." Some studies have reported a certain amount of lag between changes in trust and self-confidence and an actual change in allocation strategy; this lag is referred to as "inertia" [21]. Atoyan et al. found that interface design plays an important role in influencing users' trust in automation [2].

When the user changes the allocation strategy, the performance of the system inevitably changes. For the feedback loop to close, the user needs to observe this change in performance. Depending on the system, there might be a significant time delay before the user observes the change in performance. This time delay may result because only cumulative feedback is provided to the user [8], or it may be a result of the time required to send information to the user over the communication channel. The significance of a particular amount of time delay depends on the nature of the system. For example, a delay of a few hundred milliseconds will have a drastic effect on a USAR system but may have little effect on a Mars rover. The change in system performance is also dependent on machine accuracy, and, as explained in Section 3.1, the performance of automation in robotic systems is generally not very high.

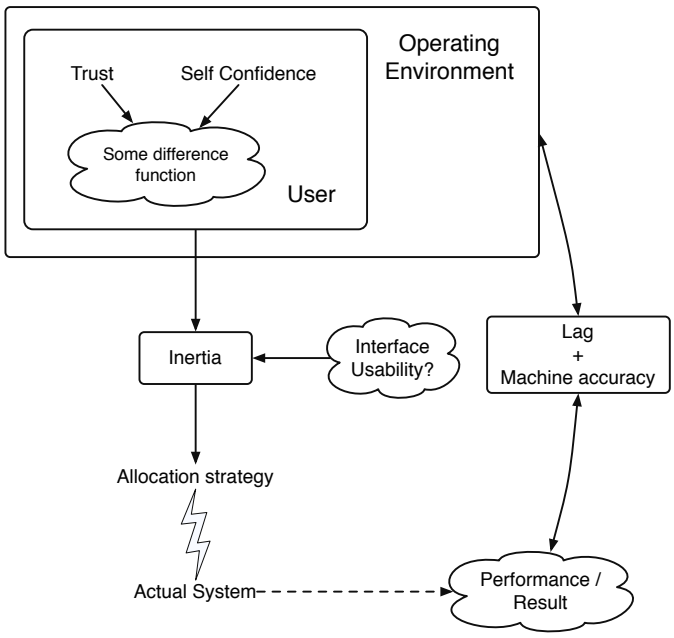

Figure 1. A basic model of trust adapted for robotics. It shows that trust and self-confidence influence automation allocation along with some factors that would be more relevant to robotics like interface usability, lag, and machine accuracy.

Table 3 at the end of this paper lists previous work related to trust and automation. For a more comprehensive coverage of trust and au-

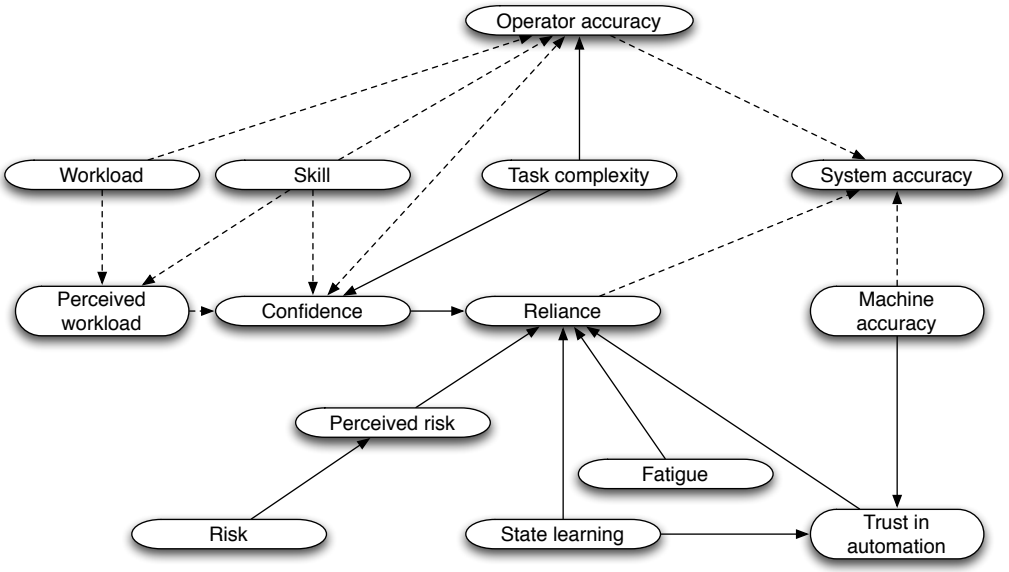

Figure 2. Factors influencing automation use, according to Riley [37]. The dashed lines indicate the relationships that are yet to be proven and the solid lines indicate the relationships that have been experimentally proven.

tomation, see Lee and See [20]. Table 3 lists experimental setup details such as the number of participants, background of participants, and experiment task. It also lists the main contribution of the papers. Most of the experimental setups were some form of automated decision aids. There were two studies that were based on questionnaires ([16, 26]). Most of the experiments were modeled after systems that were not very complex, such as orange juice pasteurization [19], character identification [37], camouflaged soldier detection [9], etc. Most of the studies recruited students. One study in which pilots participated was [37]. It is important to note that while most of the studies utilized simulations, some of the simulations were modeled after real systems ([19, 28, 30]) while others were not.

In one of their experiments with simulated pasteurization plants, Lee and Moray [22] found that automation allocation was dependent on trust and self-confidence. They also found that operators exhibited inertia, that there was a bias towards manual control, and that this bias was even more prominent during the initial stages. In another experiment, Lee and Moray also found that automation usage is dependent on individual biases [19]. The bias towards manual control was also validated by Riley in his experiments with a character identification system [37]; due to these findings in a small sample set of users, he suggests running larger sets of participants.

As can be seen in Table 3, most trust in automation studies have had relatively few participants. Riley also conducted similar experiments with students and pilots and found that automation allocation strategy followed the same pattern for both groups except that the pilots relied on automation more than the students did. This difference raises the question of testing systems with domain experts, which is seldom done in automation research but is standard in human-computer interaction for usability studies.

Dzindolet et al. conducted experiments with an automated decision aid for camouflaged soldier detection [9]. In one of their experiments, they found that participants who had little information about the reliability of the system considered the system trustworthy. This bias is referred to as the "positivity bias". The concept of the positivity bias was developed in the field of social psychology; the idea is that people tend to be biased to think highly of other people when adequate information about them is not available. If participants had a positivity bias towards an automated system, the participants would be more likely to use automated control (at least initially). However, Dzindolet's findings contradict those of Lee and Moray [22] and Riley [37], in which users showed an initial bias towards manual control.

3 Automation and Trust in Robotics

While the broad range of automation research provides a context for examining issues of trust with robots, there are a number of factors that limit how well this work generalizes to the robotics domain. For example, studies of automated systems have tended to utilize systems such as autopilots, flight management systems, vision systems for target or obstacle identification, and factory control systems [20]. Participants interact with a simulated system, which allows experimenters to inject errors and observe how participants' level of trust of and use of the system change as a result. The systems used in these experiments generally have no physical embodiment and do not interact with the physical world. Furthermore, these automated systems tend to be designed for rigid tasks; that is, each system performs only one very specific type of task. Robots are generally designed to accomplish a wider range of tasks: an assistive robot might be needed to fetch a cup of coffee from the kitchen one day and a hat from the closet the next, and a robotic system for USAR can be deployed at different disaster areas with different physical layouts and characteristics, such as a collapsed building or a mine shaft. In fact, it is quite common for robots to be used for tasks their designers never envisioned. Thus, the implications of these automation studies for physically embodied robots with noisy sensors operating in dynamic environments are less clear.

3.1 Automation Reliability

Reliability is one characteristic which has been shown empirically to affect a user's trust in a system [5, 9, 37]. Designing automation for robotics is more challenging than for other automated systems because of the difficulties in modeling the robot and its environment and the challenges posed by poor sensor data. Designing automation requires modeling an existing system. Regardless of the implementation architecture, building such a model requires that all possible states be mapped to an action. This mapping is easy to do for traditional automated systems, such as the orange juice pasteurization plant utilized in Lee and Moray's experiments [19]; however, creating this mapping is difficult to do when the number of possible states is very large or dynamic. Designing automation for such systems requires many approximations, reducing the reliability of automation even in the presence of perfect sensor information. Robots designed for urban search and rescue, assistive technology, or unmanned surveillance must operate in unknown, unstructured, and dynamic environments. This lack of environmental constraints makes designing automation to cover all possible circumstances the robot might encounter very difficult. As a result, these types of robotic systems are likely to have lower reliability than other automated systems.

Automation performance not only depends on the implementation of automation by the designers, but it is also heavily dependent on information from sensors. Most automation experiments conducted regarding trust have been in simulation (e.g., Table 3). In these experiments, the reliability of automation is artificially controlled; most often, the reliability is constant throughout the entire experiment or constant for relatively large time periods. This type of experimental design may help to highlight the effect that reliability has on trust and control allocation; however, since the resulting models are derived from simulated systems, their applicability to real, physical systems remains unclear. This issue is very important in robotics because most sensor modalities used are either unreliable (i.e., sonar, infrared, etc.) or their accuracy is dependent on environmental factors (such as light, reflectivity of surfaces, other characteristics of surfaces, etc). This suggests that at the very least, for robotics, simulated experimental setups with reliability modeled after real world systems or real world experiments are needed. While studies have been conducted which examine the effects of reliability on trust, given the dynamic nature of reliability in robotics, a comprehensive study of the effects of reliability in robotics is required to either validate existing trust models or modify them.

3.2 Automation Capability

Automation capability defines to what extent system operations can be controlled by automation, whereas automation reliability defines how well system operations can be performed. In most previous automation experiments, automation capability has not been an issue; however, in robotic systems, the maximum capability of automation must be considered separately from reliability.

In most automation experiments conducted to date, the automated system is capable of performing the entire operation. This capability is due in part to the oversimplified nature of the experimental setup. An example would be controlling the valves in a pasteurization plant or correctly indicating the presence of a camouflaged soldier in an image. From the operator's perspective, the mental model of the automation's capability is relatively static: the automation always has the capability to complete the task at hand; however, this may not be the case with robotics. In several robotic domains, systems employ discrete levels of autonomy for different reasons. These levels of autonomy define what tasks the automation can perform. Parasuraman et al. have described a classic example illustrating how different levels of autonomy can be utilized [35]. USAR systems developed by INL [3], and UML [17] have several discrete autonomy levels, requiring the user to accurately learn, remember, and use the correct mental model for automation. The need for a complicated model of automation capability has the potential to drastically increase the chance of misplaced trust in automation. UML has also implemented sliding scale autonomy for USAR [6]. According to Desai and Yanco, a sliding scale autonomy system is a continuum of autonomy levels from full teleoperation to full autonomy [7]. From an automation point of view, a sliding scale system might be better than a discrete autonomy system, but the influence of such sliding scale autonomy on trust has yet to be studied.

3.3 Changing Levels of Autonomy

Furthermore, the fact that robotic systems may operate under changing levels of autonomy is generally not addressed in this literature. In some robotic systems, the user specifies the desired level of autonomy; in others, the robot may change its own level of autonomy without specific human direction to do so. Desai [6] lists factors in the robotics application domain that could govern the minimum as well as maximum amount of automation a system can have. Some robot application domains, such as urban search and rescue (USAR), can have very unstructured environments [29], which require the presence of a human operator to assist the automated system. A robotic system with different levels of autonomy requires that the user develop an allocation control strategy to decide how much autonomy the system should have at any given time. However, most of the research regarding allocation control strategy that has been done has been mainly in the field of industrial automation or aviation automation. Most often in such situations, the automation can either be turned on or turned off (e.g., [5, 21, 37]). Most autonomous robotic systems employ a discrete autonomy system in which there are several autonomy levels to choose from, which complicates the problem of automation allocation: the user must now decide not only whether or not automation must be used but also how much; thus the level of autonomy must also be considered as a factor related to the user's trust of the system.

In robotics, adjustable autonomy systems are sometimes referred to as mixed initiative autonomy (MIA) systems. However, the term "adjustable autonomy" is also used to refer to systems in which the operator and the automated system can both change the level of autonomy. A system like this was implemented by Desai [6]. In domains like USAR, in which the operator may utilize a robot for long periods of time, it is very difficult for the operator to maintain situation awareness the entire time. The widely accepted definition of situation awareness is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [10]. Situation awareness can be easily lost if the robot is operating autonomously. This lack of good SA might result in an accident if the automation fails for some reason. MIA systems can prevent these types of failures by gradually handing over control to the operator as the automation starts to fail. By transferring control to the operator gradually, the operator gains time to regain SA. The MIA system implemented by Desai [6] can also take over some amount of control from the operator if the automation detects that the opera-

tor is performing errors that the automation could avoid if it were in control. The effects of automation failures on trust in MIA systems have not been studied. Some researchers have tried to treat robots as peers rather than just tools (e.g., Marble et al. [25] and Fong et al. [11]). Such systems would implement some sort of MIA, which makes studying the effects of MIA on trust even more important.

3.4 Prior Work on Trust in Robotic Systems

To date, there has been little work examining issues of trust directly with robots, although some work has been conducted involving simulated robots. Dassonville et al. [4] conducted a study in which participants used a joystick to control a simulated PUMA arm. Errors were introduced into the simulation, and participants were asked to rate the reliability, performance, and predictability of the joystick's behavior (as well as how difficult it was to make such ratings). The results of the study were consistent with prior work in autonomous systems that suggest that the user's self-confidence is a significant factor which influences use of such systems.

More recently, Freedy et al. [13] have examined trust in the context of mixed-initiative command and control systems using the MITPAS (Mixed-Initiative Team Performance Assessment System) Simulation Environment. The researchers constructed a quantitative measure of trust by assuming that people use a rational decision model such that "trust behavior is reflected by the expected value of the decisions whether to allocate control to the robots on the basis of past robot behavior and the risk associated with autonomous robot control" [13]. For their evaluation, participants were asked to assume the role of a controller of an Unmanned Ground Vehicle (UGV); the UGV was able to autonomously target and fire, but participants were instructed to take control of the UGV if its autonomous behaviors would lead to a time delay or a failure. The experimenters varied the competency of the UGV's firing behavior and recorded participants' choices to override the UGV. The results suggested that participants who were able to ascertain whether the UGV was very competent or incompetent adjusted their behavior accordingly, seeming to trust the system to continue to maintain the same level of competence. In the case in which the system was of indeterminate competence, it was more difficult for users to adjust their behavior. Because the entire experimental setup took place in a video game-like environment, it is unclear how these results would generalize to physical robots.

While relatively little work has been done investigating trust in robots, there is a large body of research on trust in different types of technologies. Because we are interested in developing a model of trust for human-robot interaction, we have examined trust models that were developed for other technology domains. For example, Song et al. [40] developed a neural network-based trust model for understanding users' acceptance of recommendations from a system of heterogeneous agents. Another agent-related trust model was developed by Rehak et al. [36], who used fuzzy numbers to represent trust in cooperating ubiquitous devices. McKnight et al. [27] developed a trust model to understand users' acceptance of a website offering legal advice. However, all of these systems differ from robots in the same ways that the systems previously studied in the automation literature do. To advance the field of human-robot interaction, a systematic study of trust in human-robot interaction is necessary to build trust models in this domain.

4 Web-Based Questionnaire on Attitudes Towards Robotic Automation

To examine people's perceived level of comfort with robotic automation, we have conducted a preliminary web-based questionnaire. Participants were recruited through Mechanical Turk, a website which allows individuals and companies to post human intelligence tasks which are accepted and completed by online workers [42]. Participants were asked about a (fictitious) new car that had the ability to park itself automatically. In order to examine participant's initial biases towards the system, participants were not given any information about the competence or reliability of the fictitious car. Participants were asked to envision parking at a grocery store and to rank the following situations in order of how comfortable they would feel in each situation from 1 (most comfortable situation) to 6 (least comfortable situation):

- You park your car manually.
- Another driver manually parks their car next to your car.
- Another car automatically parks itself next to your car.
- Your car automatically parks itself (and you cannot override it).
- Your car automatically parks itself (but you can override it).
- You take a taxi and the taxi driver parks the taxi.

We received 176 responses to the questionnaire (69.3% female, 30.1% male, 0.6% unknown). Participants reported their ages as follows: 18 to 25, 22.1%; 26 to 35, 36.3%; 36 to 45, 22.1%, 46 or older, 18.1%; unknown, 1.1%. 97.7% of respondents reported having prior experience driving a car. The mean rankings of each scenario are shown in Table 1. We conducted a Friedman two-way analysis of variance to compare the rankings for the scenarios, which provided evidence for significant differences among the six rankings, $\chi^2(5) = 319.79$, $p < 0.001$. In order to determine which scenarios' rankings were significantly different from each other, we used Wilcoxon matched-pairs signed-ranks tests with the Bonferroni correction (Table 2).

Overall, 65% of respondents indicated they would be most comfortable manually parking their own car (mode rank = 1), and 55% of respondents indicated they would be least comfortable if the car was parking itself and they had no means to override it (mode rank = 6). While a taxi passenger also has no means to "override" a taxi driver, participants tended to feel less comfortable with the automated system. Similarly, 38.6% of participants ranked the situation in which another driver manually parks a car next to theirs as the second most comfortable (mode rank = 2), yet 36.4% of participants ranked the situation in which another car automatically parks itself next to theirs as their second to least comfortable (mode rank = 5), even though they would have no control over the other car in either case. This suggests that people may tend to trust a robotic system less than another person even in circumstances for which there may not logically be any difference in terms of the person's control over the situation.

Nomura et al. have demonstrated empirically that people's negative attitudes towards robots will affect their interactions with robots [32]. The types of biases which we observed in this survey represent negative attitudes which may impact how robotic automation is utilized. In order to improve HRI given these biases, a robotic system will need to adjust its interactions to align with the amount of trust placed in it by the user. For example, the robotic system could alter of warning thresholds (e.g., collision warnings or status warnings such as battery level), how often it asks for help, and which level of autonomy to use. Strategies for modeling human behavior within a robot have been examined (e.g., [12]), but more work is needed.

Table 2. p-values from Wilcoxon matched-pairs signed-ranks tests between the rankings for each scenario. All values less than 0.003 (adjusted from 0.05) indicate statistical significance.

Scenario	Self Manual	Another Driver Manual	Taxi	Self Auto : Override	Another Driver Auto	Self Auto : No Override
Self : Manual		$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
Another Driver : Manual			0.81	0.60	$p < 0.001$	$p < 0.001$
Taxi				0.48	$p < 0.001$	$p < 0.001$
Self : Auto : Override					$p < 0.001$	$p < 0.001$
Another Driver : Auto						$p < 0.001$

Table 1. Mean ranking, mode ranking, and percentage of participants whose responses matched the mode ranking, where 1 = most comfortable situation and 6 = least comfortable situation.

Scenario	**Mean Rank**	**Mode Rank**	**Participants at Mode**
Self : Manual	1.74	1	65.3%
Another Driver : Manual	3.31	2	38.6%
Taxi	3.36	3	27.3%
Self : Auto : Override	3.19	4	27.3%
Another Driver : Auto	4.36	5	36.4%
Self : Auto : No Override	5.04	6	55.1%

5 Conclusion and Future Directions

Prior work on trust in automated systems provides a foundation for understanding and modeling trust in human-robot interaction, but much work remains to be done. Because of the challenges of modeling a robot's sensors, actuators, and its environment as well as the challenges of interpreting noisy sensor data, automation for robotic systems is likely to be less reliable than the systems used in previous work on automation. In addition, a robot may be used in a variety of situations for a variety of tasks, and a robot may not always have the capability to complete every aspect of the task at hand. Thus, robotic systems may have a lower level of automation capability than the systems utilized in automation research. Adjustable autonomy and mixed-initiative robotic systems, including systems which may change their autonomy level dynamically, introduce additional complexity which may affect the user's trust in the system. Our web-based questionnaire also illustrates that users may be biased against robotic autonomy, even compared with situations in which there may be little difference in terms of their own control (i.e., a person riding in a car being parked by another person as opposed to by an automated system).

Further research is needed in order to create models of trust which are specifically tailored towards human-robot systems. The field of HRI should begin to investigate the question of trust through empirical studies, particularly relating to those factors which distinguish robots from other automated systems. Studies in which participants must execute a task with a real robotic system could include measures of perceived reliability and the system's actual reliability to compare how these factors influence participants' use of the system and reported trust of the system. For tasks in which robotic automation is only sometimes helpful, a careful examination of how participants understand the system's capabilities, and how this impacted trust in the system, would also be helpful. Examining the effects of changing levels of autonomy on trust, as well as the effect of automation failures, is another possible research area. In order to build systems which promote appropriate levels of trust, HRI designers will need to consider how to design both the robot's form and its interactions such that it provides feedback which will help the user understand the robot's capabilities and limitations.

Acknowledgments

This work is supported by the National Science Foundation (IIS-0546309).

REFERENCES

[1] S. Antifakos, N. Kern, B. Schiele, and A. Schwaninger, 'Towards improving trust in context-aware systems by displaying system confidence', in *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pp. 9–14. ACM New York, NY, USA, (2005).

[2] H. Atoyan, J.R. Duquet, and J.M. Robert, 'Trust in new decision aid systems', in *Proceedings of the 18th International Conferenceof the Association Francophone d'Interaction Homme-Machine*, pp. 115–122. ACM New York, NY, USA, (2006).

[3] D.J. Bruemmer, D.D. Dudenhoeffer, and J. Marble, 'Dynamic autonomy for urban search and rescue', in *2002 AAAI Mobile Robot Workshop, Edmonton, Canada, August*, (2002).

[4] I. Dassonville, D. Jolly, and A. M. Desodt, 'Trust between man and machine in a teleoperation system', *Reliability Engineering and System Safety*, **53**, 319–325, (1996).

[5] P. de Vries, C. Midden, and D. Bouwhuis, 'The effects of errors on system trust, self-confidence, and the allocation of control in route planning', *International Journal of Human-Computer Studies*, **58**(6), 719–735, (2003).

[6] M. Desai, *Sliding scale autonomy and trust in human-robot interaction*, Master's thesis, University of Massachusetts Lowell, May 2007.

[7] M. Desai and H.A. Yanco, 'Blending human and robot inputs for sliding scale autonomy', in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pp. 537–542, (2005).

[8] M. Dzindolet, S. Peterson, R. Pomranky, L. Pierce, and H. Beck, 'The role of trust in automation reliance', *International Journal of Human-Computer Studies*, (2003).

[9] M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce, and H.P. Beck, 'The role of trust in automation reliance', *International Journal of Human-Computer Studies*, **58**(6), 697–718, (2003).

[10] M. Endsley, 'Design and evaluation for situation awareness enhancement', in *Human Factors Society, Annual Meeting, 32 nd, Anaheim, CA*, pp. 97–101, (1988).

[11] T. Fong, I. Nourbakhsh, C. Kunz, L. Fluckiger, J. Schreiner, R. Ambrose, R. Burridge, R. Simmons, L.M. Hiatt, A. Schultz, et al., 'The peer-to-peer human-robot interaction project', *Space*, **6750**, (2005).

[12] T. Fong, C. Thorpe, and C. Baur, 'Collaboration, dialogue, and human-robot interaction', in *Proceedings of the 10th International Symposium of Robotics Research*, London. Springer-Verlag.

[13] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, 'Measurement of trust in human-robot collaboration', in *Proceedings of the 2007 International Conference on Collaborative Technologies and Systems*, (2007).

[14] A. Glass, D.L. McGuinness, and M. Wolverton, 'Toward establishing trust in adaptive agents', in *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 227–236. ACM New York, NY, USA, (2008).

[15] J.L. Herlocker, J.A. Konstan, and J. Riedl, 'Explaining Collaborative Filtering Recommendations', (2000).

[16] J.Y. Jian, A.M. Bisantz, C.G. Drury, and J. Llinas, 'Foundations for an Empirically Determined Scale of Trust in Automated Systems', (1998).

[17] Brenden Keyes, *MS Thesis: Evolution of a Telepresence Robot Interface*, Master's thesis, University of Massachusetts Lowell, 2007.

[18] M.T. Khasawneh, S.R. Bowling, X. Jiang, A.K. Gramopadhye, and B.J. Melloy, 'A Model for Predicting Human Trust in Automated Systems', *Origins*, (2003).

[19] J. Lee and N. Moray, 'Trust, self-confidence and supervisory control in a process control simulation', in *Proceedings of the Conference on Systems, Man, and Cybernetics*, (1991).

[20] J. Lee and K. See, 'Trust in automation: designing for appropriate reliance.', *Human Factors*, **46**, 50–80, (2004).

[21] J. D. Lee and N. Moray, 'Trust, self-confidence, and operators' adaptation to automation', *International Journal of Human-Computer Studies*, **40**(1), 153–184, (1994).

[22] J.D. Lee and N. Moray, 'Trust, self-confidence, and operators' adaptation to automation', *International Journal of Human-Computer Studies*, **40**(1), 153–184, (1994).

[23] C.L. Liu and S.L. Hwang, 'Evaluating the Effects of Situation Awareness and Trust With Robust Design in Automation', *International Journal of Cognitive Ergonomics*, **4**(2), 125–144, (2000).

[24] P. Madhavan, D.A. Wiegmann, and F.C. Lacson, 'Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **48**(2), 241–256, (2006).

[25] J.L. Marble, D.J. Bruemmer, D.A. Few, and D.D. Dudenhoeffer, 'Evaluation of supervisory vs. peer-peer interaction with human-robot teams', in *Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences*, (2004).

[26] J. Masthoff, 'Computationally Modelling Trust: An Exploration', in *Proceedings of the SociUM workshop associated with the User Modeling conference, Corfu, Greece*, (2007).

[27] D. H. McKnight, V. Choudhury, and C. Kacmar, 'The impact of initial consumer trust on intentions to transact with a web site: a trust building model', *Journal of Strategic Information Systems*, **11**(3–4), 297–323, (2002).

[28] S.M. Merritt and D.R. Ilgen, 'Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **50**(2), 194–210, (2008).

[29] Mark Micire, 'Evolution and field performance of a rescue robot', *Journal of Field Robotics*, **25**(1–2), 17–30, (2008).

[30] N. Moray, T. Inagaki, and M. Itoh, 'Adaptive automation, trust, and self-confidence in fault management of time-critical tasks.', *J Exp Psychol Appl*, **6**(1), 44–58, (2000).

[31] B. Muir, 'Trust between humans and machines, and the design of decision aids', *International Journal of Man-Machine Studies*, **27**(5–6), 527–539, (1987).

[32] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, 'Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes Toward Robots', *IEEE Transactions on Robotics*, **24**(2), 442–451, (2008).

[33] R. Parasuraman and C. Miller, 'Trust and etiquette in high-criticality automated systems', *Communications of the Association for Computing Machinery*, **47**, 51–55, (2004).

[34] R. Parasuraman and V. Riley, 'Humans and automation: use, misuse, disuse, abuse', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **39**(2), 230–253, (1997).

[35] R. Parasuraman, TB Sheridan, and CD Wickens, 'A model for types and levels of human interaction with automation', *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, **30**(3), 286–297, (2000).

[36] M. Rehak, L. Foltyn, M. Pechoucek, and P. Benda, 'Trust model for open ubiquitous agent systems', in *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05)*, (2005).

[37] V. Riley, 'Operator reliance on automation: theory and data', in *Automation and Human Performance: Theory and Applications*, (1996).

[38] V. Riley, 'Operator reliance on automation: Theory and data', *Automation and human performance: Theory and applications*, 19–35, (1996).

[39] J. Sanchez, *Factors that affect trust and reliance on an automated aid*, Ph.D. dissertation, Georgia Institute of Technology, May 2006.

[40] W. Song, V. V. Phoha, and X. Xu, 'An adaptive recommendation trust model in multiagent systems', in *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'04)*, (2004).

[41] Transportation Research Associates, 'Project no. 99-06: Intelligent vehicle initiative (side collision warning system) operational test evaluation'. Final Evaluation Report (Revised), 2003.

[42] Mechanical Turk. `http://www.mturk.com`.

[43] A. Uggirala, A.K. Gramopadhye, B.J. Melloy, and J.E. Toler, 'Measurement of trust in complex and dynamic systems using a quantitative approach', *International Journal of Industrial Ergonomics*, **34**(3), 175–186, (2004).

[44] S. van Mulken, E. Andre, and J. Muller, 'An empirical study on the trustworthiness of life-like interface agents', in *Proceedings of the HCI International*, volume 99, pp. 152–156, (1999).

[45] H.A. Yanco, J.L. Drury, and J. Scholtz, 'Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition', *Human-Computer Interaction*, **19**(1&2), 117–149, (2004).

Table 3. A list of papers that examine factors which influence trust in automation.

Paper	Task/Automation type	Number of Participants	Participant Background	System Type	Significant Findings
Lee and Moray 1991 [19]	Orange juice pasteurization plant	12	Students	Simulation	The time that users spend using automation depends not only on trust and self-confidence but also on their past use of automation and their individual biases.
Riley 1996 [38]	Character identification	30	Students and pilots	Simulation	Among other things, their data did not reveal evidence of positivity bias over the full stimulus experience. In addition, they compared the automation curves of expert and non-expert users (pilots and students). The pilots' responses followed the automation allocation curve of students except that the pilots showed a greater bias towards automation.
Jian et al. 1998 [16]	Word elicitation	157	Students	N/A	Trust and distrust can be treated as opposites, lying along a single dimension of trust.
van Mulken et al. 1999 [44]	Multimedia presentation system	32	Not given	Real	The researchers found no statistical difference between the trustworthiness of personified and unpersonified agents.
Herlocker et at. 2000 [15]	Automated collaborative filtering (ACF)	210	Volunteer users	Real	The researchers seek to find ways to improve performance of an expert system through different explanation modalities and find that, in general, adding an explanation interface to an ACF system improves the acceptance of that system among users. Results were inconclusive.
Moray et al. 2000 [30]	Central heating system	30	Students	Simulation	The researchers answer several questions about adaptive allocation and trust with respect to time-critical systems and faults.
Dzindolet et al. 2003 [9]	Camouflaged soldier detection	219	Students	Simulation	This work provides support for the positivity bias. The results also indicated that, after considerable interaction with the automated decision aid (ADA), the operators were less forgiving of automation errors. However, by providing explanations of why the ADA might make a mistake increased the operators' trust in the ADA.
Khasawneh et al. 2003 [18]	Hybrid inspection system	12	Students	Simulation	Trust is proportional to the performance of the system and can also be predicted based on system errors.
Uggirala et al. 2004 [43]	Line length estimation	12	Students	Simulation	The authors use uncertainty as a quantitative alternative to trust and find that overall trust in a system is inversely correlated with system uncertainty.
Antifakos et al. 2005 [1]	Context aware systems	14	Students	Simulation	Explicitly displaying the current confidence of the system increases the usability of automatic / context-aware systems.
Madhavan et al. 2006 [24]	Visual inspection	45	Students	Simulation	The easiness of errors made by decision aids affects users' trust.
Masthoff 2007 [26]	Perception of trust in a person given a written scenario about that person	49	Volunteers from CS Dept.	N/A	The paper introduces a quantitative trust model that includes factors not generally considered by many researchers, such as direct experiences, stereotypes, reputation, empathy, and user characteristics.
Glass et al. 2008 [14]	Adaptive agent system	14	Employees	Real	The researchers identify and discuss eight major themes that significantly impact user trust in complex systems.
Merritt et al. 2008 [28]	X-ray screening task	255	Students	Simulation	Individual differences account for a large variation in trust when system characteristics are kept constant.