Methods for Developing Trust Models for Intelligent Systems

Holly A. Yanco
Computer Science Department, University of Massachusetts Lowell, One University Avenue, Lowell, MA 01854
and The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730
holly@cs.uml.edu

Munjal Desai
Google Inc., 1600 Amphitheater Parkway, Mountain View, CA 94043
munjaldesai@google.com

Jill L. Drury
The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730
jldrury@mitre.org

Aaron Steinfeld
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
steinfeld@cmu.edu

Abstract

Our research goals are to understand and model the factors that affect trust in intelligent systems across a variety of application domains. In this chapter, we present two methods that can be used to build models of trust for such systems. The first method is the use of surveys, in which large numbers of people are asked to identify and rank factors that would influence their trust of a particular intelligent system. Results from multiple surveys exploring multiple application domains can be used to build a core model of trust and to identify domain specific factors that are needed to modify the core model to improve its accuracy and usefulness. The second method involves conducting experiments where human subjects use the intelligent system, where a variety of factors can be controlled in the studies to explore different factors. Based upon the results of these human subjects experiments, a trust model can be built. These trust models can be used to create design guidelines, to predict initial trust levels before the start of a system's use, and to measure the evolution of trust over the use of a system. With increased understanding of how to model trust, we can build systems that will be more accepted and used appropriately by target populations.

11.1 Introduction

In just one area of the intelligent systems domain, the number of robot systems has greatly increased over the past two decades. According to a survey, 2.2 million domestic service robots were sold in 2010, and that number was expected to rise to 14.4 million by 2014 (IFR 2011). Not only is the number of robots in use increasing, but the number of application domains that utilize robots is also increasing. For example, self-driving cars have been successfully tested on US roads and have driven over 300,000 miles autonomously (e.g., Thrun 2010; Dellaert and Thorpe 1998). Telepresence robots in the medical industry constitute another example of a new application domain for robots (e.g., Michaud et al. 2007; Tsui et al. 2011).

As the use of such systems increases, there is a push to introduce or add additional autonomous capabilities for these robot systems. For example, the Foster-Miller (now QinetiQ) TALON robots used in the military are now capable of navigating to a specified destination using GPS. The unmanned aerial vehicles (UAVs) deployed by the military are also becoming more autonomous (Lin 2008); the Global Hawk UAV, for example, completes military missions with little human supervision (Ostwald and Hershey 2007).

Robots are not the only examples of automated systems. IBM's intelligent agent Watson is now being used as an aid for medical diagnosis (Strickland 2013). Additionally, many of the trading decisions in the stock and

commodities markets are being made by automated systems. Automation has been in use for decades as autopilot systems in airplanes and as assistants for running factories and power plants.

Utilizing autonomous capabilities can provide benefits such as reduced time to complete a task, reduced workload for people using the system, and a reduction in the cost of operation. However, existing research in the domains of plant, industrial, and aviation automation highlights the need to exercise caution while designing autonomous systems, including robots. Research in human-automation interaction (HAI) shows that an operator's trust of the autonomous system is crucial to its use, disuse, or abuse (Parasuraman and Riley 1997).

There can be different motivations to add autonomous capabilities; however, the overall goal is to achieve improved efficiency by reducing time, reducing financial costs, lowering risk, etc. For example, one of the goals of the autonomous car is to reduce the potential of an accident (Guizzo 2011). A similar set of reasons was a motivating factor to add autonomous capabilities to plants, planes, industrial manufacturing, etc. However, the end results of adding autonomous capabilities were not always as expected. There have been several incidents in HAI that have resulted from an inappropriate use of automation (Sarter et al. 1997). Apart from such incidents, research in HAI also shows that adding autonomous capabilities does not always provide an increase in efficiency. The problem stems from the fact that, when systems or subsystems become autonomous, the operators that were formerly responsible for manually controlling those systems are relegated to the position of supervisors. Hence, such systems are often called supervisory control systems.

In supervisory control systems, the operators perform the duty of monitoring and typically only take over control when the autonomous system fails or encounters a situation that it is not designed to handle. A supervisory role leads to two key problems: loss of skill over time (Boehm-Davis et al. 1983) and the loss of vigilance over time in a monitoring capacity (Endsley and Kiris 1995; Parasuraman 1986). Due to these two reasons, when operators are forced to take over manual control, they might not be able to successfully control the system.

As such systems are developed, it is important to understand how people's attitudes about the technology will influence its adoption and correct usage. A key factor shaping people's attitudes towards autonomous systems is their trust of the system; hence, we are striving to learn the factors that influence trust, whether for all autonomous systems or for particular domains. Without an appropriate level of trust or distrust, depending upon the circumstances, people may refuse to use the technology or may misuse it (Parasuraman and Riley 1997). When people have too little trust, they are less likely to take full advantage of the capabilities of the system. If people trust systems too much, such as when challenging environmental conditions cause the systems to operate at the edge of their capabilities, users are unlikely to monitor them to the degree necessary and therefore may miss occasions when they need to take corrective actions.

Thus it is important to understand how develop appropriate levels of trust prior to designing these increasingly capable autonomous systems. Without understanding the factors that influence trust, it is difficult to provide guidance to developers of autonomous systems or to the organizations commissioning their development. In contrast, a knowledge of the way particular factors influence trust can allow a system to be designed to provide additional information when needed to increase or maintain the trust of the system's user in order to ensure the correct usage of the system.

We have seen that trust of intelligent systems is based on a large number of factors (Desai et al. 2012). In our prior work (Desai et al. 2012; Desai et al. 2013), we have found that the mobile robotics domain introduces some different trust-related factors than have been found in the industrial automation domain. There is some overlap, however: a subset of trust factors appears in both domains. Given our prior results, we believe that there is a core set of factors across all types of intelligent system domains that has yet to be codified. Further, it may be necessary to identify factors specific to each application domain.

Our ultimate goal is to understand the factors that affect trust in automation across a variety of application domains. Once we have identified the factors, our objective is to develop a core model of trust. In this chapter, we present two methods for identifying factors influencing trust and for building a trust model.

In the first method, we used online surveys of potential system users to identify the factors that most influence people's trust in two domains: automated cars and medical diagnosis systems. Our goal was to determine the factors influencing trust for these domains and compare them to determine the degrees of overlap and dissimilarity. Based upon these findings, we present a method for developing a core trust model.

In the second method, we used a series of human subjects experiments on a real robot to explore the influence of a number of variables upon people's trust of a robot system. Based upon the findings from these experiments, we built a model of trust.

This chapter describes our research methodology and findings for both methods of modeling trust, concluding with a discussion of the pros and cons of each method.

11.2 Prior Work in the Development of Trust Models

Sheridan and Verplank (1978) were among the first researchers to mention trust as an important factor for control allocation. According to their research, one of the duties of the operator was to maintain an appropriate trust of the automated system. However, the first researcher to investigate the importance of trust on control allocation was Muir (1989). According to Muir, control allocation was directly proportional to trust: i.e., the more trust the operator had in a system, the more likely he/she was to rely on it and vice versa. If the operator's trust of the automated system is not well calibrated, then it can lead to abuse (over-reliance) or disuse (under-reliance) on automation. Since this model of trust was first proposed, significant research has been done that indicates the presence of other factors that influence control allocation either directly or indirectly via the operator's trust of the automated system. Some of the factors that are known to influence trust or have been hypothesized to influence trust are explained in brief below.

- Reliability: Automation reliability is one of the most widely researched and one of the most influential trust factors. It has been empirically shown to influence an operator's trust of an automated system (Dzindolet et al. 2003; Riley 1996; deVries et al. 2003). Typically, lower reliability results in decreased operator trust and vice versa. However, some work with varying reliability indicates that the timing of the change in reliability can be critical (Prinzel III 2002).
- Risk and reward: Risk and reward are known to be motivating factors for achieving better performance. Since lack of risk or reward reduces the motivation for the operator to expend any effort and over-reliance on automation reduces operator workload (Dzindolet et al. 2003), the end result for situations with low or no motivation is abuse of automation.
- Self-confidence: Lee and Moray (1991) found that control allocation would not always follow the change in trust. Upon further investigation, they found that control allocation is dependent on the difference between the operator's trust of the system and their own self-confidence to control the system under manual control.
- Positivity bias: The concept of positivity bias in HAI research was first proposed by Dzindolet et al. (2003). They borrowed from the social psychology literature, which points to a tendency of people to initially trust other people in the absence of information. Dzindolet et al. showed the existence of positivity bias in HAI through their experiments. The theory of positivity bias in the context of control allocation implies that novice operators would initially tend to trust automation.
- Inertia: Researchers observed that when trust or self-confidence change, it is not immediately followed by a corresponding change in control allocation (Moray and Inagaki 1999). This delay in changing can be referred to as inertia. Such inertia in autonomous systems can be potentially dangerous, even when the operator's trust is well calibrated. Hence, this is an important factor that warrants investigation to help design systems with as little inertia as possible.
- Experience: In an experiment conducted with commercial pilots and undergraduate students, Riley (1996) found that the control allocation strategy of both populations was almost similar with one exception: pilots relied on automation more than the students did. He hypothesized that the pilots' experience with autopilot systems might have resulted in a higher degree of automation usage. Similar results were found in our work (Desai 2012) when participants familiar with robots relied more on automation than those participants not familiar with robots.
- Lag: Riley (1996) hypothesized that lag would be a potential factor that could influence control allocation. If there is a significant amount of lag between the operator providing an input to the system and the system providing feedback to that effect, the cognitive work required to control the system increases. This increased cognitive load can potentially cause the operator to rely on the automated system more.

11.2.1 Trust Models

In the process of investigating factors that might influence operator's trust and control allocation strategy, researchers have modeled operator trust on automated systems (e.g., Muir 1987; Lee and Moray 1992; Riley

1996; Cohen et al. 1998; Farrell and Lewandowsky 2000; Moray et al. 2000). Over a period of two decades, different types of trust models have been created. Moray and Inagaki (1999) classified trust models into five categories for which they explain the pros and cons of each type in brief: regression models, time series models, qualitative models, argument based probabilistic models, and neural net models.

Regression models help identify independent variables that influence the dependent variable (in most cases trust). These models not only identify the independent variables but also provide information about the relationship (directly proportional or inversely proportional) between each of the independent variables and the dependent variable and the relative impact of that independent variable with respect to that of other variables. The model presented in Section 11.4.3 is an example of a regression model. These models, however, cannot model the dynamic variances in the development of trust and hence must be used only when appropriate (e.g., simply identifying factors that impact operator trust). Regression models can be used to identify factors that impact trust but do not significantly vary during interaction with an automated system, and, based on this information, appropriate steps can be taken to optimize overall performance. This information can potentially be provided to the automated system to allow it to better adapt to each operator. Regression models have been utilized by other researchers (Muir 1989; Lee 1992; Lee and Moray 1992).

Time series models can be used to model the dynamic relationship between trust and the independent variables. However, doing so requires prior knowledge of the factors that impact operator trust. Lee and Moray (1992) used a regression model to initially identify factors and then used a time series model (AutoRegressive Moving AVerage model: ARMAV) to investigate the development of operator trust. Through that model, Lee and Moray found that the control allocation depends on prior use of the automated system and individual biases, along with trust and self-confidence. Using a time series model requires a large enough data set that can be discretized into individual events. For example, in an experiment conducted by Lee and Moray, each participant operated the system for a total of four hours, which included twenty-eight individual trials (each six minutes long). Qualitative data was collected at the end of each run that might have had a faulty system throughout the run. Unlike most other types of models, time series models can be used online to predict future trust and control allocation and perhaps initiate corrective action if needed. However, to our knowledge, no such models exist.

In qualitative models, the researchers establish relationships between different factors based on quantitative data, qualitative data, and their own observations. As Moray and Inagaki (1999) point out, such models can provide valuable insight into how trust, control allocation, and other factors interact. A model of trust partly based on the human-human model of trust developed by Muir (1989) and the model of human-automation interaction by Riley (1994) takes advantage of two well-established qualitative models. Given the heuristic nature of these models, they cannot be used to make precise predictions about trust and control allocation; however, they can and often have been used to create a set of guidelines or recommendations for automation designers and operators (e.g., Muir 1987; Chen 2009).

Farrell and Lewandowsky (2000) trained a neural net to model the operator's control allocation strategy and be able to predict future actions by the operator. The model, based on connectionist principles, was called CONAUT (Connectionist Model of Complacency and Adaptive Recovery). Their model received digitized information as sets of 10 bits for each task. Using that model, the authors predicted that cycling between automatic and manual control could eliminate operator complacency. While such models can accurately model trust and control allocation strategies, they require large data sets. Due to the very nature of neural networks, it is not feasible to extract any meaningful explanation about how the model works.

11.2.2 Trust in Human-Robot Interaction (HRI)

Human-Robot Interaction (HRI) is a diverse field that spans from medical robots to military robots to social robots to automated cars. While it would be ideal to create a model of trust that generalizes to all of HRI, it is important to narrow the scope of investigation because we hypothesize that the application domain is a significant factor in the trust model. Various taxonomies have been defined for HRI (e.g., Dudek et al. 1993; Yanco and Drury 2004). One such taxonomy for robots defines the system type by their task (Yanco and Drury, 2004). Another possible classification for robots is their operating environment: ground, aerial, and marine robots. The scope of the research described in this paper is limited to remotely controlled unmanned ground robots that are designed for non-social tasks. Unmanned ground robots represent a significant number of robots being developed and hence the contributions of this chapter should impact a significant number of application domains within HRI.

Several application domains within the realm of unmanned ground robots are classified as mobile robots, such as factory robots (e.g., Kiva Systems 2011; CasePick Systems 2011), consumer robots (e.g., iRobot 2011; Neato Robotics 2011), and autonomous cars (e.g., Thrun 2011; Dellaert and Thorpe 1998). However, one of the more difficult domains is urban search and rescue (USAR). USAR robots typically operate in highly unstructured environments (Burke et al. 2004), involve a significant amount of risk (to the robot, operating environment, and the victims), and are remotely operated. These factors that make operating USAR robots difficult also make USAR the ideal candidate for examining different factors that influence trust in HRI.

Along with the models of operator reliance on automation (Riley 1996), the models of trust, the list of known factors, and the impact of these factors on operator trust have been well researched in HAI (e.g., Muir 1989; Moray and Inagaki 1999; Dzindolet et al. 2001). However, the automated systems used for research in HAI and in real world applications differ from the typical autonomous robot systems in HRI and therefore necessitate investigating trust models in HRI. Some of the key differences between typical HAI systems and HRI, along with unique characteristics of HRI relevant to operator trust, are explained in brief below.

- Operating environment: The operating environment of most systems in HAI is very structured and well defined (e.g., automated plant operation or automated anomaly detection). On the other hand, the operating environment for USAR can be highly unstructured (Burke et al. 2004) and unfamiliar to the operator. The lack of structure and a priori knowledge of the environment can limit the autonomous capabilities and can also impact the reliability of the autonomous robots.
- Operator location: When operators are co-located with the autonomous system, it is easy for the operator to assess the situation (e.g., auto-pilots). However, with teleoperated robots, the operator can be up to a few hundred feet or more away from the robot. This physical separation between the robot and the operator makes it difficult to assess the operating environment and can impact the development of trust. While sensors and actuators are not unique to robots, remotely controlling actuators is more difficult with noisy sensors. In most of the experimental methodologies used in HAI, noisy sensors are not used and hence their impact on automation or the operator are not investigated.
- Risk: The level of risk involved in HAI domains varies widely, ranging from negligible (e.g., automated decision aids (Madhani et al. 2002; Dzindolet et al. 2001) to extremely high (e.g., autopilots, nuclear plants). However, the research that does exist mostly involves low risk scenarios (Muir 1989; Riley 1996; Sanchez 2006). In contrast, domains like USAR carry a significant amount of risk that the operator needs to understand and manage accordingly.
- Lag: Unlike HAI, where the input to the system and the feedback from the system is immediate, the delay in sending information to the robot and receiving information from the robot can vary based on the distance to the robot and the communication channel. This delay, ranging from a few hundred milliseconds to several minutes (e.g., in the case of the Mars rovers) can make teleoperating a robot incredibly difficult, forcing the operator to rely more on the autonomous behaviors of the robot.
- Levels of autonomy: Automated systems typically studied in HAI operate at one of two levels of autonomy on the far ends of the spectrum (i.e., completely manual control or fully automated). In HRI, robots can often be operated at varying levels of autonomy (e.g., Bruemmer et al. 2002; Desai and Yanco 2005).
- Reliability: Due to the nature of noisy and often failure prone sensors used in robotics, the reliability of automated behaviors that rely on those sensors is often lower than typically high reliability levels used for HAI research (Bliss and Acton 2003; Dixon and Wickens 2006).
- Cognitive overload: Teleoperating a remote robot can be a cognitively demanding task. Such demands can impact other tasks that need to be carried out simultaneously. Cognitive load can also result in operators ceasing to switch autonomy modes (Baker and Yanco 2004).

Along with these differences, the experimental methodology used for most of HAI research has either been based on abstract systems, micro-worlds, or low fidelity simulations (Moray and Inagaki 1999). These setups cannot be used to investigate the subtle effects of different characteristics listed above. Hence, a real-world experimental scenario will be used to examine trust in HRI in one of the methods presented in this chapter. Section 11.4.1 explains the details of the experimental methodology along with the different factors that will be examined and a motivation for examining them.

11.3 The Use of Surveys as a Method for Developing Trust Models

While experiments that allow people to use real systems can produce valuable insights into the factors that influence trust, the nature of the experimental procedures do not allow for very large sets of people to be included. To allow for a larger set of people to be queried, we decided to explore the use of surveys for developing trust models. We selected two domains to begin with: automotive and medical. Specifically, we focused on driverless cars (e.g., Google Cars) and automated medical diagnoses (e.g., IBM's Watson). There were two dimensions for each survey: the safety criticality of the situation in which the system was being used and name-brand recognition. We designed the surveys and administered them electronically, using Survey Monkey and Amazon's Mechanical Turk. We then performed statistical analyses of the survey results to discover common factors across the domains, domain-specific factors, and implications of safety criticality and brand recognition on trust factors. We found commonalities as well as dissimilarities in factors between the two domains, suggesting the possibility of creating a core model of trust that could be modified for individual domains.

11.3.1 Methodology

We chose the automotive and medical domains for several reasons. The successful completion of over 300,000 miles by Google's driverless car, as well as the rulings in three states and the District of Columbia legalizing the use of driverless cars (Clark 2013), holds much promise for these cars becoming commonplace in the near future. Watson, a question-answering agent capable of referencing and considering millions of stored medical journal articles, is also promising. Little research has been conducted about the public's opinion on IBM's Watson, so the relationship between humans and medical diagnosis agents is uncharted territory.

We felt that the general public could be expected in the future to interact with both automated cars and Watson (in conjunction with their physicians). Thus, we developed computer-based survey instruments that could be administered over the Internet to a wide audience. The surveys resided in Survey Monkey and were accessed via Amazon's Mechanical Turk so that respondents could be paid for their time. The surveys were administered in two rounds, with the first round being exploratory. After making improvements to the surveys, including the addition of factors identified by the initial participants in the "other" category, we released the second round, the results of which are reported in this paper.

Each round of surveys consisted of eight different variations: four for each of the two domains. All of the surveys began with the same demographic questions, including gender, age, computer usage, video game playing, and tendencies to take risks. Then each survey variant had a unique scenario designed to capture differences in public opinions depending on the seriousness of the situation ("safety critical" versus "non-safety critical") and the brand of the automated machine (well-known brand from a large organization versus a brand from an unknown startup). Thus there are four variations for each domain: safety critical and well-known brand ("branded"); safety critical and unknown brand ("non-branded"); non-safety critical and branded; and non-safety critical and unbranded.

In the automotive safety critical scenario, the environment was described as high-speed, with lots of traffic. In the non-safety critical scenario, the environment was described as low-speed with little traffic. While one might argue that all driving is safety critical, clearly it is more difficult to ensure safe travel at higher speeds and with more traffic. It is also more difficult to imagine oneself taking over control from such an autonomous system at high speeds in difficult driving conditions.

In the medical safety critical scenario, the task described was to determine diagnoses and treatments of three possible types of cancer. In the non-safety critical scenario, the respondent was given ample information to be certain that the affliction was not life threatening. The three possible afflictions in the non-safety critical scenario include mononucleosis, influenza, or the common cold. Cancer denotes a greater level of importance and urgency whereas the latter situation seems less dire.

In addition to the severity of the situation, we wanted to see whether the brand of the automated machine affected people's trust level as well. For the automotive domain, we explicitly described the automated system as being a Google Car for the two branded surveys. For the medical domain, we specified that Watson was a product of IBM in two survey variants. In the remaining survey variants, we did not label the automated machine as either the Google Car or IBM's Watson; instead, we said that a small, startup company developed the systems. In this way, we hoped to identify the extent to which the reputation of the company influences trust in an intelligent system in these domains.

Each survey in the automotive domain presented a list of 29 factors that could influence a person's trust of an automated system; surveys in the medical domain presented 30 factors. This list of factors was determined initially from a literature search, including the factors from Desai (2012) discussed below in the results section. We started with a shorter list in the initial design of our surveys; we released each of these initial surveys to small sample sizes (25 per survey; 100 in each domain, for a total of 200). Based upon these preliminary results, we added some additional factors, which were identified by respondents in a free-text "other" field. This process resulted in the full list factors for each domain used in the second version of the surveys, some of which were specific to the particular automation domain and others that were common to the two. The results presented in this paper are from the second version of the surveys, with 100 respondents for each of the eight survey variants.

The surveys also included three test questions used to ensure that respondents were actually reading the survey and answering to the best of their ability: "this sentence has seven words in it," "most dogs have four legs," and "the influence of the color of one's shirt" on their trust of an autonomous system. If a respondent answered one or more of these test questions incorrectly, their data was removed from the dataset.

We created each survey on Survey Monkey and utilized Amazon Mechanical Turk to disseminate them to the public. We narrowed our pool to residents of the United States with a minimum age of 18. We paid each respondent $0.90 to complete the survey. This human subjects research was approved by MITRE's IRB.

11.3.2 Results and Discussion

We released 100 HITs (Human Intelligence Tasks) on Mechanical Turk for each of the versions of our surveys. Each survey had 83 questions, similar except for the wording that pertained to branding/not and safety critical/not. After discarding responses that had one or more of the test questions described in Section 11.3.1 answered incorrectly, we had 382 responses in the medical diagnosis domain (231 male, 151 female; mean age 31.1 (9.0)) and 355 in the car domain (191 male, 164 female; mean age 35.6 (12.6)).

For the medical domain, we had 91 valid responses for the branded and safety critical version, 101 for branded and not safety critical, 97 for non-branded and safety critical, and 93 for non-branded and not safety critical. For the automotive domain, we had 90 valid responses for the branded and safety critical version, 92 for branded and not safety critical, 82 for non-branded and safety critical, and 91 for non-branded and not safety critical. The gender and age demographics were not significantly different between the survey versions for each domain.

For each of the trust factors, respondents were asked to rank how the factor would influence their trust in the system on a 7 point Likert scale, with 1 meaning "strongly disagree" and 7 meaning "strongly agree." The results for the trust factors were aggregated for the automotive domain and for the medical domain. In Tables 11.1 and 11.2, we present the list of factors sorted on the mean score from the Likert scale; while a Likert scale is not a continuous scale and averaging the responses is not strictly correct, it does allow us to see which factors have greater influence on trust across the respondents. Due to this limitation of a Likert scale, we discuss our results in terms of the top, middle and bottom thirds, rather than a strict ordering based upon the mean.

In both domains, the ability of a system to stay up-to-date, statistics about its past performance, and the extent of the research on the system's reliability are important factors for influencing trust in the system, appearing in the top third in both domains. In the middle third, both domains included the person's own past experience with the system, the reputation of the system, the effectiveness of the system's training and prior learning, and observing a system's failure. These common factors could form the basis of a model of trust for automated systems; of course, we need to expand our work to many other domains in order to discover the true core.

In the bottom third, both domains include the system's possibility of being hacked, the system's user-friendliness, its ability to communicate effectively, the popularity of the system, and the aesthetics of the system. These factors are being judged as unimportant to trust by respondents in both domains. However, there may be some domains where issues related more to user interface and the usability of the system could come into play. For example, in a social robot domain such as companion robots for the elderly, the way the system looks could have a greater influence on the user's trust of the system: a pet-like robot covered in fur might be more trusted than a more machine-like system showing metal and wires, for example.

We found that there are domain specific factors present in the top third of the list. For the medical domain, respondents ranked the accuracy of the diagnosis, verification of the diagnosis, and the doctor's ability to use the machine in the top third. In the automotive domain, reliability also ranked in the top third through several of the factors. In our survey design, we elected to have a number of questions about reliability to determine if there were different aspects of reliability. While we did see some differences, the list of factors could be reduced by using

	Automotive Domain				
	Rank	**Ref**	**Influence Factor**	**Mean**	**Std dev**
Top Third	1	A	Statistics of the car's past performance	5.98	1.32
	2	B	Extent of research on the car's reliability	5.87	1.39
	3		My own research on the car	5.82	1.33
	4		Existence of error/problem indicators	5.79	1.49
	5		Possibility that the hardware or software may fail	5.69	1.70
	6		Credibility of engineers who designed the car	5.69	1.53
	7	C	The car's ability to stay up-to-date	5.64	1.53
	8		Technical capabilities of the car	5.55	1.55
	9		Your understanding of the way the car works	5.54	1.48
	10		Your past experience with the car	5.53	1.66
Middle Third	11	D	Reputation of the car	5.49	1.57
	12		Level of accuracy of the car's routes	5.49	1.48
	13		Amount of current roadway information available to the car (e.g., weather, traffic, construction, etc.)	5.43	1.65
	14	E	Effectiveness of the car's training and prior learning	5.41	1.63
	15		Amount of information that the car can access	5.41	1.64
	16	F	Observing a system failure (e.g., making a wrong turn, running a stop light)	5.37	2.00
	17		Accuracy of the route chosen	5.36	1.60
	18		User's familiarity with the car	5.29	1.58
	19		The reputation of the car manufacturer	5.27	1.66
Bottom Third	20		Agreement of routes between car and my knowledge	5.26	1.55
	21		The car's methods of information collection	5.24	1.52
	22	G	Possibility of the car being hacked	5.09	1.94
	23	H	The user-friendliness of the car	5.04	1.69
	24		Amount of verification by your friend of the car's proposed route and driving ability	4.73	1.71
	25		Your friend's training to use the car effectively	4.68	1.99
	26	I	The car's ability to communicate effectively (e.g., accurate grammar, breadth of vocabulary)	4.60	1.88
	27	J	Popularity of the car	3.38	1.74
	28	K	Aesthetics of the car	3.01	1.72

Table 11.1. Rankings of the factors that can influence trust of an automated system in the automotive domain. Factors ranked in the same thirds for both the automotive (this table) and medical (Table 11.2) domains are cross-referenced with letters in the "Ref" column. These common factors appearing in the same third of the rankings give evidence that a core model of trust factors could be developed. The other factors, which are common to both domains but ranked in different thirds or which are domain specific, would be the domain specific factors used to customize the core trust model for a particular domain.

	Medical Domain				
	Rank	**Ref**	**Influence Factor**	**Mean**	**Std dev**
Top Third	1		Accuracy of the diagnosis	6.33	1.04
	2		Level of accuracy of the machine's diagnosis	6.07	1.16
	3	A	Statistics of machine's past performance	6.04	1.20
	4	C	The machine's ability to stay up-to-date	5.97	1.17
	5		Amount of your information available to the machine (e.g., x-rays, physicals, cat scans, etc.)	5.85	1.26
	6		Amount of verification by your doctor of the machine's suggestions	5.84	1.19
	7		Agreement of diagnoses between doctor and machine	5.83	1.32
	8		Doctor's training to use the machine effectively	5.80	1.24
	9		Amount of information that the machine can access	5.79	1.25
	10	B	Extent of research on the machine's reliability	5.79	1.3
Middle Third	11	E	Effectiveness of the machine's training and prior learning	5.63	1.35
	12		Technical capabilities of the machine	5.62	1.31
	13		Existence of error/problem indicators	5.52	1.49
	14	D	Reputation of the machine	5.50	1.41
	15		The machine's methods of information collection	5.46	1.29
	16		Possibility that the hardware or software may fail	5.34	1.64
	17		Credibility of engineers who designed the machine	5.31	1.52
	18		Your past experience with the machine	5.25	1.44
	19	F	Observing a system failure (e.g., making an incorrect diagnosis)	5.15	1.88
Bottom Third	20		User's familiarity with the machine	5.07	1.52
	21	G	Possibility of the machine being hacked	5.06	1.88
	22		My own research on the machine	5.04	1.52
	23		Your understanding of the way the machine works	5.03	1.59
	24		The reputation of the machine's manufacturer	4.87	1.65
	25		Amount of time the doctor consults other doctors	4.74	1.72
	26	I	The machine's ability to communicate effectively (accurate grammar, breadth of vocabulary)	4.61	1.68
	27	H	The user-friendliness of the machine	4.07	1.66
	28	J	Popularity of the machine	3.77	1.72
	29	K	Aesthetics of the machine	2.64	1.74

Table 11.2. Rankings of the factors that can influence trust of an automated system in the medical domain. Factors ranked in the same thirds for both the automotive (Table 11.1) and medical (this table) domains are cross-referenced with letters in the in the "Ref" column.

reliability in place of this group of factors; we will do this when we move to the next phase where we ask respondents to rank trust factors in order of importance.

Of note is where the responsibility of system verification and understanding lies between the two domains. In the top third of the factors in the medical domain, we see that people are looking to the doctor to mediate the results of the automated system. However, respondents are relying more on themselves in the automotive domain. This responsibility can be demonstrated by the fact that "your understanding of the way the [system] works" ranks in the top third for the automotive domain, but in the bottom third for the medical domain. Models of trust for automated systems will need to take into account whether the system is used directly by an end-user or whether it is utilized by a mediator for the end-user. Other such domains might include automated stock trading systems.

In some of our earlier work (Desai 2012), we also utilized Amazon's Mechanical Turk to determine factors that would influence human-robot interaction for novice robot users. To obtain these results, Desai created a series of videos showing robots moving in a hallway environment, which were watched by the survey respondents. Test questions included the color of the robot shown in the video to ensure that the video had been watched. There were 386 valid responses received.

Desai (2012) reports the top six factors that influence trust of a robot system are reliability, predictability, trust in engineers that designed the robot, technical capabilities of the robot, system failure (e.g., failing sensors, lights, etc.), and risk involved in the operation. The factors in the middle third were error by automation, reward involved in the operation, interface used to control the robot, lag (delay between sending commands and the robot responding to them), and stress. The factors in the bottom third were training, situation awareness (knowing what is happening around the robot), past experience with the robot, size of the robot, and speed of the robot.

While our surveys presented 30 possible factors to respondents and Desai's had a total of 17, we see some similarities between the factors in the top third, most notably reliability (although, as discussed above, our surveys presented several questions about aspects of reliability). We also found that trust in the engineers who designed the system was important to our respondents, largely through the different surveys presented for branded vs. non-branded automated systems.

For both application domains, we found a significant difference in people's trust of the system based upon whether the system was made by a well-known company (Google for the automotive domain; IBM's Watson for the medical domain) vs. a "small, startup company." Our surveys had two questions about branding, to which participants answered on a 7 point Likert scale, with 1 meaning "strongly disagree" and 7 meaning "strongly agree." In the first, participants were asked to rate the statement "I trust the machines' capabilities because it was created by ['IBM', 'Google', or 'a small, startup company']." The second statement asked if the participant's "trust in a fully-autonomous system similar to this machine would decrease if it was created by ['a lesser-known company' for the IBM and Google versions or 'a more established company' such as Google or IBM]." Results are shown in Tables 11.3 and 11.4.

Clearly, given these findings, it will require additional work for designers of automated systems to convince users to trust the systems made by small companies. However, one could note that Google was a small, startup company not long ago. Other factors such as past performance of the system can also be used to assist with the trust of a non-branded automated system.

11.3.3 Modeling Trust

A core model of trust can be formulated by looking at the common factors in the top third of the survey results. In the case of this work, the core factors would be statistics about the system's past performance, the extent of research on the system's reliability, and the system's ability to stay up to date. While this last factor will have different meanings in each domain, either referring to the medical information necessary to perform a diagnosis or to road information needed to drive safely and accurately, it is still a core factor that influences people's trust in an automated system.

Factors that rank highly in one domain, but low in the other, could be used as factors to customize the core model of trust for the particular application domain. For example, in the medical domain, we see top ranked factors addressing the accuracy of the system and the doctor's interpretation of the system. In contrast, in the automotive domain, we see more reliance on self-knowledge and the car's ability to convey information.

It is important to note that, given the fact that the users of Amazon's Mechanical Turk skew towards having more education than the average population (Ross et al. 2010), the responses reported in this paper might not be applicable to the general population but instead are only applicable to the population with an undergraduate

	Questions		Mean		T Value
	Brand (Watson)	Non-Brand	Brand	Non-Brand	
Safety Critical	I trust the machine's capabilities because it was created by IBM.	I trust the machine's capabilities because it was created by a small, upstart company.	3.04	2.54	0.006
	My trust in a fully autonomous system similar to this machine would decrease if it was created by a lesser-known company.	My trust in a fully autonomous system similar to the machine would decrease if it was created by a more established company such as IBM.	3.42	2.81	0.003
Non Safety Critical	I trust the machine's capabilities because it was created by IBM.	I trust the machine's capabilities because it was created by a small, upstart company.	3.24	2.67	0.002
	My trust in a fully autonomous system similar to this machine would decrease if it was created by a lesser-known company.	My trust in a fully autonomous system similar to the machine would decrease if it was created by a more established company such as IBM.	3.46	2.87	0.003

Table 11.3. Branded vs. Non-branded Technology: Medical Domain.
Reputation matters: Significant differences were seen for responses for branded automated systems in the medical domain.

	Questions		Mean		T Value
	Brand (Google)	Non-Brand	Brand	Non-Brand	
Safety Critical	I trust the car's capabilities because it was created by Google.	I trust the car's capabilities because it was created by a small, upstart company.	3.19	2.41	0.000
	My trust in a fully autonomous system similar to cars would decrease if it was created by a lesser-known company.	My trust in a fully autonomous system similar to cars would decrease if it was created by a more established company such as Google.	3.67	2.89	0.001
Non Safety Critical	I trust the car's capabilities because it was created by Google.	I trust the car's capabilities because it was created by a small, upstart company.	3.33	2.55	0.000
	My trust in a fully autonomous system similar to cars would decrease if it was created by a lesser-known company.	My trust in a fully autonomous system similar to cars would decrease if it was created by a more established company such as Google.	3.88	2.82	0.000

Table 11.4. Branded vs. Non-branded Technology: Automotive Domain
Reputation matters, part II: Significant differences were also seen for responses for branded automated systems in the automotive domain.

degree or greater. We need to conduct an analysis of the data with respect to education level to determine if there are differences between responses for different levels of education. However, despite this potential limitation of our survey population, we believe surveys like ours can identify factors that will influence trust.

While our surveys allowed people to specify the relative importance of the factors, they did not provide people with the means to indicate whether that factor would result in an increase or decrease in trust. Our next step will be to conduct surveys asking people to choose the top factors which influence their trust, ranking them from most to least important. We will also explore the influence that these factors have upon each other; for example, a system's ability to explain its action influences the system's understandability.

We are also expanding this research to other automated system domains. Our methodology will need to change for some of these domains, as we have been relying on people from the population of Mechanical Turk workers. While such people are well qualified to answer questions about cars and doctor's visits, they will be less qualified to answer questions about the use of automated systems in very specialized domains such as the military or power plants. However, we believe that the use of surveys, whether completed by "average" people or people working in specialized domains, will allow us to identify the top factors influencing trust in automated systems in each domain. As we explore more domains, we will be able to identify those factors that are common to many domains; these factors will form the common core of a trust model.

11.4 Robot Studies as a Method for Developing Trust Models

The primary goal of the research described in this section was to create a better understanding of different factors that impact operator trust and control allocation when interacting with an autonomous remote robot. We also wanted to investigate how certain attributes central to remote robot teleoperation (e.g., situation awareness, workload, task difficulty) impact operator behavior. By observing the variations in the different factors and how they affect operator trust and control allocation strategy, a model of operator interaction specifically for teleoperation of an autonomous remote robot was constructed and has been used to create a set of guidelines that can improve the overall system performance.

11.4.1 Methodology

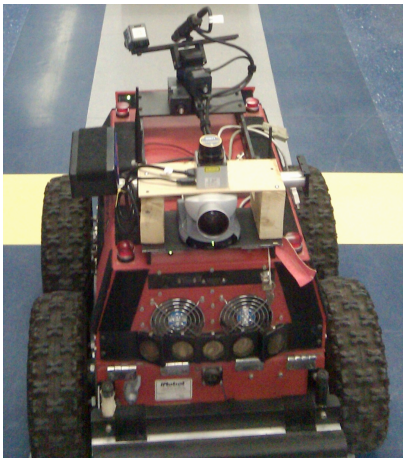

Figure 11.1: The ATRV-JR used in the robot experiments to explore trust factors in human-robot interaction.

The robot used is an iRobot ATRV-JR platform, shown in Figure 11.1. The ARTV-JR has differential drive and a wide array of sensors. These sensors include a front facing SICK LMS-200 laser range finder that can scan 180 degrees, a rear facing Hokuyo URG-04LX laser range finder with a field of view of 240 degrees, a Directed Perception PTU-D46-17 pan-tilt unit with a Sony XC-999 camera mounted on it, and a rear facing Canon VC-C4 camera mounted on the back of the robot. The robot also has a 3.0 GHz Intel Core2Duo processor with 4GB of memory and runs Ubuntu 8.04. It has an 802.11n radio capable of operating on both the 2.4GHz and 5.0GHz range. The client code to control the robot is written in C++ using Player (Gerkey et al., 2003) and compiled using GCC.

Almost all of the prior research in HAI has focused on using two autonomy modes on the far ends of the spectrum. In accordance with this existing research, we decided to provide the participants with two autonomy modes. One of those autonomy modes was at the high end of the autonomy spectrum. Rather than selecting the second autonomy mode to be manual teleoperation mode we decided to opt for a similar autonomy mode where the robot would assist the participants. The key reason was to always keep the participant informed about the robot's behavior, something that would not be possible with a pure manual teleoperation mode. The participants could operate the robot in one of two autonomy modes: robot-assisted mode or fully autonomous mode. The participants were free to select either mode and could switch between them as many times as they wanted. They were also told that there were no benefits or penalties for selecting either mode. When each run was started, no autonomy mode was selected by default, thereby requiring the participants to make an explicit selection. The maximum speed at which the robot moved was the same in both modes and was restricted to approximately 0.41 feet per second. These configurations ensured that the performance of both autonomy modes was similar.

In the fully autonomous mode, the robot ignored the participant's input and followed the hard coded path. The obstacle avoidance algorithm ensured that the robot never hit any object in the course. In the robot-assisted mode, the participant had a significant portion of the control and could easily override the robot's movements, which were based on the path it was supposed to follow. The robot's vectors were calculated the same way in both autonomy modes and were displayed on the user interface (UI) on the laser display to show the participant the robot's planned direction.

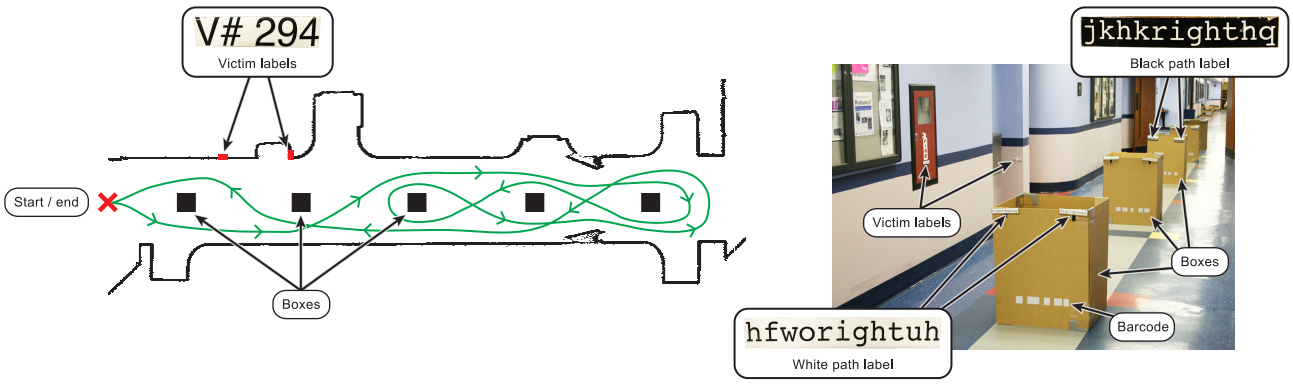

Figure 11.2: The test course used in the experiments.

Figure 11.2 shows the test course designed for these experiments. The course was approximately 60 feet long and had 5 obstacles (boxes) placed about 9 feet from each other. The width of the course was 8 feet. The clearance on either side of the boxes was 3 feet, versus a robot width of approximately 26 inches. Therefore the small clearance on either side of the boxes made it difficult to drive. The course had moderate foot traffic, as it was placed in a hallway in a public building; however, we found that people walking through the area were able to do so in a way that did not interrupt the experiment. Additionally, during each experiment, one of our researchers was with the robot at all times, so could ask people to avoid the robot, if necessary.

The robot started and ended each run at the same location. For each run, the participants had to follow a preset path. Since we planned five runs, we designed five different paths (also referred to as maps) based on the following criteria:

- The length of each map must be the same (~200 feet).
- The number of U-turns in a map must be the same (3 U-turns).
- The number of transitions from the left side of the course to the right and vice versa must be the same (3 transitions).

Since the maps were similar in difficulty and length, they were not counter-balanced. Instead, the maps were selected based on a randomly generated sequence. A sample map is shown in green in Figure 11.2.

Each box on the course had text labels to provide navigational information to the participants. Text labels were placed on top of the boxes to indicate the path ahead. Since the boxes were wide, similar labels were placed on both edges of the face as shown in Figure 11.2, to make it easy for the participants to read the labels as the robot moved past the boxes. The labels indicated one of three directions 'left', 'right', or 'uturn'. These directions were padded with additional characters to prevent the participants from recognizing the label without reading them.

Two sets of labels were necessary to prevent the participants from driving in an infinite loop. Figure 11.2 shows the two types of labels that were used. The labels with a white background (referred to as white labels) were followed for the first half of the entire length and then the labels with a black background (referred to as black labels) for the second half. The transition from following the white labels to black labels was indicated to the participants via the user interface (UI).

The boxes also had barcodes made from retro-reflective tapes that the robot could read using its laser rangefinder (Figure 11.2). While the robot did not actually use these barcodes in the experiments (the localized pose of the robot was used instead to encode the paths), the participants were told that the robot reads the barcodes to determine the path ahead, just like they read the labels. The robot displayed the contents of the bar code on the UI. The path for each run was predefined via a set of navigation waypoints because the robot could not consistently read the barcodes, making it difficult to have a controlled experiment. Based on a constant video compression rate, sampling resolution, and the font size, the labels could be read from about 3 feet away by a participant. The robot simulated reading the labels from approximately the same distance, thereby reducing the potential for a bias to rely on the robot or vice versa. The participants were informed that the robot at times might make a mistake in reading the barcodes and that they should ensure that the direction read by the robot was correct. Participants were also told that if the robot did make a mistake in reading the barcode, it would then

proceed to pass the next box on the incorrect side, resulting in the participant being charged with an error on their score (described below).

The course also had four simulated victims. These victims were represented using text labels like the one shown in Figure 11.2. The victim tags were placed only on the walls of the course between 2.5 feet and 6 feet from the floor. The victim locations were paired with the paths and were never placed in the same location for any of the participant's five runs. While there was a number associated with each victim, the participants were told to ignore the number while reporting the victims. Whenever participants found a new victim, they were told to inform the experimenter that they have found a victim. They were explicitly instructed to only report victims not reported previously. The experimenter noted information about victims reported by the participants and also kept track of unique victims identified.

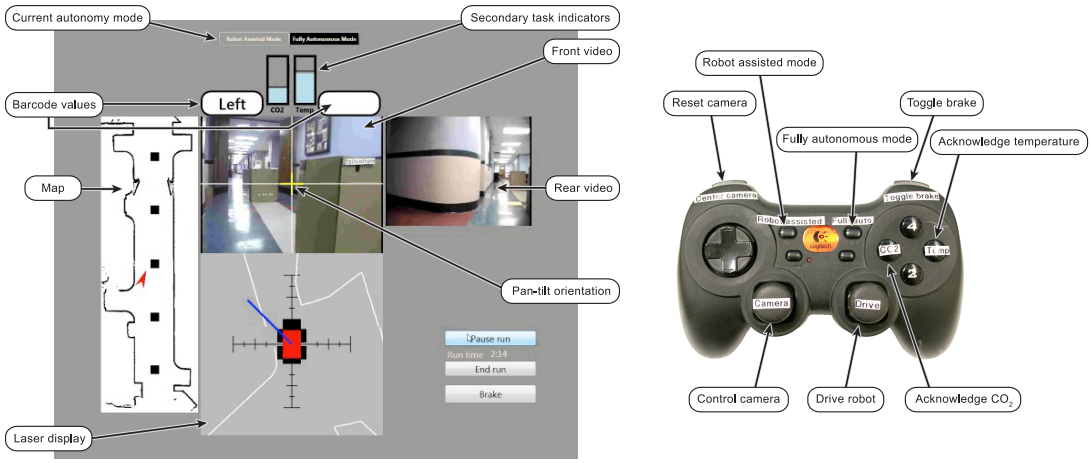

Figure 11.3: The user interface for the robot

Figure 11.3 shows the UI utilized for controlling the robot. The video from the front camera was displayed in the center and the video from the back camera was displayed on the top right (mirrored to simulate a rear view mirror in a car). The map of the course with the pose of the robot was displayed on the left. The distance information from both lasers was displayed on the bottom around a graphic of the robot just under the video. There were vectors that originate from the center of the robot and extend out. These vectors indicated the current magnitude and orientation of the participant's input via the gamepad and the robot's target velocity. The participant's vector was displayed in light gray and the robot's vector was displayed in blue.

The participants provided input using the gamepad shown in Figure 11.3. Participants could drive the robot, control the pan tilt unit for the front camera, select the autonomy modes, turn the brakes on or off, re-center the camera, and acknowledge the secondary tasks.

The participants were asked to drive the robot as quickly as they could along a specified path, while searching for victims, not hitting objects in the course, and responding to the secondary tasks. To create additional workload, simulated sensors for CO_2 and temperature were used. The participants were not told that the sensors were not real. They were also told that the robot's performance was not influenced in any way by changes in temperature and CO_2. The values from the sensors were displayed on the UI (Figure 11.3), which the participants were asked to monitor. Participants were asked to acknowledge high CO_2 and temperature values by pressing the corresponding buttons on the gamepad. The values were considered high when their values are above the threshold lines on the secondary task indicators (Figure 11.3); values over the threshold were indicated by changing the color of the bars from light blue to red, to assist the participants in recognizing the change. The level of workload was varied by changing the frequency with which the values crossed the threshold during multiple robot runs; all participants experienced the same patterns across their runs. The simulated sampling rate for the sensors was kept steady.

In the feedback (Section 11.4.2.2), reduced task difficulty (Section 11.4.2.3) and long-term (Section 11.4.2.4) experiments, the simulated sensor readings were removed from the interface. In their place, participants were asked at regular intervals (every 25 seconds) whether their trust in the robot had increased, stayed the same, or decreased. The answers given were plotted over time during the run. We then defined the area under the trust

curve (AUTC) as a metric that could be used to measure on-line trust, as opposed to end of run measures such as were used by Muir and Jian.

Using higher levels of automation can reduce workload and hence is desirable, especially under heavy workload from other tasks. To prevent participants from using high levels of autonomy all the time, regardless of the autonomous system's performance, it is typical to introduce some amount of risk. Hence, in line with similar studies (e.g., Riley 1996; Lee and Moray 1992; Dzindolet et al. 2002), the compensation was based in part on the overall performance. The participants could select a gift card to a local restaurant or Amazon.com. The maximum amount that the participants could earn was $30. Base compensation was $10. Another $10 was based on the average performance of 5 runs. The last $10 was based on the average time needed to compete the 5 runs, provided that the performance on those runs was high enough.

The performance for each run was based on multiple factors, with different weights for each of these factors predetermined. The participants are told there was a significant penalty for passing a box on the incorrect side, regardless of the autonomy mode. If the participants passed a box on the wrong side, they were heavily penalized (20 points per box). In addition to the loss of score, participants were told that time would be added based on the number of wrong turns they took, but the specific penalties were not revealed. For the first box passed on the wrong side, no additional time was added, to allow participants to realize that the reliability of the system had dropped. For the second incorrect pass, 60 seconds were added, with an additional 120 seconds for the third and an additional 240 for the fourth, continuing with a cumulative increase. Finding the victims was also an important task, so 10 points were deducted for each victim missed.

The scoring formula was not revealed to participants, although they were told about the factors that influence their score. The score for each run was bounded between 0 and 100. If the score was 50 or more, the participants were eligible for a time bonus; if they completed all of the runs in an average of under 11:45 minutes, they received an additional $10. If they had a score of 50 or more and averaged between 11:45 and 15 minutes, they received a $5 bonus. Participants were told about this interdependence between score and time, which was designed to prevent participants from quickly running through the course, ignoring the tasks, while also providing a significant motivation to perform the task quickly.

At the end of each run, the score was calculated and the participants were informed about the amount of compensation that could be received based only on that run. At the end of five runs, the average compensation was calculated and given to the participant.

There were three sets of questionnaires. The pre-experiment questionnaire was administered after the participants signed the consent form; it focused on demographic information (i.e., age, familiarity with technology similar to robot user interfaces, tendency towards risky behavior, etc.). The post-run questionnaire was administered immediately after each run; participants were asked to rate their performance, the robot's performance, and the likelihood of not receiving their milestone payment. Participants were also asked to fill out previously validated trust surveys (Muir 1989; Jian et al. 2000) and a NASA Task-Load Index (TLX) questionnaire (Hart and Staveland 1988) after each run. After the last post-run questionnaire, the post-experiment questionnaire was administered, which included questions about wanting to use the robot again and its performance. These human subjects studies were approved by the University of Massachusetts Lowell's IRB.

After participants signed the informed consent form, they were given an overview of the robot system and the task to be performed. Then participants were asked to drive the robot through the trial course in fully autonomous mode. The experimenter guided the participants during this process by explaining the controls and helping with tasks if necessary. The trial course was half the length of the test course. Once participants finished the first trial run, they were asked to drive the robot again through the same course in the robot-assisted mode. Since there were multiple tasks that participants needed to perform, we decided to first show them the fully autonomous mode, as that would be a less overwhelming experience. Once the participants finished the second trial run, they were asked to fill out the post-run questionnaire. While the data from this questionnaire was not used, it allowed participants to familiarize themselves with it and also helped to reinforce some of the aspects of the run that they needed to remember.

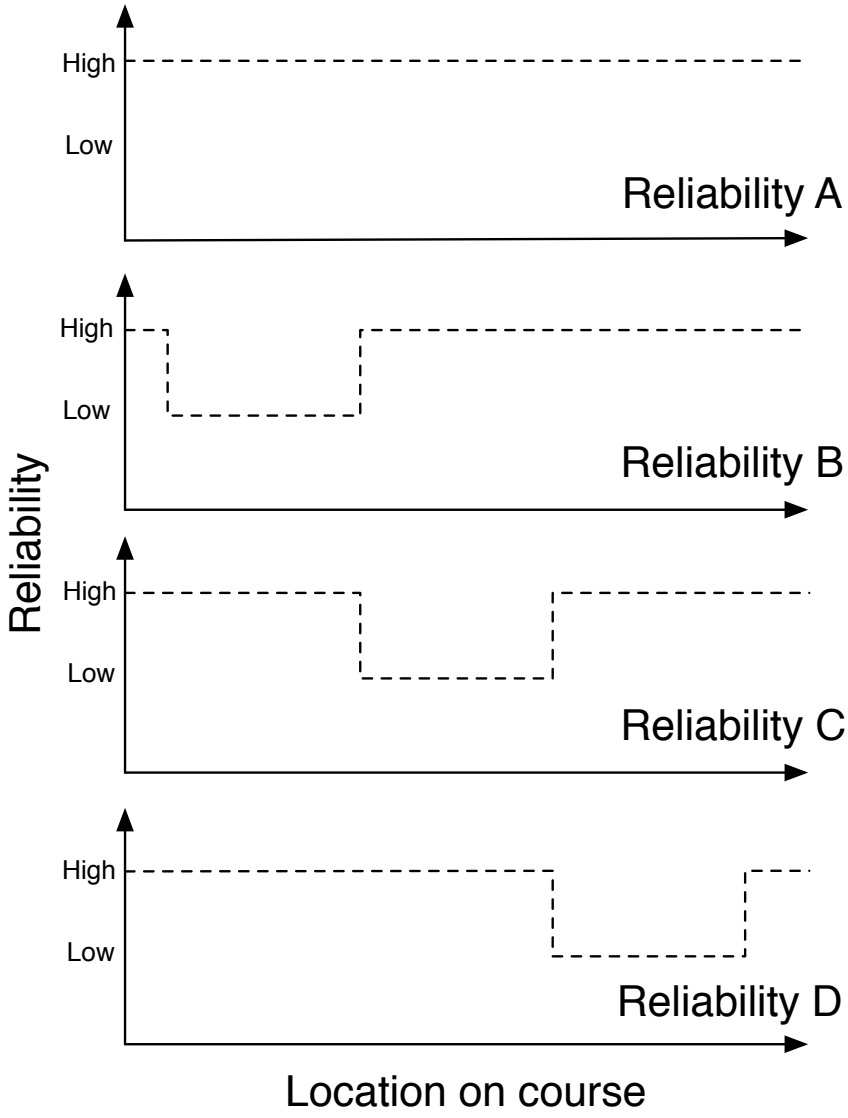

Figure 11.4: Reliability configurations for the robot's runs.

After the two trial runs, the participants were asked to drive the robot for five more runs. In each run, a different map was used. During these runs the reliability of robot autonomy was either held high throughout the run or was changed, according to four pre-planned reliability configuration, shown in Figure 11.4. The changes in reliability were triggered when the robot passed specific points in the course. These locations were equal in length and there were no overlaps. For all four patterns, the robot always started with high reliability. The length of each low reliability span was about one third the length of the entire course. Using different dynamic patterns for reliability allowed us to investigate how participants responded to a drop in reliability at different stages and how the changes influenced control allocation. Every participant started with a baseline run under full reliability (Reliability A in Figure 11.4). Then, the four reliability profiles were counter-balanced for the remaining four runs.

The methodology explained in this section was utilized for all of the experiments. Since multiple factors (e.g., reliability, situation awareness, long-term use) needed to be investigated, it was not feasible to design a within-subjects experiment. Hence, a between-subjects experiment was designed. The overall concept was to conduct multiple experiments, each with two independent variables (e.g., reliability and situation awareness). The dependent variables were the operator's trust and the control allocation strategy. To discern the influence of

reliability and other factors being investigated, a baseline experiment with dynamic reliability (DR) as the only independent variable was conducted first. Data from that experiment was used as a baseline for comparison with data from other experiments.

11.4.2 Results and Discussion

Using the methodology described in the prior section, we conducted experiments to determine how a number of factors would influence an operator's trust and control allocation strategy: lowering situation awareness (SA), providing feedback about the robot's confidence in its current operations, reducing task difficulty, and long-term interaction on operator trust and control allocation. This section presents qualitative models based on the impact of those factors as determined by the experiments (for the full results of the experiments, see Desai 2012)) as well as a set of guidelines, proposed to help better design autonomous robot systems for remote teleoperation and to improve system performance during operation. These models are presented in the context of the Human interaction with Autonomous Remote Robots for Teleoperation (HARRT) model described in the next section.

11.4.2.1 Reducing Situation Awareness (SA)

Figure 11.5 shows the impact of comparing our baseline dynamic reliability (DR) experiment with our low SA experiment (LSA) where the user interface was modified to impact the participant's SA.

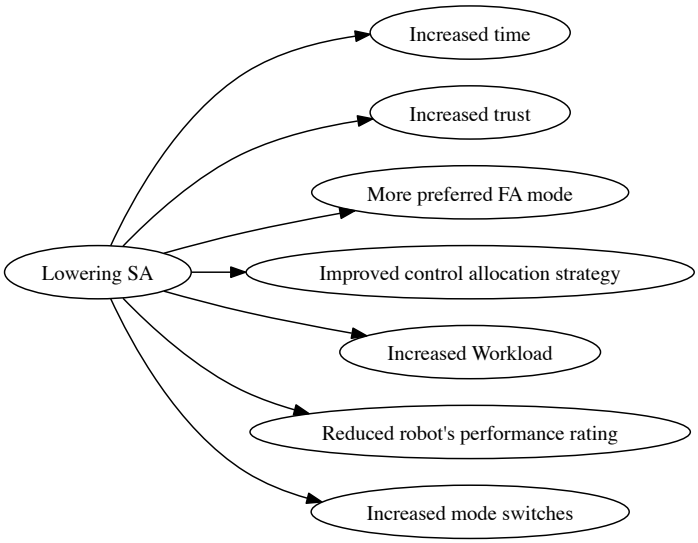

Figure 11.5: The impact of reducing situation awareness (SA) on different factors. All of the effects shown are based on significant differences between the Low Situation Awareness (LSA) and Dynamic Reliability (DR) experiments

As the participants' SA was reduced, it increased their workload. We suspect the increase in workload was due to the additional effort (cognitive and otherwise) required to maintain the minimum required level of SA. Additionally, lowering SA makes the task of remote teleoperation more difficult, which could also increase workload. The combination of increased workload and poor SA increased the time needed to finish the task.

We suspect that lowering SA forced participants into relying more on the fully autonomous (FA) mode. Higher reliance on the FA mode improved the control allocation strategy, since the ideal control allocation strategy required the participants to rely more on FA than the robot-assisted (RA) mode. While the increase in trust was unexpected, it can be explained by the higher reliance on FA for a task that was difficult to perform manually.

Lowering SA also reduced the participants' rating of the robot's performance, even though there was not a significant difference in performance. We suspect this was due to two reasons: poor SA made it difficult to

correctly judge the robot's performance and the participants could have blamed the robot for providing inadequate information needed for teleoperation.

Guidelines based on the SA model, shown in Figure 11.5, are as follows.

Guideline 1: Reduced SA leads to higher reliance on autonomous behaviors. Intentionally reducing SA to force operators to rely on autonomous behaviors is not recommended as a design strategy due to the other undesirable side effects. However, such influence does remain a possibility, but should only be exercised when absolutely necessary, since doing so can potentially impact safety and performance.

Guideline 2: Suspend or defer non-critical tasks when SA is reduced. Even with higher reliance on automation, the workload is expected to increase, so tasks that are not critical should be suspended or deferred to offset the increased workload and to prevent an overall detrimental impact on performance.

Guideline 3: Switch functions unaffected by reduced SA to automation. Functions not impacted by reduced SA can be switched over to automation in an attempt to reduce workload.

Guideline 4: Educate operators about SA. Operators associate robot performance with SA and therefore operators must be informed (during training or during the interaction) that low SA does not necessarily impact the robot's performance.

11.4.2.2 Providing Feedback

Figure 11.6 shows the results of comparing results of the baseline Real-Time Trust (RT) experiment with that of the Feedback (F) experiment where the participants were provided with feedback concerning the robot's confidence in its own sensors.

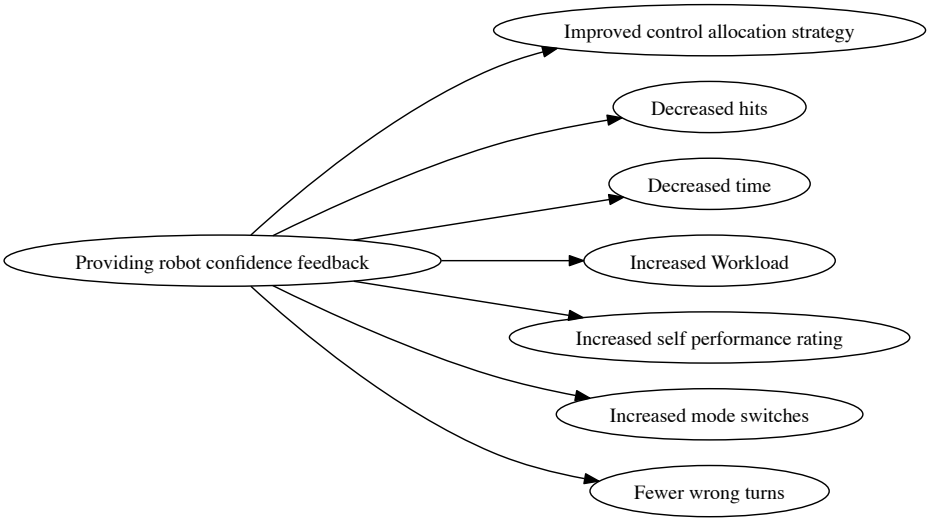

Figure 11.6: The impact of providing feedback on different factors. All of the effects shown are based on significant differences between the Feedback (F) and Real-Time Trust (RT) experiments.

Providing information about the robot's confidence in its own sensors and decision making to the participants increased their workload, as they were given additional information that needed to be processed. Also, participants reacted to the change in robot's confidence by aggressively changing autonomy modes and therefore increased the number of autonomy mode switches. We suspect these autonomy mode changes were another reason that resulted in an increase in workload.

However, increased autonomy mode switches and better robot supervision due to the variations in the robot's confidence resulted in a better control allocation strategy, which in turn led to better performance. Despite the better performance, the participant's trust of the robot did not increase; we suspect this lack of increase in trust was due to the type of feedback provided to the participants.

It is often conjectured that providing feedback should improve an operator's trust in the system by helping operators better align their mental model with that of the system's architecture and operation. However, in this case, the information provided to the participants could not have helped achieve more synchronized mental models. We suspect this discrepancy occurred because no information was provided that could sufficiently explain why the robot made a mistake in reading the labels. Providing such information requires feedback that provides details about the robot's internal processes. For example, informing the user that the robot cannot read labels accurately at certain angles would explain the decrease in the robot's confidence and help the operators better understand the robot's internal operation. The feedback also provided negative information to the participants. It informed the participants that the robot's confidence was medium, low, or at best functioning as intended.

Providing feedback seems to directly impact workload and the operator's control allocation strategy and the impact of feedback on other attributes aligned with the HARRT model. (Figure 11.9 incorporates all of the models described in this section.) Guidelines based on the feedback model are described below.

Guideline 5: Provide feedback only when necessary. There is a cost associated with providing information to operators during their interaction with a remote robot. Therefore, information that is not only important, but also essential for immediate operation should be provided.

Guideline 6: Select the type of feedback based on the desired effect. The type of feedback being provided to the operators must be considered carefully, since it can impact an operator's behavior. The corollary is, that based on the desired effect on operator behavior, different types of feedback can be provided. For example, a temporal impact on control allocation can be expected if the robot's confidence is being presented to the operators. However, if a long-term effect is desired, other means of providing information must be selected. For example, explaining the typical causes for reduction in the robot's confidence could provide the operators with better understanding of the robot and result in a permanent effect. Guideline 5 must be considered while doing so.

11.4.2.3 Reducing Task Difficulty

Figure 11.7 shows the results of comparing data from the baseline Real-Time Trust (RT) experiment with that of the Reduced Difficulty (RD) experiment where the complexity of the teleoperation task was reduced.

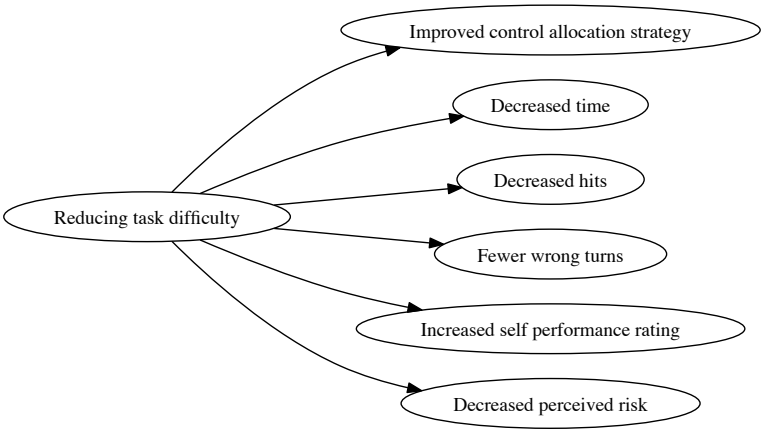

Figure 11.7: The impact of reducing task difficulty on different factors. All of the effects shown are based on significant differences between the Reduced Difficulty (RD) and RT experiments.

With the teleoperation task easier to perform, we expected the participants to not rely on the fully autonomous mode as much, and, consequently, a poor control allocation strategy was expected. However, the control allocation strategy improved along with an increase in autonomy mode switches. We suspect the reduced difficulty of the teleoperation task reduced the participants' workload and allowed them to better observe the robot's performance in the fully autonomous mode. This better robot supervision allowed them to switch autonomy modes appropriately and improve the control allocation strategy. We suspect that improvement in supervision and the resulting increase in autonomy mode switches increased the workload enough to offset the initial reduction in workload due to the easier task.

The easier teleoperation task and the better robot supervision improved performance and safety by reducing the number of hits, reducing the time needed to finish, and reducing the number of wrong turns. Reducing the difficulty of the task seems to primarily impact an operator's control allocation strategy. The impact on other attributes aligns with the HARRT model. Guidelines based on the reduced difficulty model are described below:

Guideline 7: Tasks with reduced difficulty result in better robot supervision and no reduction in workload. If the difficulty of the task reduces during an interaction or for interactions that involve a relatively easy remote robot teleoperation task, operators should be expected to allocate the additional available cognitive resources towards better supervision of the robot's behavior or secondary tasks.

Guideline 8: Do not expect operators to assume manual control for easier tasks. Operators will not necessarily opt for lower autonomy modes, at least in scenarios involving multiple tasks or a relatively high workload. While a reduction in the difficulty of the task will improve performance and safety, the operator's trust of the system will not be affected.

11.4.2.4 Long-Term Interaction

The long-term interaction experiment (LT) was conducted to investigate if an operator's trust and control allocation strategy change over a longer period of time. We looked for trends to incorporate into the model and a set of guidelines. Another goal of the LT experiment was to investigate if there is a difference between operators who are familiar with robots and those who are not.

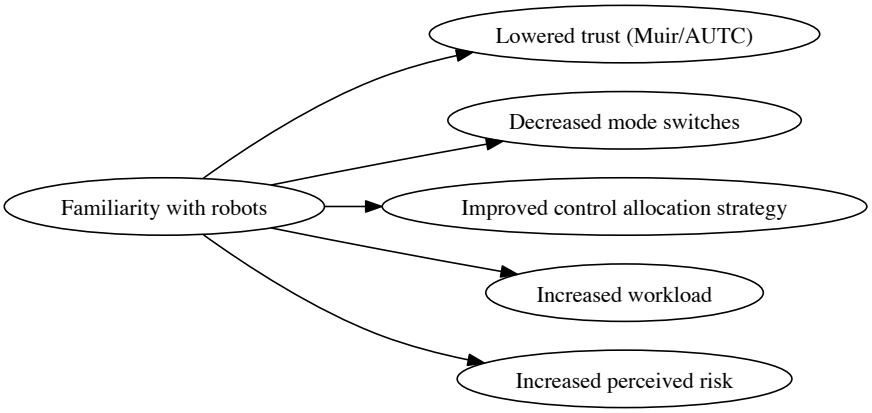

Figure 11.8: The impact of familiarity with robots on different factors. All of the effects shown are based on significant differences between the two participant groups in the Long-term (LT) experiment.

Interestingly, no significant differences were found between sessions two through six for any attribute. This lack of a difference between sessions and the significant similarities found between sessions indicates that an operator's behavior during initial interaction can predict his or her behavior over the short term.

With respect to the impact of familiarity with robots, several significant differences were found. Figure 11.8 shows the impact familiarity with robots has on operator behavior. It shows that while there was not a difference in performance, participants who were familiar with robots trusted them less and had an increased workload, perhaps due to feeling the need to execute better robot supervision, in accordance with the HARRT model. The better robot supervision in turn positively affected their control allocation strategy and also is consistent with the HARRT model. Figure 11.9 shows the familiarity model incorporated into the HARRT model and guidelines based on the reduced long-term and familiarity model are described below:

Guideline 9: Initial operator behavior does not change over the short term. It is possible to quickly assess and predict an operator's behavior over a longer period of time, based on their initial interactions with the robot.

Guideline 10: Familiarity with robots does not impact performance. Familiarity with robots should not be interpreted as or confused with expertise in remote robot teleoperation. While familiarity with robots impacts trust, it does not impact performance.

11.4.2.5 Impact of Timing of Periods of Low Reliability

Periods of low reliability early in the interaction not only have a more immediate detrimental impact on trust, but that effect lasts throughout the interaction as it also impedes the recovery of trust. Since the experimental setup was designed to require participants to rely more on the fully autonomous mode, the impact of decreased trust on other parameters was not as noticeable. However, for most balanced operations, the impact on trust would also be accompanied by a similar impact on control allocation, performance, and workload. Guidelines based on the impact of periods of low reliability early in the interaction are described below:

Guideline 11: Operator's initial interactions must always be stable. The implications of the timing data are that initial segments of every interaction must be stable and reliable. If needed, this experience should be facilitated by conducting a short, controlled interaction.

Guideline 12: In the event of a reliability drop early in the interaction, corrective measures must be taken. These steps (e.g., providing information explaining the cause for the reduction in reliability) must essentially minimize or prevent erratic operator behavior due to confusion or other factors. There are costs associated with these preventive steps, along with other implications associated with different measures, so caution must be exercised while selecting corrective measures.

11.4.2.6 Impact of Age

As people grow older, their attitude towards risk changes: they are willing to take fewer risks (e.g., Mather et al. 2009). We also found a significant correlation with age to answers to several questions about risk asked of participants. This unwillingness to take on more risk is shown in robot use through the fact that they prefer some autonomy modes and do not switch out of their comfort zone as often. Attitudes towards risk change with age, but so does the view or the definition of risk. It was often mentioned by the older participants that the compensation did not matter to them as much. However, it must also be said that they were still motivated to perform well. The inertia in control allocation exhibited by the older participants could potentially also have increased their workload and ultimately performance. Guidelines based on the impact of age are described below:

Guideline 13: Know your target audience. It is important to take into account the different population groups that will be interacting with a robot. Understanding the motivations of the operators can help explain their view on potential risks and better predict their behavior.

Guideline 14: Accommodate operators of all ages. Due to a higher probability of poor control allocation and poor performance for older operators, more time should be spent training them. To counteract the inertia observed, additional steps can also be taken. However, caution must be exercised to ensure that these steps do not increase their workload. For the other end of the age spectrum, given their tendency to take more risk, the risks involved in the scenario must be explained carefully. Since the younger population has the ability to better

manage workload and better robot management, it should be easier to influence their control allocation strategy if needed.

11.4.3 Modeling Trust

Using the experimental methodology, multiple experiments were conducted to examine the impact of different factors on operator trust and control allocation. These factors were selected based on different criteria. Some factors were selected based on the results of the initial surveys (i.e., reliability and risk). In fact, to better model real world scenarios, we ensured that dynamic reliability and risk were inherent in all of the experiments. Other factors like situation awareness (SA) and reduced task difficulty (RD) were selected based on their significance to the remote robot teleoperation task and also on our observations of other experiments involving remote robot teleoperation. Factors like feedback and long-term interaction were selected based on conjectures and commonly held beliefs. For example, it is often assumed that providing feedback to the operator should increase their trust of the robot and improve performance.

The results from these experiments showed interesting, sometimes unexpected, but overall insightful data. Using that data we were able to find different attributes that are relevant to human interaction with remote autonomous robot and the mediating relationships between them.

These results were used to create the Human interaction with Autonomous Remote Robots for Teleoperation (HARRT) model, a regression based model. Based on the HARRT model and the specific experiments, guidelines were proposed that should help improve overall performance by better managing the different tradeoffs (e.g., workloads, situation awareness, feedback) to influence operators' control allocation strategy. These results also highlight some of the differences between HAI and HRI. For example, a primary difference between HAI and HRI was the lack of direct correlation between trust and control allocation, a result always observed in HAI research.

11.5 Conclusions and Future Work

Our ultimate goal is to build models of the factors that influence people's trust in automated systems, across many domains, building a common core model of trust for automated systems and identifying factors specific to particular domains. Such models will serve to inform the designers of automated systems, allowing the development of systems that address the key factors for developing and maintaining a person's trust of an automated system. This chapter presents some of our initial work towards this goal, identifying the factors that most influence people's trust of automated cars, medical diagnosis systems, and remotely operated robot systems.
We have presented two methods for developing different types of models of trust in automation. The HARRT model is a regression model that used data from extensive experimentation. However, given the limitations of regression modeling, the HARRT model is not able to be used dynamically to predict current levels of user trust during the use of the robot. It is best used as a predictive model of how trust will evolve. However, it could be modified into a trust model that could be run dynamically, which would provide the robot with the means to determine the moment-by-moment modifications in its behavior that are necessary to elicit appropriate trust levels from the user. In contrast, the survey-based method results in a list of factors that are important to generating trust. Because the survey-based model is not designed for real-time execution, it is best used prior to system design to generate requirements that are trust-related.

The choice of the modeling approach for any given situation can thus be based on the time and resources available, and whether real-time adjustments in the robot's behavior are desired so as to elicit the optimal level of trust at any given point. The survey-based modeling method is more appropriately used when a large number of potential respondents are available; for example, the user pool consists of a large segment of the general population. If the system is to be used by specialized populations (e.g., doctors, first responders), it is best to conduct studies within those user groups.

The characteristics of each modeling approach are summarized in Table 11.5.

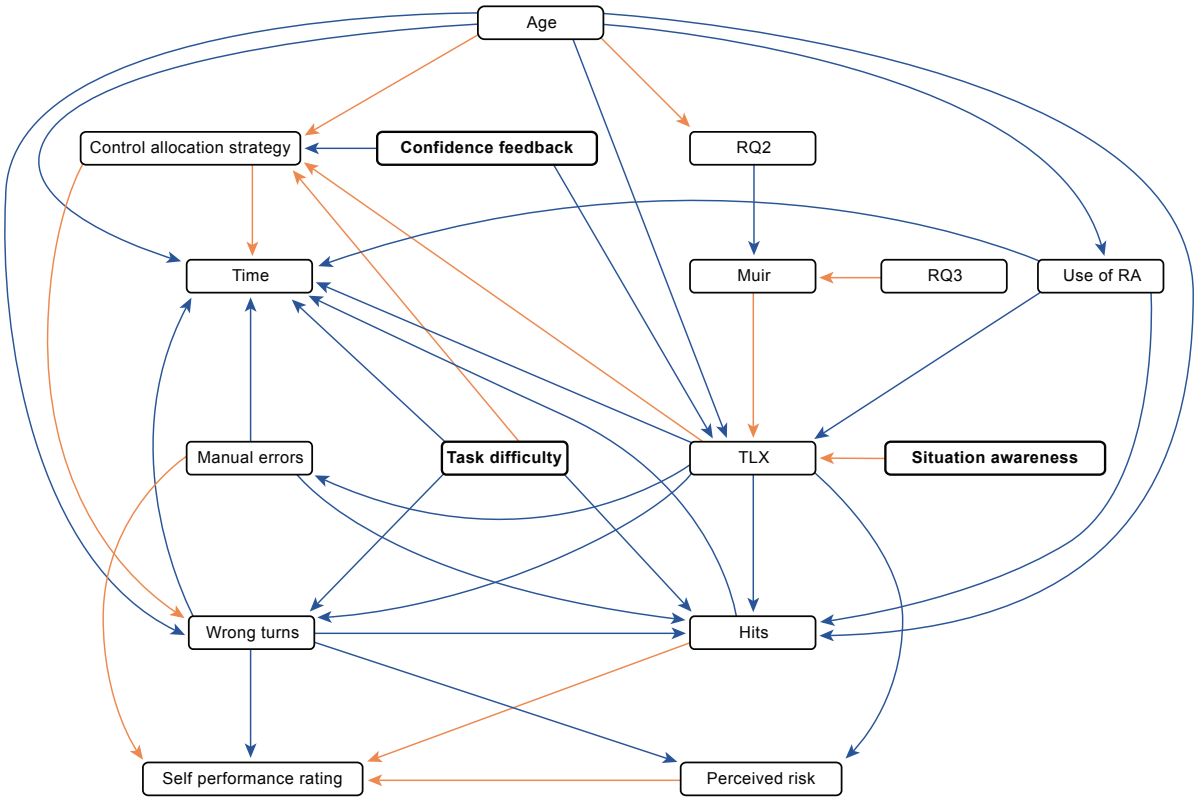

Figure 11.9: The original human and autonomous remote robot teleoperation (HARRT) model augmented with all of models described in Section 11.4.2. The orange and blue arrows indicate an inverse relationship or a proportional relationship, respectively. RQ2 is the second question about risk from Grasmick et al. (1993): "Sometimes I will take a risk just for the fun of it." RQ4 is the fourth question from Grasmick et al. (1993): "Excitement and adventure are more important to me than security." Participants were asked to answer these questions using a 6-point Likert scale.

Characteristic	**HARRT**	**Survey-Based Model**
Time and cost to create	High	Low
Method to create	Experimentation with system	Surveys
Format	Directed graph	List of factors
Real-time execution possible?	Yes, with modifications	No
Potential use	Modify robot's behavior in real time to elicit appropriate trust	Requirements generation

Table 11.5. Characteristics of HARRT and Survey-Based Modeling Approaches

We are planning to create an example of an executable HARRT model as part of our future work. This work will involve developing alternative behaviors when pre-identified conditions occur. For example, if the robot determines that a sensor is becoming unreliable, this condition may trigger the robot to provide an error message and modify the user interface layout to make an alternative sensor's readouts more salient. This will be a

successful strategy if the user trusts the unreliable sensor less and the alternative sensor more, as compared to earlier trust levels.

Additional future work will use a survey-based approach to examine the effects of mediated versus unmediated autonomy on the relative importance of the trust factors. By mediated, we mean that there is a human expert user who interprets the robot's results or who actually operates the robot on behalf of the end user (that is, the user who is benefiting from the robot's work). The use of IBM's Watson by an oncologist represents a mediated use from the standpoint of the cancer patient. We hypothesize that the trust factors will be rated differently from users in mediated versus unmediated situations.

11.6. Acknowledgements

This research has been supported in part by the National Science Foundation (IIS-0905228 and IIS-0905148) at the University of Massachusetts Lowell and Carnegie Mellon University, respectively, and by The MITRE Corporation Innovation Program (Project 51MSR661-CA; Approved for Public Release; Distribution Unlimited; 13-3767).

Munjal Desai conducted the research described in this chapter while a doctoral student at the University of Massachusetts Lowell. Michelle Carlson of The MITRE Corporation assisted with the design and analysis of the survey-based research. Hyangshim Kwak and Kenneth Voet from the United States Military Academy assisted with the survey-based work during their internships at The MITRE Corporation. Many people in the Robotics Laboratory at the University of Massachusetts Lowell have assisted with the robot testing over several years, including Jordan Allspaw, Daniel Brooks, Sean McSheehy, Mikhail Medvedev, and Katherine Tsui. At Carnegie Mellon University, robot testing was conducted with assistance from Christian Bruggeman, Sofia Gadea-Omelchenko, Poornima Kaniarasu, and Marynel Vázquez.

All product names, trademarks, and registered trademarks are the property of their respective holders.

11.7 References

Baker, M. and Yanco, H. (2004), Autonomy mode suggestions for improving human-robot interaction, IEEE International Conference on Systems, Man and Cybernetics, 2948-2953.

Bliss, J. P. and Acton, S. A. (2003), Alarm mistrust in automobiles: how collision alarm reliability affects driving, Applied Ergonomics, 34(6), 499-509.

Boehm-Davis, D. A., Curry, R. E., Wiener, E. L., and Harrison, L. (1983), Human factors of flight-deck automation: report on a NASA-industry workshop, *Ergonomics*, 26(10), 953-961.

Bruemmer, D. J., Dudenhoeffer, D. D., and Marble, J. L. (2002), Dynamic autonomy for urban search and rescue, AAAI Mobile Robot Workshop.

Burke, J. L., Murphy, R. R., Rogers, E., Lumelsky, V. L., and Scholtz, J. (2004), Final report for the DARPA/NSF interdisciplinary study on human-robot interaction, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 34(2), 103-112.

CasePick Systems (2011), retrieved 12/30/11 from http://www.casepick.com/company.

Chen, J. Y. (2009), Concurrent performance of military tasks and robotics tasks: effects of automation unreliability and individual differences. 4th Annual ACM/IEEE International Conference on Human-Robot Interaction, 181-188.

Clark, M. (2013, 7/29), "States Take the Wheel on Driverless cars," *USA Today*, from http://www.usatoday.com/story/news/nation/2013/07/29/states-driverless-cars/2595613/.

Cohen, M. S., Parasuraman, R., and Freeman, J. T. (1998), Trust in decision aids: a model and its training implications, Command and Control Research and Technology Symposium.

Dellaert, F., and Thorpe, C. (1998), Robust car tracking using Kalman filtering and Bayesian templates, Intelligent Transportation Systems Conference, 72-83.

Desai, M. (2012), Modeling trust to improve human-robot interaction, Ph.D thesis, University of Massachusetts Lowell, November.

Desai, M., Medvedev, M., Vazquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., and Yanco, H. (2012), Effects of changing reliability on trust of robot systems., 7th Annual ACM/IEEE International Conference on Human-Robot Interaction.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013), Impact of robot failures and feedback on real-time trust, 8th Annual ACM/IEEE International Conference on Human-Robot Interaction.

Desai, M. and Yanco, H. (2005), Blending human and robot inputs for sliding scale autonomy, IEEE International Workshop on Robots and Human Interactive Communication, 537-542.

deVries, P., Midden, C., and Bouwhuis, D. (2003), 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.

Dixon, S. R. and Wickens, C. (2006), Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload, Human Factors, 48(3), 474-486.

Dudek, G., Jenkin, M., Milios, E., and Wilkes, D. (1993), A taxonomy for swarm robots, IEEE/RSJ International Conference on Intelligent Robots and Systems, 441-447.

Dzindolet, M., Pierce, L., Beck, H., Dawe, L., and Anderson, B. (2001), Predicting misuse and disuse of combat identification systems, Military Psychology, 13(3), 147-164.

Dzindolet, M., Pierce, L., Beck, H., and Dawe, L. (2002). The perceived utility of human and automated aids in a visual detection task, Human Factors: The Journal of the Human Factors and Ergonomics Society, 44(1), 79-94.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003), The role of trust in automation reliance, International Journal of Human Computer Studies, 58(6), 697-718.

Endsley, M., and Kiris, E. (1995). The out-of-the-loop performance problem and level of control in automation, Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(2), 381-394.

Farrell, S. and Lewandowsky, S. (2000), A connectionist model of complacency and adaptive recovery under automation, Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(2), 395-410.

Gerkey, B., Vaughan, R., and Howard, A. (2003), The Player/Stage project: tools for multi-robot and distributed sensor systems, 11th International Conference on Advanced Robotics, 317-323.

Grasmick, H., Tittle, C., Bursick, Jr, R., and Arneklev, B. (1993), Testing the core empirical implications of Gottfredson and Hirschi's General Theory of Crime, Journal of Research in Crime and Delinquency, 30(1), 5-29.

Guizzo, E. (2011, 10/18), "How Google's Self-driving Car Works," IEEE Spectrum, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works.

S. G. Hart and L. E. Staveland (1988), Development of NASA-TLX (Task Load Index): results of empirical and theoretical research, Human Mental Workload, 1(3), 139-183.

International Federation of Robotics (IFR) (2011), Statistics about service robots, retrieved 12/30/11 from http://www.ifr.org/service-robots/statistics/.

iRobot (2011), iRobot Roomba, retrieved 12/30/11 from http://www.irobot.com/roomba.

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

Kiva Systems (2011), Kiva Systems, retrieved 12/30/11 from http://www.kivasystems.com/.

Lee, J. (1992). Trust, self-confidence and operator's adaptation to automation, PhD thesis, University of Illinois at Urbana-Champaign.

Lee, J. and Moray, N. (1991), Trust, self-confidence and supervisory control in a process control simulation, IEEE International Conference on Systems, Man, and Cybernetics, 291-295.

Lee, J. D. and Moray, N. (1992), Trust, control strategies and allocation of function in human-machine systems, Ergonomics, 31(10), 1243-1270.

Lin, P. (2008), Autonomous military robotics: risk, ethics, and design, Technical Report, Defense Technical Information Center.

Madhani, K., Khasawneh, M., Kaewkuekool, S., Gramopadhye, A., and Melloy, B. (2002), Measurement of human trust in a hybrid inspection for varying error patterns, Human Factors and Ergonomics Society Annual Meeting, 46, 418-422.

Mather, M., Gorlick, M. A., and Lighthall, N. R. (2009), To brake or accelerate when the light turns yellow? Stress reduces older adults' risk taking in a driving game, Psychological Science, 20(2), 174-176.

Michaud. F., Boissy, P., Corriveau, H., Grant, A., Lauria, M., Labonte, D., Cloutier, R., Roux, M., Royer, M., and Iannuzzi, D. (2007), Telepresence robot for home care assistance. AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics, March.

Moray, N. and Inagaki, T. (1999), Laboratory studies of trust between humans and machines in automated systems, Transactions of the Institute of Measurement and Control, 21(4-5), 203-211.

Moray, N., Inagaki, T., and Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks, Journal of Experimental Psychology: Applied, 6(1), 44-58.

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

Muir, B. M. (1989), Operators' trust in and use of automatic controllers in a supervisory process control task, PhD thesis, University of Toronto.

Neato Robotics (2011), Neato XV-11, retrieved 12/30/11 from http://www. neatorobotics.com/.

Ostwald P., and Hershey, W. (2007), Helping Global Hawk fly with the rest of us, Integrated Communications, Navigation, and Surveillance Conference, April/May.

Parasuraman, R. (1986), Vigilance, monitoring, and search. Handbook of perception and human performance: cognitive processes and performance. Boff, K.; Thomas, J.; and Kaufman, L. John Wiley & Sons.

Parasuraman, R., and Riley, V. (1997), Humans and automation: use, misuse, disuse, abuse, Human Factors, 39(2), 230-253.

Prinzel III, L. J. (2002), The relationship of self-efficacy and complacency in pilot-automation interaction, Technical report, Langley Research Center.

Riley, V. (1994), Human use of automation, PhD thesis, University of Minnesota.

Riley, V. (1996). Operator reliance on automation: theory and data. Automation and human performance: theory and applications. Parasuraman, R. and Mouloua, M. Lawrence Erlbaum Associates.

Ross, J., Irani, I., Silberman, M., Zaldivar, A., and Tomlinson, B. (2010), Who are the crowdworkers?: shifting demographics in Amazon Mechanical Turk, ACM CHI Conference on Human Factors in Computing Systems Extended Abstract, 2863-2872.

Sanchez, J. (2006), Factors that affect trust and reliance on an automated aid, PhD thesis, Georgia Institute of Technology.

Sarter, N., Woods, D., and Billings, C. (1997), Automation surprises, Handbook of Human Factors and Ergonomics, 2: 1926-1943, Wiley.

Sheridan, T. B. and Verplank, W. L. (1978), Human and computer control of undersea teleoperators, Technical report, Department of Mechanical Engineering, Massachusetts Institute of Technology.

Strickland, G. E. (2013), Watson goes to med school, IEEE Spectrum, 50(1): 42-45, January.

Thrun, S. (2010, 10/9), What we're driving at, from http://googleblog.blogspot.com/2010/10/what-were-driving-at.html.

Tsui, K., Norton, A., Brooks, D., Yanco, H., and Kontak, D. (2011), Designing telepresence robot systems for use by people with special needs, International Symposium on Quality of Life Technologies.

Yanco, H. A. and Drury, J. (2004), Classifying human-robot interaction: an updated taxonomy, IEEE International Conference on Systems, Man and Cybernetics, 2841-2846.