EFFECTS OF PROACTIVE EXPLANATIONS BY AUTONOMOUS SYSTEMS ON HUMAN-ROBOT TRUST

by

Lixiao Zhu

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Computer Science).

Golden, Colorado

Date _____

Signed: _____
Lixiao Zhu

Signed: _____
Dr. Thomas Williams
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Tracy Camp
Department Head and Professor
Department of Computer Science

ABSTRACT

Human-Robot Interaction (HRI) seeks understanding, designing, and evaluating of robots for human-robot teams. Previous research has indicated that the performance of human-robot teams depends on human-robot trust, which in turn depends on appropriate robot-to-human transparency. In this thesis, we explore one strategy for improving robot transparency, proactive explanations, and its effect on the human-robot trust. We also introduce a resource management testbed, in which human participants engage in a resource management exercise while a robot teammate performs an assistive task. Our results suggest that there is a positive relationship between providing proactive explanations and human-robot trust.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Human-Robot Interaction . . . . . HRI

No Explanations . . . . . NE

Proactive Explanations . . . . . PE

Situation Awareness . . . . . SA

Urban Search and Rescue . . . . . USAR

CHAPTER 1

INTRODUCTION

1.1 Human-Robot Trust

Human-Robot Interaction focuses in part on understanding, designing, and evaluating robots for human-robot collaboration tasks. Previous research has demonstrated that human-robot team performance depends on human-robot trust. Billings et al. introduce that the definition of human-robot trust refer to "exprectations, confidence, risk or uncertainty, reliance, and vulnerability" [1]. Previous studies suggest that human-robot trust in robots can be influenced by the task performance of a robot [2]. Human-robot trust may encourage or discourage people to use automation [3] and operators' trust is positively related to the use of automation. Human teammates must accept and trust their robot partners to have successful interactions [4]. Human-robot teams tend to have undesirable performance when human-robot trust is either too low or too high, which means that human-robot trust must be maintained at an appropriate level rather than directly maximized [5]. One key factor in establishing human-robot trust is the transparency of robots' internal beliefs, desires, and intentions [6].

1.2 Transparency and Trust

One way to achieve an appropriate human-robot trust is by designing robots that exhibit an appropriate level of transparency [7, 8]. It will help a human user to believe in a robot's capability and its ability to perform adequately for a certain type of task if the design of the robot is clearly explained to the user. Earlier findings suggest that improving transparency may minimize misunderstanding and confusion in a human-human interaction by considering the methods, such as verbal input, body orientation, facial expression, and eye behavior [9]. The same theory may also apply to human-robot interaction: two principal methods to improve the transparency of an autonomous system for a human participant are through

non-verbal or verbal communications. Wortham et al. [10] show that a simple real-time visualization of a robotic system can help a human to understand the design of the robot. Researchers have also developed a human-like verbal communication system for robots to improve transparency for human-robot interaction [11].

Non-verbal communication is able to positively increase human understandability of robots to improve human-robot task performance [12]. Non-verbal communication within a human-robot team can be divided into two types, which are explicit and implicit [12]. For explicit communication, an autonomous system deliberately informs the receiver about its goal of sharing information. For example, a robot may move around and point to a particular object to notify humans. Similarly, explicit non-verbal communication can also direct a human's attention during human-robot conversation. An example of implicit communication, on the other hand, is that a human can implicitly recognize a robot's current interests by monitoring the robot's gaze. Breazeal et al. [12] suggests that the transparency of a human-robot interaction can be improved by giving implicit non-verbal communication. Explicit non-verbal communication can benefit a human-robot team because it informs humans of the reason for the robot's action. Implicit non-verbal communication can also improve the efficiency and robustness of the HRI.

Another method of improving transparency of human-robot teamwork is through human-like verbal communication. The next section provides more information about giving verbal communication within a human-robot team.

1.3 Explanation Generation

Non-verbal communication can positively affect transparency within a human-robot team. Ultimately, human-robot trust will be increased by a high level of transparency. This can be accomplished through dialogue strategies such as explanation generation.

Sridharan et al. [13] discuss that there are different ways to formulate explanation generation. Sridharan divides existing approaches of explanation generation into two classes. The first class is to use minimal heuristics with elaborate system descriptions and observations

of system behavior to generate explanations, such as Answer Set Programming (ASP) [14]. Another class is to depend more on heuristics to provide explanations, such as the method that is introduced by Meadows et al. [15].

In this thesis, we specifically consider differences between two types of explanations: proactive explanations and reactive explanations. Reactive explanations are explanations generated in reaction to a request for explanation from a a human teammate, to help users understand past robot actions. Previous studies on robot explanation generation have mainly focused on these sorts of explanations that are provided by robots on demand. In contrast, little work has been performed on proactive explanations: explanations that are generated before an action is performed. This provides information to human users so that they can understand the actions taken by these automation systems. One challenge of generating proactive explanations is to avoid communicating too much information, which may overload users [16]. Hence, explanations generated by a robot will need to provide accurate information for users without unnecessary details about the robot's actions.

We further divide proactive explanations into two categories that differ according to the level of information provided.

- **'What' Proactive Explanation**: The first type of proactive explanation is to give explanations that only inform the user of *what* the robot is going to do. For example, a navigation robot can inform the user about its next turn by stating *that* it is going to make a turn.

- **'Why' Proactive Explanation**: The second type of proactive explanation will notify the user of both what the robot is going to do, and why the robot is going to do it. For example, a navigation robot will tell the user about its next turn and *why* it is making that decision (traffic issue, shorter path, etc).

The ultimate goal of our study is to understand the effects of proactive explanation on human-robot trust. In addition, we wish to understand how different types of proactive

explanations influence human-robot trust separately. Another important question we would like to answer is whether human monitoring of a robot is related to human-robot trust.

1.4 Contents and Contributions of This Thesis

This remainder of the thesis is organized as follows:

Chapter 2 presents further related work that informs our approach. The background section discusses the categories of human-robot interaction and relevant works in human-robot trust.

Chapter 3 introduces the resource management testbed we designed and used for this human-subject study. Chapter 4 presents the design and results of a human-subject study conducted to answer our research questions.

Chapter 5 and Chapter 6 give the discussion and conclusion from the results of this study. Our study suggested that there is a positive relationship between providing proactive explanations by robots and an increase of human-robot trust. Finally, we discuss directions for future work.

CHAPTER 2

BACKGROUND

2.1 Human-Robot Interaction

Human-robot interactions can be categorized into two main types [17]. The first type is remote interaction, in which the locations of the human and the robot are not the same. For example, researchers from Arizona State University simulate a remote human-robot interaction for an urban search and rescue (USAR) task [18], in which a participant interacted with a simulated Nao robot without being in the same location. As another example, a mobile robot was designed and implemented as a teleoperated tool for remotely accessing objects on requests by a remotely located operator [19]. The second type of interaction described by Goodrich is proximate interaction in which the human and the robot stay in the same location during the interaction. Most research in HRI focuses on proximate interaction. For example, Tran et al. [20] presented work on mixed-reality HRI in which a participant interacts with the Pepper robot while wearing a Microsoft HoloLens in the same room. In this work, we focus on proximate human-robot interaction.

Proximate HRI can be categorized into three approaches as follows: [21]:

1. **Robot-centered HRI**: This category of HRI focuses on the view of an autonomous system as the main entity that is acting based on its own decisions and motivations. Human-robot interaction serves to help the robot to meet its needs. For example, a robot may politely ask a person to open a door for the robot in order to deliver an important document. In this case, the human-robot interaction (a person opens a door for the robot) serves to achieve the robot's goal (to deliver a document to the destination). As another example, a group of researchers from Harvard investigated human-robot overtrust [22] by placing a robot outside a student residence hall and let the robot to request students for access. In this experiment, the robot is the main

entity to ask people (students) to open the door for it.

2. **Robot Cognition-centered HRI**: This category of HRI emphasizes that a robot should be treated as an intelligent system that solves problems by making it own decisions for a specific application domain [21]. Robot cognition-centered researchers focus on designing robots with an appropriate level of cognitive capacity [23]. For example, a study recently introduced a self-driving robot that could properly climb a mountain path with a deep convolutional neural network [24].

3. **Human-centered HRI**: This view of HRI mainly focuses on developing and designing an autonomous system that can work with a person or a group of people in a manner that is pleasant to human teammates. Researches in this area study the relationship between human perceptions of robots and their actions [21]. Previous studies introduce the importance of designing a robot that is safe and friendly for humans [25]. Since robots are more and more commonly used in public workspaces, such as hospitals [26, 27], restaurants [28, 29], and colleges or schools [30], there are extraordinary demands of social robots to interact with human and make them feel comfortable to work with the robots. From this point of view, the design of robots should pay attention to robots' usability, acceptability, and believability [23]. Our work focuses on human-centered HRI in order to develop robots that can work with people more efficiently and effectively.

2.2 The Use of Social Robots

Much of human-centered HRI comes from the area of Social Robotics. Previous work has defined social robots as autonomous systems developed to interact with human beings socially or to provide social responses in a natural way [31]. In other words, a service robot is designed to support a human-like interaction [32]. There are many applications of using a social robot as a social agent, who interacts with humans in a human-like way. For example, Satake et al. [33] proposed a model which allows an autonomous system to initiate

conversations with consumers who are walking in a shipping mall. The robot they used in the study, Robovie, is designed to give human-like expressions. One of the key findings from the work is that robots need to guess people's behavior, such as willingness to interact with the robot in a public environment. Hence, the design of a social robot has to consider not only the robot's performance and ability but also its social or communication skills with humans. Social robots have become more and more popular the everyday environments. While robots were previously limited to industries or labs for research purposes, service robots or social robots are now being used in many fields to enhance the quality of human lives. The number of service robots exceeds the number of industrial ones in 2008 according to the World Robotics Survey [34], which was provided by the International Federation of Robotics (IFR).

Social robots are being used in education. Mubin et al. [35] present that educational robotic systems are developed to provide social supportive behaviors in the learning activities, such as language, science, and technology educations. In this work, robots are categorized into three different roles with an associated case study. The three roles of robots in educations are a tutor, a peer, and a tool. Saerbeck et al. implemented a tutor robot to help students with a language learning task [36]. The results in the study suggest that the robot tutor iCat, which is developed by Philips Research, improved the young students' learning performance. Han and Kim [37] propose that using robots as a teaching peer by praising students can effectively motivate students' willingness to learn.

Robots have also been used specifically for science learning activities. Students interacted with a robot playing an game to learn arithmetic [38]. The robot tutor dynamically adapts the complexity of the assignments based on the learning outputs of the student. To use a robot as a peer, an android robot SAYA and a human-like robot RoboThespian collaborate with two sixth grade classes to work on science assignments [39]. Robots are also treated as tools to teach students about science-related topics. Church et al. [40] described a robotic system to enable high school students to learn physics. In addition, Mubin et al. [35] gave

instances of using robotic systems in technology education. Preceding work describes a Beebot programmable toy as a tool to help young kids to develop excellent problem-solving skills for the advanced technology [41]. Robots can play an important role in education to motivate students, eventually improving students' learning performance.

Robotics have also been used in the context of urban search and rescue (USAR). The purpose of USAR is to locate stranded victims and provide initial medical support to those trapped under a structural collapse. Robots designed for USAR can enter disaster environments, which are either too dangerous for rescue personnel or too small for a human to enter [42]. The most famous example of using robots for an urban search and rescue mission is the 2001 World Trade Center (WTC) collapse [43]. From September 11 until October 2 2001, mobile robots were used to provide supports for the disaster. There are four main targets for rescue robots. The first is to locate the victims and report their locations back to the rescue center. The second is to find optimal paths in the building for a more expeditious excavation. The third is to inspect the collapse. Finally, the robots work to detect any harmful or poisonous gases or materials by sensors.

Intelligent hospital service robotic system becomes another innovative area for interactive robots. The goal of assistant robots in hospitals is to save human resources to improve hospital efficiency [44]. Takahashi et al. [26, 45] implemented a transport robot in a hospital in Hong Kong. The omnidirectional mobile robot has a truck to transfer important specimens and other supplies. Another case study of a transport robot in hospitals is introduced by Ljungblad et al. [46] who use a mobile robot to deliver goods and blood samples in a semi-public hospital environment.

Although there are numerous applications of social robots, the common design of robotic systems is to interact with human beings to finish human-robot tasks. In other words, a service robot is expected to perform and act in a human-like fashion to interact with a real person. Service robotic systems are designed to understand human requests, provide suitable solutions to meet human needs, and improve the work efficiency of a human-robot

team. Therefore, the study of human-robot interaction becomes crucial for the design of modern robots.

2.3 Common Metrics in Human-Robot Interaction

Steinfeld et al. [47] introduce common metrics for evaluation across a large-scale range of human-robot tasks and studies, including:

2.3.1 System Performance

System performance focuses on how to evaluate a robot-human team performance instead of a task-specific performance. The metrics of system performance contains three subareas. The first type is the quantitative performance, which measures the effectiveness and efficiency of the human-robot team for a human-robot task. The definition of the effectiveness of a human-robot team is the percentage of the human-robot task that was finished with the designed autonomous systems. For example, a navigation robot is designed to guide a person from place A to place B. The robot fails to finish the task by only guiding the person to arrive at the destination. The human teammate has to intervene 20% of the task time to help the robot to finish the task. Therefore the system only has 80% effectiveness. This metric can be calculated by using the number and duration of human interventions for a human-robot task. Another important metric for quantitative performance is efficiency. The amount of time spent to accomplish a task represents the efficiency of a robot. The number of completed tasks can also be used to compute the efficiency of a human-robot team.

The second sub-metric of system performance is subjective ratings. Subjective ratings are used in addition to quantitative measures of the performance of a human-robot team. The same navigation robot can be used as a great example to explain this concept. In addition to the efficiency and effectiveness of guiding a human teammate to a destination, the quality of the selected path or route is also valuable for the system performance.

The last sub-metric is the appropriate utilization of mixed-initiative, which is one of the main challenges in task-oriented HRI. A robot is required to maintain more self-awareness

and awareness of their human teammate. The recommended measures are the percentage of commands requested by a robot or a human and the number of unnecessary interruptions of the human teammate. Consequently, the ability to self-regulate who possesses initiative is crucial for a human-robot team.

2.3.2 Operator Performance

The second common metric is the evaluation of an operator's performance. Operator performance has three factors: situation awareness (SA), workload, and accuracy of mental models of device operation. Operator situation awareness is categorized into three levels [48]. The first level of situation awareness is a perception of items in the surrounding environment. The second level of SA is the comprehension of the noticed items. The third level is to project the future of the items. Endsley [49] proposes a well-known measurement tool, called the Situation Awareness Global Assessment Technique (SAGAT), for evaluating operator SA. Other than operator SA, operator workload is another significant measurement for a human-robot team. The workload can be assessed by relating operator cognitive load to their SA. Hart and Staveland [50] introduce a multi-dimensional rating scale to estimate the human workload. The assessment technique is known as the NASA-Task Load Index (NASA-TLX).

2.3.3 Robot Performance

Three aspects can be considered to evaluate the performance of a robot for a human-robot task. The first measurement is to obtain a robot's self-awareness, which is the ability of a robot to evaluate itself. Self-awareness plays a very vital role to influence the efficiency of human-robot interaction. When a robot has less awareness of its capabilities to complete a certain task and fails to report its potential troubles to its human teammate, the human has to take more time to monitor the robot's condition to prevent system failures. The following characteristics of a service robot are proposed to increase a robot's self-awareness. First, the robot should understand its intrinsic limitations, which include the limitation of sensors connected to the robot. Second, the robot should be able to self-monitoring and

identify its system failures. Finally, the robot with high self-awareness should be responsible for recovering from errors.

Human awareness is another relevant aspect of the robot's performance. Drury et al. [51] present a HRI awareness framework to define the need for a robot's human awareness. From this framework, autonomous systems have to acknowledge humans' commands for direct interaction between robots and a group of individuals. The HRI awareness framework also proposes a metric to assess human awareness by counting the number of awareness violation, which is when the robot fails to provide needed information.

The performance evaluation of an autonomous system should also examine the robot's autonomy, which is the ability of a service robot to operate without helps from human operators. The common metric for measuring autonomy is by applying the notion of neglect tolerance [52]. Neglect tolerance is a measure of how a robot's effectiveness drops below a desirable level of performance when the operator is neglecting the robot.

As Steinfeld et al. [47] mentioned in the performance evaluation of a robot, a robot's self-awareness can affect human interventions and monitoring of the robot. A human operator may spend more time on checking the robot's current status and conditions if a robot has low self-awareness to report its errors to the human operator. In other words, the robots can deliberately provide explanations to the operator about its status to decrease the human monitoring of the robot. In this thesis, we present a human-subject study to understand the relationship between human monitoring and proactive explanations. Based on our research, no previous study in HRI proposes that the human monitoring of a robot is related to explanations by robots. Therefore, we explore the relationship between human monitoring and proactive explanation to confirm whether human-robot trust is correlated to human monitoring of autonomous systems.

2.4 Basis of Human-Robot Trust

As robots are designed to work with people, human-robot trust plays an important role in a human-robot team to accomplish its goal. It is more necessary to consider the

effect of human-robot trust in robots during high-risk tasks. Robots can be used to enter dangerous areas that are too small or unsafe to humans to rescue others during an emergency situation [42]. In high-risk situations, people tend to have significantly less interest in working with poorly performing robots [53]. Meanwhile, Salem et al. [54] suggest that a robot's performance does not affect human decisions to accept or reject its requests. The nature of a robot's requests has a significant effect on human teammates' decision to obey its instructions. There are other factors can describe the basis of human-robot trust. These characteristics can be divided into three major classes [55]. The first type is the ability [56] and expertise [57] of a robot. Another type is the performance of a robot. Competence [58] and reliability [59] of a robot can also influence human-robot trust.

2.5 Human-Robot Trust Measurement

There are a large number of previous research studies focused on human-robot trust measurement [55, 60–65]. One of the most commonly practiced methods of measuring human-robot trust within a human-robot team is using self-report measures. Jian et al. [66] proposed a trust survey to assess human-robot trust in autonomous systems, which become more and more popular in our daily life. Jian et al. performed a three-phased human-subject experiment relating human-robot trust in other people or robotic systems. Another well-known trust scale proposed by Schaefer [60] is a 40 item trust scale, which was proved to be more accurate than Jian's work. The 40 item trust scale generally take between five to ten minutes to finish. A 14 item sub-scale of the 40 item scale is introduced by Schaefer. Schaefer compared the performance of these two types of scale with Jian's work. Both 14 item and 40 item trust scales successfully represent the human-robot trust change between pre- and post-interaction [60], while Jian's trust survey failed to measure the trust change.

CHAPTER 3

RESOURCE MANAGEMENT TESTBED

In order to assess our research questions, we require an experimental testbed that meets the following three requirements. The first requirement is that participants and robots must be co-located in the same room. This will enable a proximate human-robot interaction, in which the robot is physically embodied and can be treated as a real teammate. Previous research suggests that it is difficult to assess the social and cognitive aspects of human-robot interaction [17] when robots are not physically co-located with their human teammates. The second requirement is that the testbed must include a robot that the human teammate is depending on to perform a portion of their shared task. Moreover, these task-relevant actions must be observable and understandable by the robot's teammate, so that the teammate has the motivation to physically monitor the robot's actions in order to know what the robot is doing and how it contributes to the human-robot task. The third requirement is that the testbed must also require the human to perform a portion of the shared task in order to ensure task success. This requirement prohibits human teammates from monitoring the robot consistently, because they must spend the majority of their time on their portion of the shared task. In this chapter, we present an experimental testbed designed to satisfy these three requirements.

In order to study the effect of proactive explanations on human robot trust, we had participants engage in a resource management task with the help of a robot, in which participants spend different types of resources while exploring an environment. The robot behind the player was responsible for "collecting" these resources. Figure 3.1 displays the user interface design of the resource management testbed. For the resource management testbed, a participant partners with a turtlebot robot, which consistently collects resources for the human player. The player's goal is to consume the collected resources efficiently by pursuing

all found resources request blocks, which can be found on the board. The user can manually determine which type of resources is most needed and directly make a request to the robot to collect this resource type. However, the robot also autonomously decides to collect different types of resources on the user's behalf, depending on the resources that it believes are most needed. In this task context, proactive explanations can be used to explain these autonomous decisions to the human teammate.

As shown in Figure 3.2, the red circle represents the player's current location. The participant can move around by clicking on the yellow squares. They can also move to any visited cells in any direction by selecting the teal cells. Light gray cells indicate cells that are visible based on the current location with potential resources. Some dark gray cells are undiscovered and can be discovered as the player moves around the board.

Resources are dynamically distributed throughout the environment as the user explores, according to a pre-defined resource distribution policy. The general idea of the distribution algorithm is to allocate more blocks of a certain type of resource in the early portion of the task and to distribute fewer blocks of the same type of resource in the latter portion of the task. Table 3.1 presents the number of resource blocks in each color. Each cell holds two resource blocks. We expect a human player to request more red resource blocks at the beginning of the task. For the last 10 resource blocks, Table 3.1 shows that only one block is assigned to be in red and more blocks in pink are requested. Early in the task, more red blocks are needed, whereas later in the task, more pink blocks will be needed.

After the player moves from one cell to another, a simple math question will pop up for him or her to solve (designed to keep the user's attention on the resource management game, rather than the robot situated behind them). The participant will need to type the correct answer to the math problem to keep moving forward. Figure 3.1(a) gives an example of a math question during the task. Here, the different colored blocks represent different types of resources. If the user has enough resources to clear the blocks in an occupied cell, the resource request blocks on the square will turn black to indicate that resources have been

successfully spent to clear that block. An example of successfully cleared resource request blocks can be seen in Figure 3.1(b). To cancel out a whole cell, the player will need enough resources for both request blocks within the cell.

The resources needed to clear blocks on the user's screen are collected by the user's robot teammate. As shown in both Figure 3.2(a) and Figure 3.2(b), on the right side of the testbed interface are monitors to show the player the range of each resource being collected. These allow the user to determine for themselves which resources are in short supply, but do not allow them to explicitly determine which resource is currently being collected. In order to determine which resource is currently being collected, the user must physically turn their body around to observe the robot and determine which it is collecting (a process we describe later on when describing our human-subject experiment).

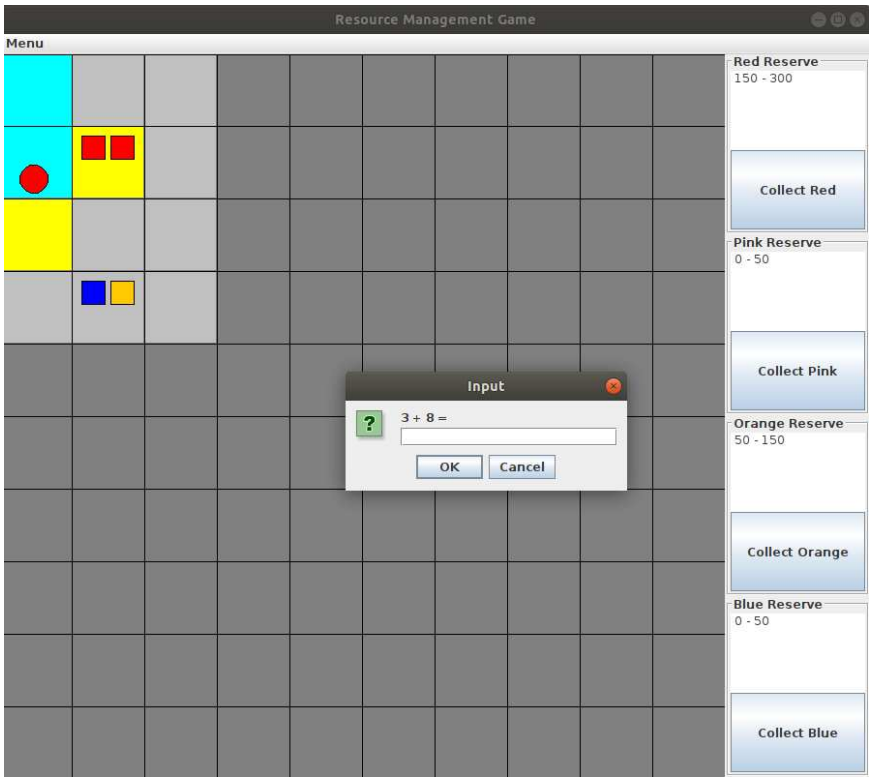
The robot determines which resource type to collect through two different processes First, the user may selecting a button located on the right side of the testbed interface to command the robot to collect a particular type of resource. Second, the robot will periodically decide autonomously to switch to collecting a different type of resource that it believes is more critically needed, as decided by a resource selection algorithm.

Table 3.1: Pre-defined Resource Request Blocks Distribution

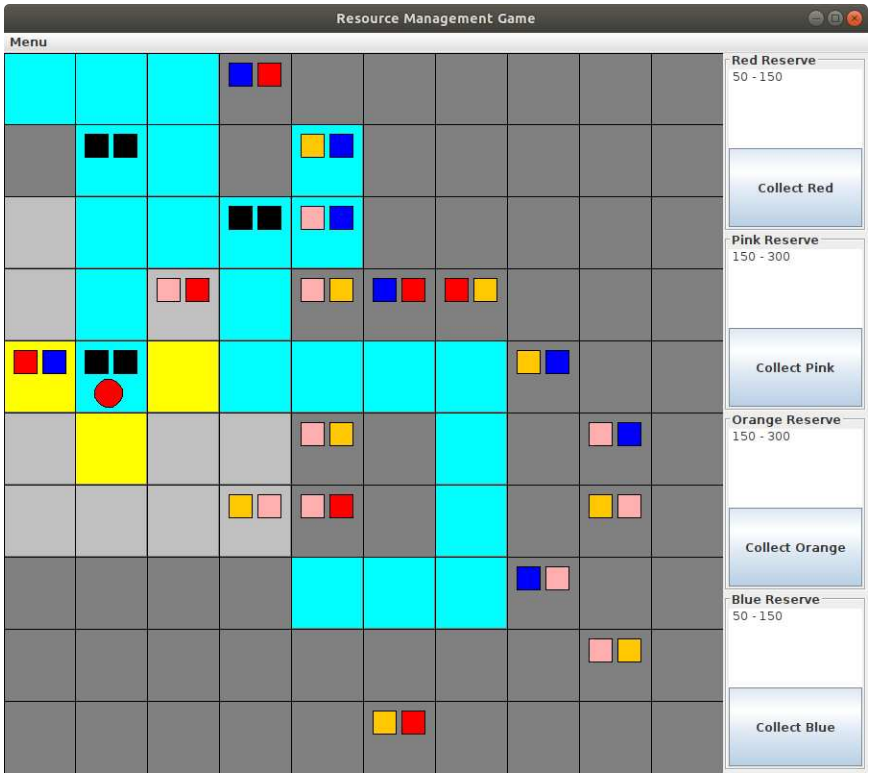
	Red	Pink	Orange	Blue
5 Cells	4	1	2	3
10 Cells	3	2	1	4
15 Cells	2	3	4	1
20 Cells	1	4	3	2

3.1 Resource Selection Algorithm

A resource selection algorithm is designed and implemented to help the robot to determine which type of resource is most needed with the least collected amount. The selection algorithm calculates ratios between the number of collected resources and the number of



(a) UI Design with Math Question



(b) Resource Request Blocks Consumption

Figure 3.1: UI Design of the Resource Management Testbed

resource blocks on the board for each type of resource.

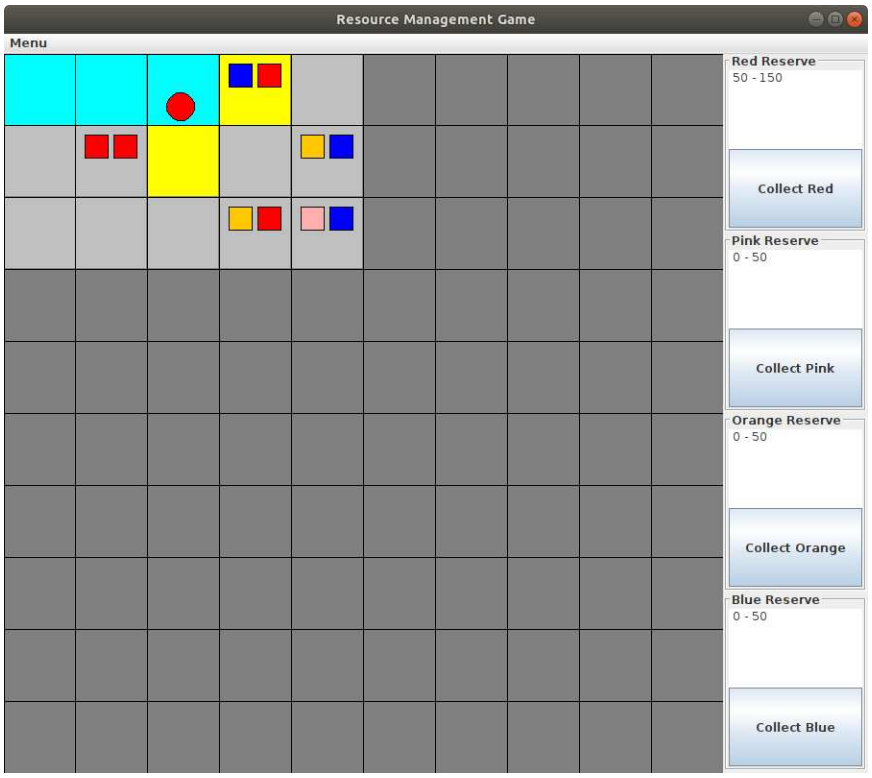
$$argmin_{color \in red,orange,pink,blue} \frac{collected_{color}}{needed_{color}} \tag{3.1}$$

The turtlebot will pick the resource type with the smallest ratio value (Equation 3.1), which implies that the type of resources is most needed with the least collected amount. The collection algorithm allows the turtlebot robot to dynamically make the optimal decision during the resource management task. More information about the implementation can be found under Section 3.2.4.

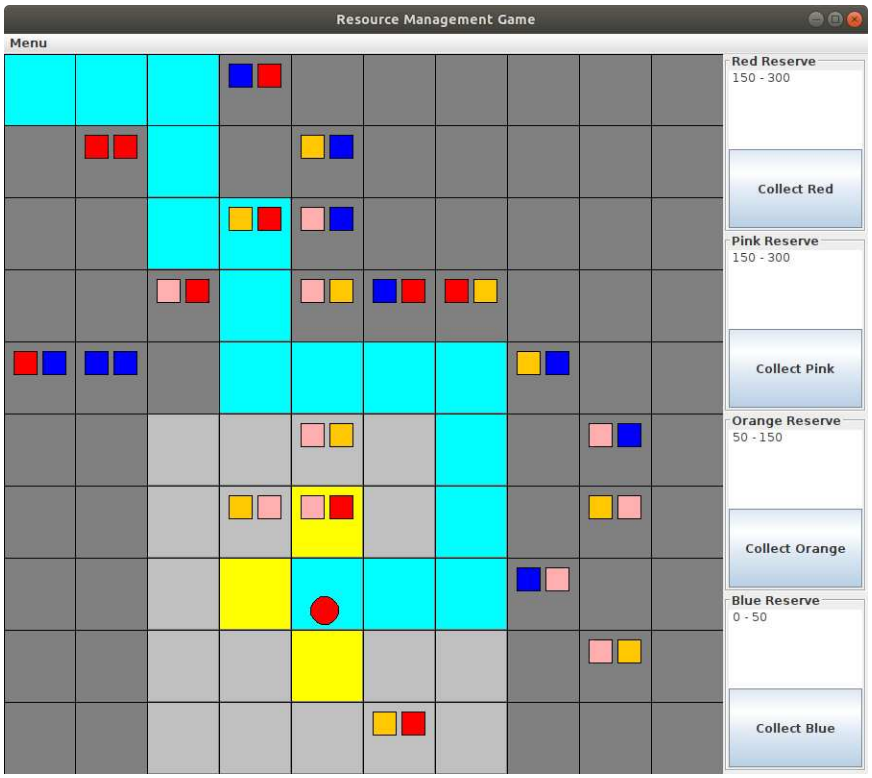
3.2 Software Architecture

The next sections describe the software architecture of the resource management testbed. The resource management testbed was implemented using the Agent Development Environment (ADE) [67], through a total of seven Java classes that are together utilized by the Resouce Management Game ADE Component:

- ResourceManagementGameComponent.java: The main method of the resource management component. This class is integrated as an ADE component.
- ResourceManagementGameComponentImpl.java: The implementation of ResourceManagementGameComponent.java contains the methods to communicate with other ADE components, such as turtlebotcomponent, which is used to control the movement of a turtlebot robot.
- AudioPlayer.java: The class is used to play pre-recorded wav files for proactive explanations and welcome sentences.
- Board.java: The GUI of the board for the resource management task plays a crucial role to build other GUIs. The resource distribution algorithm is implemented. The method of pop up math questions is also included.
- BoardCell.java: The GUI of each board cell within the board.



(a) 5 Resource Cells



(b) 20 Resource Cells

Figure 3.2: Resource Distribution Algorithm

- Resource.java: The GUI of resource request blocks.
- ResourceCollectionAlgorithm.java: The class of pre-defined resource collection algorithm maintains the amounts of each type of resources.
- ResourceMonitor.java: The GUI of the monitoring system with buttons for the user to make requests for the turtlebot robot to collect different types of resources.
- Player.java: The GUI of the player is represented as a red circle on the board.

The following sections include more details about the implementation of each of the ADE components used in this implementation.

3.2.1 Resource Management Component

The resource management component is integrated with the Java-based Agent Development Environment (ADE), which already includes Components for controlling the Turtlebot robot used in our testbed and experiment. When the participant or the robot makes a new decision to change the current resource type from one to another, The Resource Management Component executes a Java RMI service call to ADE's Turtlebot component to trigger its movement. The Resource Management Component also maintains a GUI that serves as the main JFrame to hold other GUIs, such as the testbed's Board and MenuBar. Finally, this component maintains two threads during the task. The first thread updates the testbed GUIs, calculates the levels of each resource, and consumes resource blocks as necessary. This thread runs every 1 second. The second thread is used to call the turtlebot to update the current resource type to collect based on the resource selection algorithm (and, potentially, triggers generation of a proactive explanation as described later in this thesis). This thread updates every 45 seconds.

3.2.2 Board GUI

The Board GUI displays 10 columns and 10 rows. A mouse listener is implemented to enable a user's movement by clicking on the screen.

3.2.3 Board Cell Manager

The BoardCell Manager coordinates with the Board GUI by storing and updating the information on each board cell, including the resource request blocks contained within the cell and the color of the cell; yellow if it is the current target, light grey if it is visible but not the target, and dark grey if it has not yet been revealed, as shown in Figure 3.1.

3.2.4 Resource Manager and Selector

The resource manager and selector maintains the number of collected blocks available for each resource, automatically collects 5 resources per second for the robot's currently targeted resource, and provides the resource selection algorithm described in the previous section.

3.2.5 Resource Monitor

The resource monitor displays the range of collected resources for each type of resource, and provides buttons for participants to select the resource type to collect. All four buttons are disabled for three seconds after one of the buttons is pressed, so that the robot has sufficient time to rotate towards the new block type. Figure 3.1 shows these monitors and buttons.

In the next section we describe how this resource management testbed is deployed in the context of a human-subject experiment.

CHAPTER 4
EXPERIMENT

4.1 Research Goal

The main research goal of our study is to understand the fundamental relationship between human-robot trust of robots and proactive explanations. Specifically, we seek to assess the following three hypotheses.

- **H1**: Robots that generate proactive explanations will be more trusted than robots that do not.
- **H2**: There is a negative correlation between monitoring of a robot and human-robot trust.
- **H3**: Proactive explanations that explain why a robot is about to perform an action will have more influence on human-robot trust than proactive explanations that only inform the participant about the robot's action.

For the effects of proactive explanations on human-robot trust, we believe that compared with providing no explanations, providing proactive explanations can bring high transparency which can in turn lead to high human-robot trust. Similarly, proactive explanations with more information give higher level of transparency than proactive explanations with less details, which we would thus expect to lead to an equivalently higher level of human-robot trust. Finally, this research seeks to confirm or refute the existence of a negative relationship between human monitoring of robots and state human-robot trust (cp. [68]).

4.2 Experiment Design

In order to study the effects of robot's proactive explanations, we designed a human-subject experiment in which a robot used one of three explanation behaviors:

1. **'What' Proactive explanations (PE1)**: During the human-subject experiment, the robot informed participants of *what* resource it planned to collect. For example, a turtlebot robot would tell the player which type of resource will be collected ("I am going to collect blue resources"). The proactive explanation will provide to the human player whenever the turtlebot robot decides to change the current collecting resource type from one to another.

2. **'Why' Proactive explanations (PE2)**: This type of proactive explanation includes the information of both the next step and the reason behind the future action. For this study, the turtlebot robot will not only inform its next action but also why it makes the decision to the participant. For example, the turtlebot robot will explain to the human player which type of resources and why it picks the resources ("I am going to collect red resources because you are low on red resources, but it seems that you may need a lot of them."). This type of explanation is given to the human teammate when the turtlebot robot changes the current resource type.

3. **No explanations (NE)**: During the experiment, the turtlebot robot does not provide any explanation about its decision/action to the human player. This control condition is used to against other conditions by providing proactive explanations with partial or full descriptions.

Each participant experienced all three conditions through a within-subjects Latin square design. Table 4.1 shows the details of the well-designed sequences. Dividing participants into diverse orders of proactive explanation types allows measuring the changes in trust among different experiment conditions. One of the main advantages of the application of this method is to control variation in more than one different directions without a large number of runs [69]. The expected number of participants for each sequence is 15-20. The anticipated total number of participants is 45-60. Each participant was randomly assigned to one of the three conditions.

Table 4.1: Three Sequences of Different Proactive Explanation Conditions

Sequence 1 (S1)	PE1	PE2	NE
Sequence 2 (S2)	NE	PE1	PE2
Sequence 3 (S3)	PE2	NE	PE1

For the human-subject experiment, our original plan was to have 45-60 participants. This human-subject experiment was approved by the Human Subjects Research (HSR) Team at Colorado School of Mines, with the title: HRI 011: Effects of Proactive Explanation on human-robot trust.

4.3 Procedure

Participants were recruited through our experiment ad on campus and contacted us to participate in the experiment. The experiments were performed at Brown hall with a real robot. Upon arriving at the lab and providing informed consent, participants were guided through the following experimental phases:

1. Pre-interaction: An introduction session is used to provide the instructions of the experiment to the participant. The participant is invited to experiment with the testbed and get familiar with the setup of the experiment before he or she starts the testing session. The participant is asked to fill a trust survey check the participant's initial trust of robots as a ground level of trust. The participant is allowed to take unlimited time to answer the survey. Once he or she finishes the survey, the participant moves to the next step.

2. First interaction: The participant joins the first experiment session. The experiment condition is picked from the assigned sequence (Table 4.1). The estimated time for this session is less than 10 minutes.

3. First post-interaction survey: The participant fills another trust survey to measure the participant's trust after the first session. All participants are given enough time to

answer the questionnaires.

4. Second interaction: Experiment session 2 is the same as the experiment session 1. The test condition is based on the selected sequence.

5. Second post-interaction survey: Participant trust after the second experiment session is measured through an additional trust survey.

6. Third interaction: Finally, participants participate in the final experimental condition.

7. Third post-interaction survey: The participant then takes the final trust survey.

8. Explanation and Compensation: The goal of the study is explained to the participant, who is then paid and thanked.

4.4 Interaction Design

Each experiment session lasted less than 10 minutes, during which participants perform the resource management task described in Chapter 3. During this task, the robot used the following exact phrases when generating explanations:

1. Proactive explanations contain partial description to explain the robot's next action.

    - I am going to collect red resources.
    - I am going to collect pink resources.
    - I am going to collect orange resources.
    - I am going to collect blue resources.

2. Proactive explanations with full description explain the robot's next target and the reason behind its decision.

    - I am going to collect red resources because you are low on red resources, but it seems that you may need a lot of them.

- I am going to collect pink resources because you are low on pink resources, but it seems that you may need a lot of them.
- I am going to collect orange resources because you are low on orange resources, but it seems that you may need a lot of them.
- I am going to collect blue resources because you are low on blue resources, but it seems that you may need a lot of them.

4.5 Measures

A 14 item scale [60] is selected for the human-subject experiment to measure the human-robot trust of the turtlebot robot. During the human-subject experiments, all participants are asked to fill a trust survey with 14 questions four times. Figure 4.1 displays the trust survey for the human-subject experiment. The trust survey is stored on Google Drive and allows a participant to click the click-box to record his or her responses. All participants were given unlimited time to answer the questionnaire during the experiments. We also record the amount of time that a user is monitoring a robot as an objective measures for human-robot trust.

		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
	What % of the time will this robot be …											
1	**Dependable**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
2	Reliable	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
3	**Unresponsive**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
4	Predictable	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	What % of the time will this robot …											
5	**Act consistently**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
6	Function successfully	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
7	**Malfunction**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
8	Have errors	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
9	**Provide feedback**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
10	Meet the needs of the mission/task	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
11	**Provide appropriate information**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
12	Communicate with people	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
13	**Perform exactly as instructed**	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
14	Follow directions	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

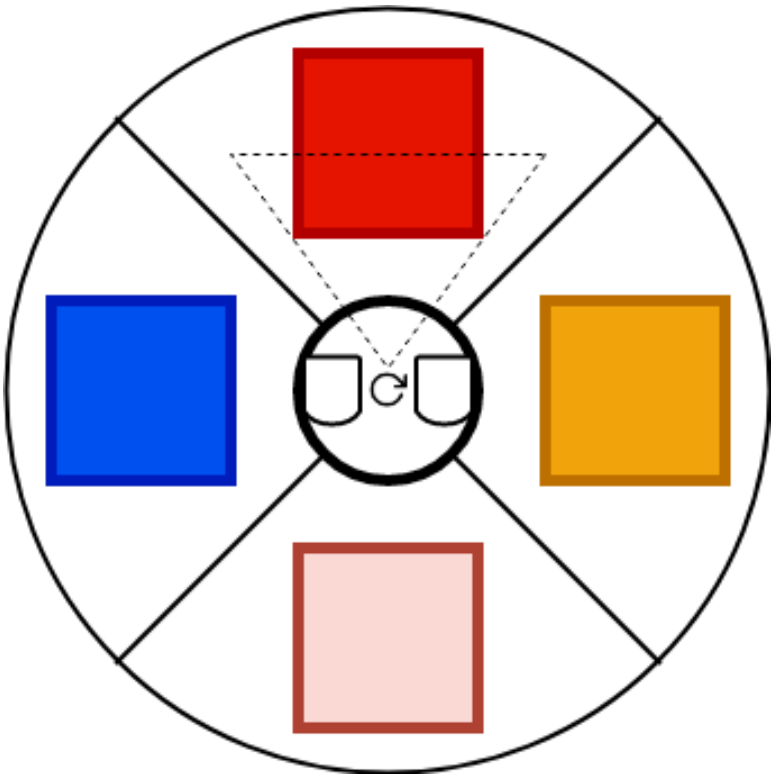
Figure 4.1: 14-Items Trust Survey

4.6 Turtlebot Robot Setup

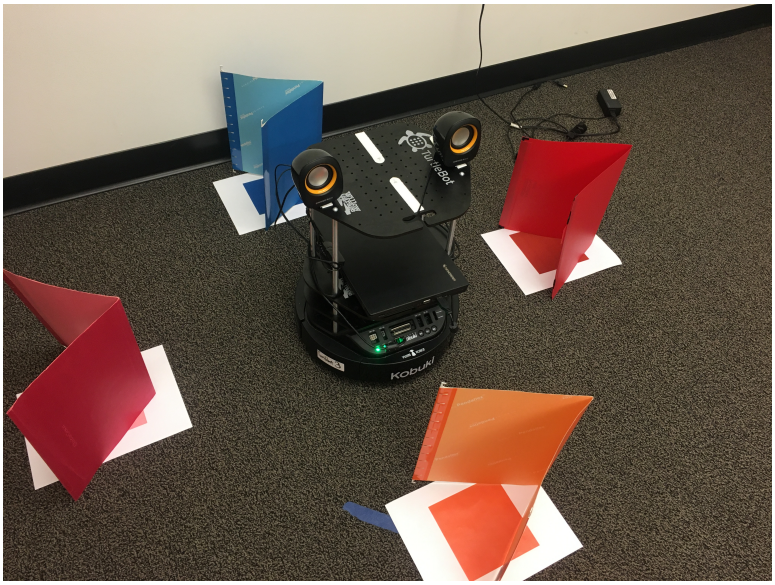
The robot was surrounded by four colored resources and turned to face whichever block it was currently collecting. This allowed the user to identify what resource was currently being collected by turning around and examining the robot. When the robot is facing one of the resource types, this indicates the type of resource that is being collected. Two speakers were placed on the top of the turtlebot to play pre-recorded proactive explanations to participants during the experiment. All participants were informed that the front face of the turtlebot robot is where the speaker is pointing. Figure 4.2(a) presents the locations of the turtlebot robot and the resources in the 2D view. Since the turtlebot robot is surrounded by different types of resources, the only necessary movement of the robot is to rotate itself to face the correct resource type.

The turtlebot robot is controlled by the resource management testbed autonomously during the task. A Thinkpad laptop is connected and carried by the turtlebot. The laptop runs all the ADE components locally and shares the screen with the participant's computer to allow him or her to perform the resource management task and control the turtlebot remotely.

Figure 4.2 illustrates the relative locations of the robot and four different resources. Since the location of each resource is fixed during the experiment, we can program the robot to rotate from the current resource type to the target resource type. For example, when the type of resources changes from red to pink, the robot will turn 180 degrees clockwise. The same logic applies to resource type changes for pink-red, orange-blue, and blue-orange. When the resource type is turned from red to orange, the robot will need to turn 90 degrees clockwise. For the change from red resources to blue resources, the robot will turn 90 degrees counterclockwise instead of turn 270 degrees clockwise. The turtlebot robot maintains and updates the current type of resources before and after it finishes the rotation.



(a) Turtlebot Robot Setup in the 2D View



(b) Turtlebot Robot Setup With Resources

Figure 4.2: Turtlebot Robot Setup for the Human-subject Experiment

4.6.1 Experiment Setup

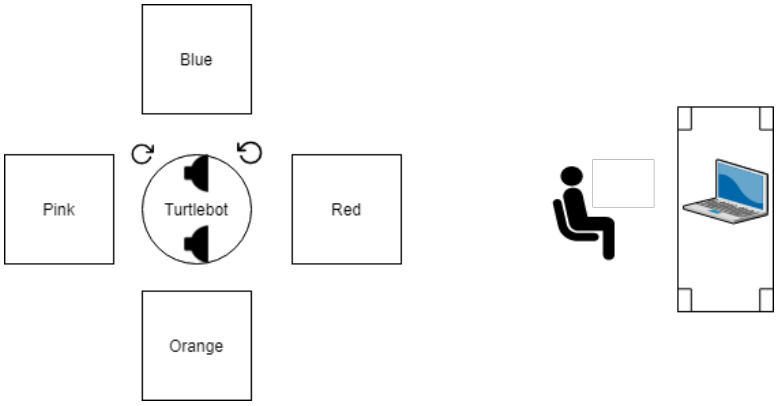
The described experiment required one experimenter, one participant, and one turtlebot robot. Participants were guided to sit in front of our lab's desktop. The turtlebot robot and resource blocks were located behind the participant. The amount of time that the player was monitoring the turtlebot was recorded by using cameras mounted in the corners of the laboratory. Figure 4.3(a) presents the positions of participants and the turtlebot robot in the 2D view. Figure 4.3 presents the general setup of the human-subject experiment in real-world conditions. Hence, we can easily record the amount of time that the player is monitoring the turtlebot by using the cameras mounted in the lab. The major advantage of this setup is that researchers can directly measure participants' monitoring time by checking the movement of their upper bodies.

4.7 Results

For this human-subject experiment, I ran a total of 32 participants on campus. The majority of the participants were university students. 32 experiments were performed and 21 valid datasets were used after removing data from participants who performed actions that required experimenter intervention (e.g., accidentally closing the testbed window). For Sequence 1 (PE1-PE2-NE) from Table 4.1, we have 8 valid datapoints. The number of valid datapoints for Sequence 2 (NE-PE1-PE2) is 6. The last sequence S3 (PE2-NE-PE1) has 7 valid human-subject datapoints.

4.7.1 human-robot trust

Table 4.2 and Table 4.3 present summary statistics on raw survey responses and gain scores (i.e. per-participant differences from pre-experiment to after each within-subjects experimental condition) for human-robot trust throughout the experiment. Participants' gain scores were analyzed using a Bayesian Analysis of Variance, followed by calculation of Baws scores (Bayes Inclusion Factors on Matched Models). Before presenting the results of this analysis we briefly explain how to interpret the Bayes Factors produced by this analysis.



(a) Experiment Setup in the 2D View

(b) Experiment Setup in the Real World Condition

Figure 4.3: Turtlebot Robot Setup for the Human-subject Experiment

Table 4.2: Reported Human-Robot Trust

					95% Credible Interval	
	N	Mean	SD	SE	Lower	Upper
Pre-Interaction	21.000	7.078	1.023	0.223	6.612	7.544
NE	21.000	6.684	1.109	0.242	6.179	7.189
PE1	21.000	7.524	1.042	0.227	7.049	7.998
PE2	21.000	7.844	1.034	0.226	7.373	8.315

Table 4.3: Trust Gain Scores

				95% Credible Interval	
Trust Gain	Mean	SD	N	Lower	Upper
NE	-0.395	0.963	21.000	-0.833	0.044
PE1	0.446	1.343	21.000	-0.166	1.057
PE2	0.766	1.446	21.000	0.107	1.424

Bayes factors can roughly be interpreted as ratios of evidence in favor of alternative hypotheses relative to competing (e.g., null) hypotheses [70]. Bayes factors between 0.33 and 3 are generally taken as anecdotal evidence [71] insufficient to confirm or refute a hypothesis. Bayes factors between 0.33-0.10 or 3-10, in turn provide substantial evidence against or for the hypothesis in question; Bayes factors between 0.03-0.10 and 10-30 provide strong evidence; and Bayes factor less than 0.01 or greater than 100 provide decisive evidence to assess hypotheses.

As shown in Table 4.4, our analysis provides decisive evidence in favor of an effect of proactive explanation condition on human-robot trust. Here, the Bayes Factor of 222.649 indicates that the collected data were 222 times more likely to have been generated under a model accounting for proactive explanation condition than one that does not. To interrogate this difference, we performed post-hoc pairwise comparisons between experimental conditions.

The results of these pairwise post-hoc analyses are summarized in Table 4.5. As shown in Figure 4.4(b), our results specifically suggest that the trust gains in the two proactive

Table 4.4: Bayesian ANOVA Results on Trust Gain Scores

Models	P(M)	P(M—data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.500	0.004	0.004	0.004	0.975
trust gain	0.500	0.996	222.649	1.000	

Table 4.5: Post Hoc Comparisons of Human-Robot Trust Gain

		Prior Odds	Posterior Odds	$BF_{10,U}$	error %
NE	PE1	0.587	4.578	7.794	0.001
	PE2	0.587	24.784	42.193	1.402e-4
PE1	PE2	0.587	0.742	1.264	7.902e-4

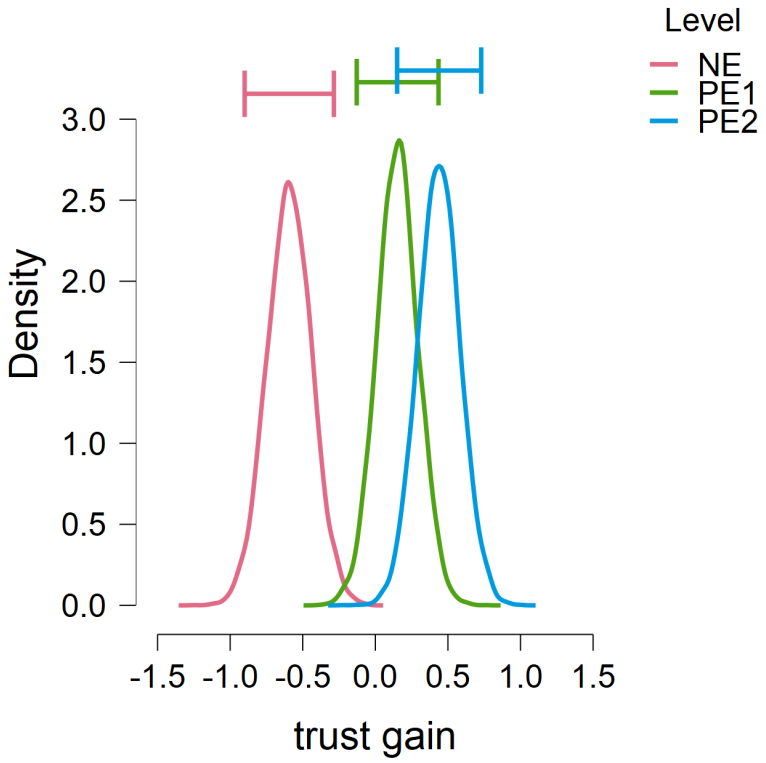
explanation conditions were greater than in the no explanation condition, with stronger evidence in favor of significant gains in trust when the robot delivered a proactive explanation explaining why an action was being taken (PE2, Bf 24.784) than when the robot delivered a proactive explanation that only explained what action was being taken (PE1, Bf 4.578). Comparison between these conditions, however, produced weak evidence insufficient to either confirm or refute a difference between those strategies (Bf 1.264). Overall, these results strongly suggest that human-robot trust is more likely to increase when robots proactively explain their actions.

4.7.2 Human Monitoring of Robots

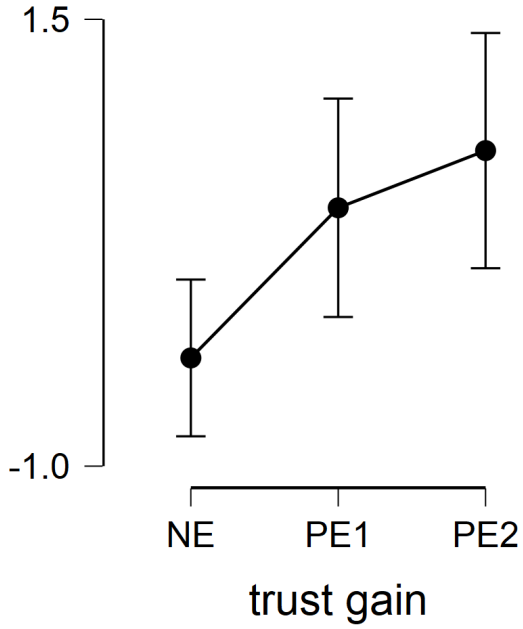
Table 4.6: Percentage of Turn Around Time Data Description

				95% Credible Interval	
Percentage of Turn Around	Mean	SD	N	Lower	Upper
Turn Around - NE	4.160	3.052	17.000	2.591	5.729
Turn Around - PE1	3.268	3.158	17.000	1.645	4.892
Turn Around - PE2	2.488	2.925	17.000	0.984	3.991

Participants' time spent monitoring the robot was assessed by watching the video files recorded by the camera system mounted in the lab. Unlike the datasets for human-robot



(a) Distribution of Trust Gain



(b) Trust Gain among Three Experiment Conditions

Figure 4.4: Bayesian ANOVA for Experiment Conditions on human-robot trust Gain

trust, we have a total of 17 available video recordings, as four video recordings were missing due to the failure of the camera system. From the 17 video recordings, two types of observation are measured. The first measurement is the amount of monitoring time, which is when a participant turns around and checks the condition of the robot. The second measurement is the number of monitoring of the turtlebot robot for each participant. Table 4.6 presents the data description of percentage of turn around time. A percentage of turn around time was calculated by finding the ratio between the amount of monitoring time for a specific task session and the time that the participant finished the task session.

Table 4.7: Model Comparison of Monitoring Time

Models	P(M)	P(M—data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.500	0.488	0.951	1.000	
Percentage of Turn Around	0.500	0.512	1.051	1.051	0.803

As shown in Table 4.7, our data analysis provides anecdotal evidence in favor of an effect of proactive explanations on the percentage of monitoring time. The Bayes Factor of 1.051 suggests the collected data were only 1 times more likely to have been generated under a model with proactive explanation condition the the model without proactive explanation condition. Post-hoc pairwise comparisons were performed between experimental conditions.

Table 4.8: Post Hoc Comparisons of Monitoring Time

		Prior Odds	Posterior Odds	$BF_{10,U}$	error %
Turn Around - NE	Turn Around - PE1	0.587	0.258	0.440	6.640e-4
	Turn Around - PE2	0.587	13.766	23.436	5.425e-5
Turn Around - PE1	Turn Around - PE2	0.587	0.215	0.366	0.003

Table 4.7 presents the results of these pairwise post-hoc analyses. As shown in Figure 4.5(b), our results suggest that percentage of monitoring time in the proactive explanations conditions had a larger value than in the condition of no explanation. However, the

Bayes Factor of 0.440 suggests anecdotal evidence to either confirm or refute a difference between percentage of monitoring time between in the no explanation condition and proactive explanation explaining what action was being taken by a robot. The Bayes Factor of 23.436 proposes strong evidence in favor of increase in percentage of human monitoring when the robot provides a 'why' proactive explanation (PE2, Bf 23.436) than when the robot gave no proactive. Comparison between two proactive explanation conditions suggested weak evidence to either confirm or refute a difference between percentage of monitoring time between when the robot delivered a 'what' proactive explanation and when the robot delivered a 'why' proactive explanation. In conclusion, these results cannot confirm or refute that human monitoring of a robot is more likely to increase when robots gives proactive explanations explaining what action was being taken.

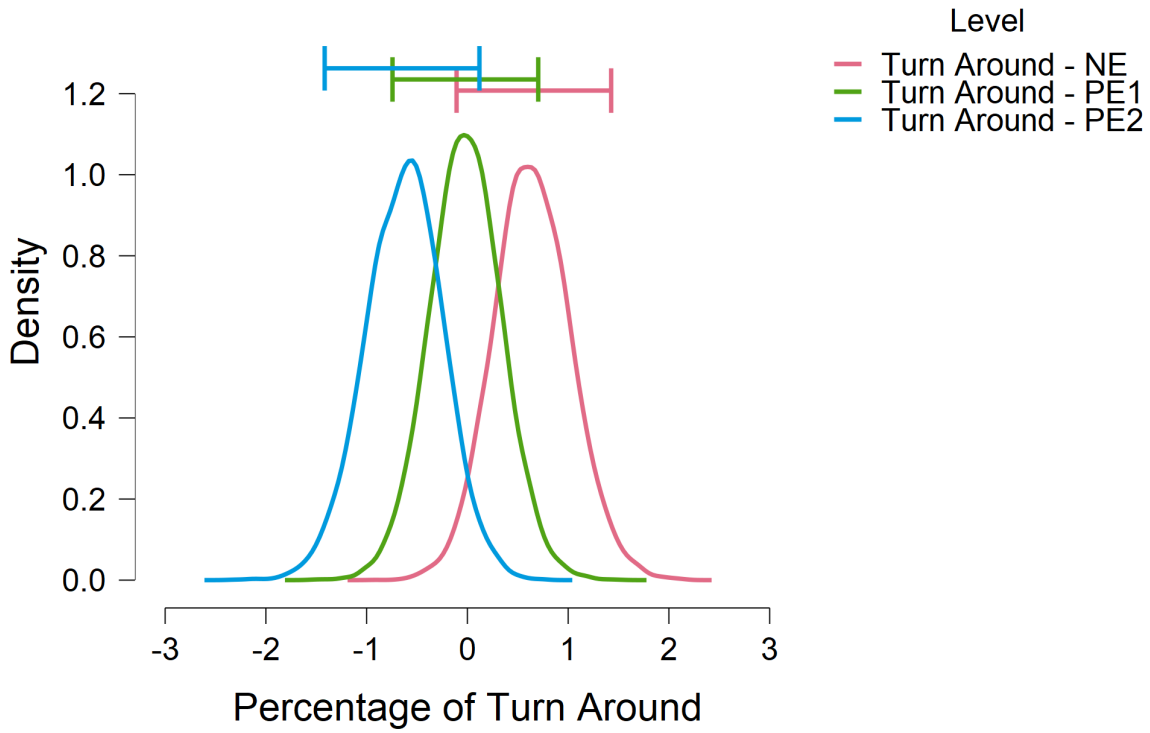
Table 4.9: Turn Around per Second Data Description

				95% Credible Interval	
Turn Around per Second	Mean	SD	N	Lower	Upper
NE	0.026	0.019	17.000	0.017	0.036
PE1	0.021	0.018	17.000	0.012	0.030
PE2	0.016	0.018	17.000	0.007	0.025

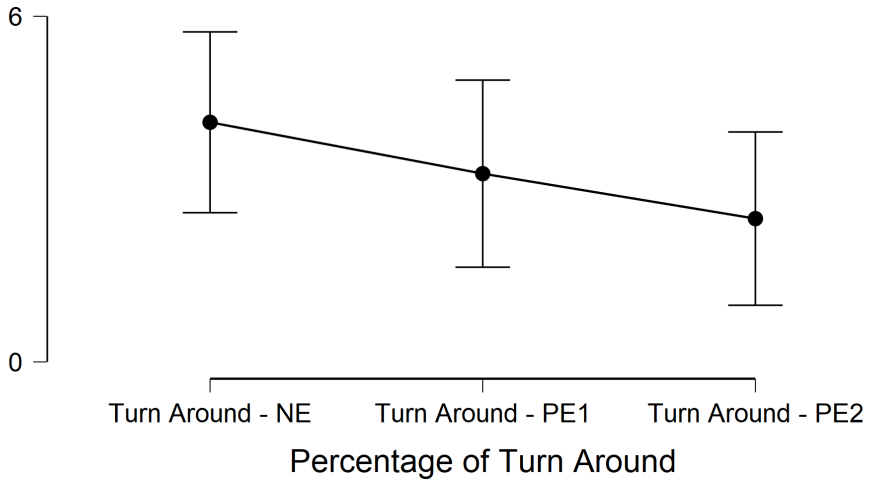
Table 4.10: Bayesian ANOVA Results on Turn Around per Second

Models	P(M)	P(M—data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.500	0.435	0.771	1.000	
Turn Around per Second	0.500	0.565	1.298	1.298	0.498

Table 4.9 presents summary statistics on the number of turn around per second throughout the experiment. As shown in Table 4.10, the Bayes Factor of 1.298 proposes anecdotal proof for the role of proactive explanations in turn around per second. The Bayes Factor of 1.298 suggests that the collected data were only 1 times more likely to have been generated under a model considering proactive explanation conditions than one that does not. Post-hoc



(a) Distribution of Percentage of Monitoring Time



(b) Percentage of Monitoring Time among Three Experiment Conditions

Figure 4.5: Bayesian ANOVA for Experiment Conditions on Monitoring Time

pairwise comparisons between experimental conditions are provided in the following table.

Table 4.11: Post Hoc Comparisons of Turn Around per Second

		Prior Odds	Posterior Odds	$BF_{10,U}$	error %
NE	PE1	0.587	0.343	0.583	0.002
	PE2	0.587	3.130	5.328	1.591e-4
PE1	PE2	0.587	0.222	0.377	0.002

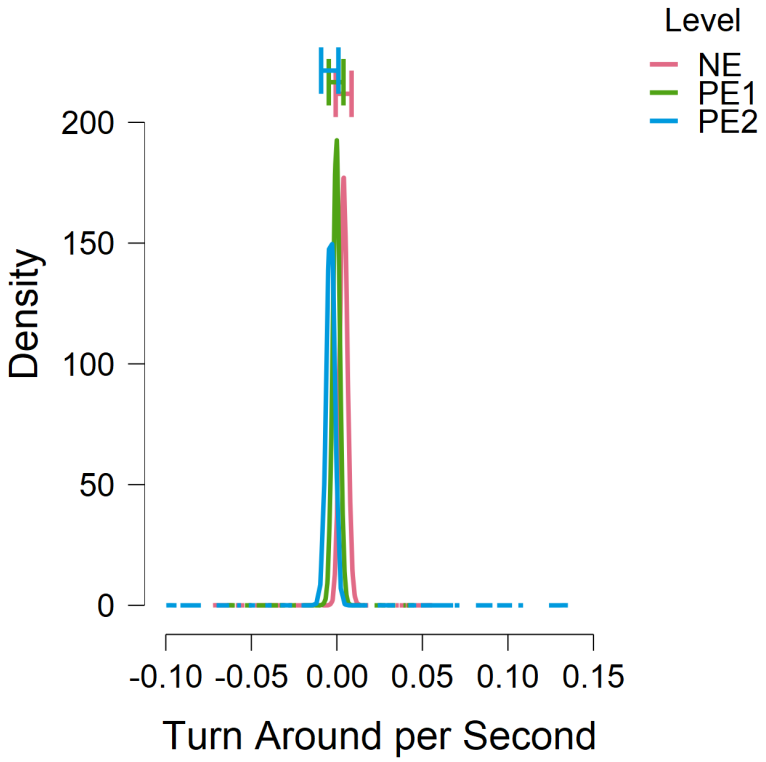
The results of these pairwise post-hoc analyses are summarized in Figure 4.6(b). As shown in Figure 4.6(b), our results specifically suggest that turn around per second when a robot provided proactive explanations explaining why an action was being taken (PE2, Bf 5.328) was less than in the no explanation condition. The Bayes Factor of 0.583 suggested weak evidence insufficient to either confirm or refute a difference of turn around per second when the robot delivered a 'what' proactive explanation. Comparison of turn around per second in the two proactive explanation conditions suggested anecdotal evidence to either confirm or refute that there is a difference of the turn around per second between those strategies (Bf 0.377). Overall, these results weakly suggests that humans' turn around per second is more likely to decrease when robots proactively explain their actions.

4.7.3 Monitoring vs Trust

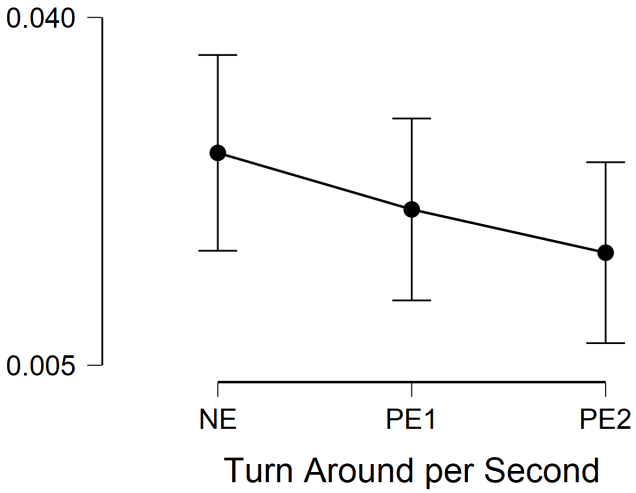
Table 4.12: Bayesian Pearson Correlation Between Monitoring and Human-Robot Trust

			r	BF_{10}
Trust Gain	-	Percentage of Monitoring	-0.276	1.128
	-	#Turn Around per Second	-0.217	0.544

As shown in Table 4.12, our results failed to either confirm or refute either the relationship between percentage of time spent monitoring the robot's behavior and human-robot trust (Bf 1.128), or a relationship between turns per second and human-robot trust (Bf 0.544).



(a) Distribution of Turn Around per Second



(b) Turn Around per Second among Three Experiment Conditions

Figure 4.6: Bayesian ANOVA for Experiment Conditions on Turn Around per Second

CHAPTER 5
DISCUSSION

Our first experimental hypothesis was that proactive explanations would increase human-robot trust relative to when no explanation was given (H1). Our results confirm H1, clearly showing that human-robot trust gain is larger when proactive explanations are presented than the trust gain when no explanations are offered. In other words, proactive explanations provided by a robot to its human teammate can increase the human-robot trust of the robot within a human-robot team. This finding is consistent with the findings from previous studies, which suggested that verbal communication can in general serve to increase the level of transparency for human-robot interaction [9, 11]. Verbal communications, such as the proactive explanations in this study, are able to promote appropriate human-robot trust by maintaining an appropriate level of transparency [8]. Accordingly, the fundamental finding from our investigation reconfirms these findings from previous research. Another advantage of giving proactive explanations is to make humans feel that they work in a team. Anecdotally, several times during the human-subject study, the participants discussed that the task sessions with proactive explanations made them believe they work with a teammate instead of a tool, whereas when the turtlebot robot provided no explanation, participants felt the robot was more like a tool. This is a positive result given the research of Billings et al. [4], which suggests that people should treat robots as real teammates instead of tools to have an effective interaction.

The second hypothesis was that proactive explanations would decrease humans' monitoring relative to when no explanation was given (H2). Our results fail to confirm a correlation between human monitoring and different type of proactive explanations. However, our results (Table 4.8) provide strong evidence that the percentage of monitoring time for PE2 is less than the percentage of turn around time for NE (Bf 23.436). For turn around per

second, we have the same finding, which indicates that participants are more likely to turn around less for experiment conditions of PE2 than the task sessions with no proactive explanations. We had discussed this phenomenon with a few participants at the very end of their experiments. They explained the reason to why they checked more frequently when there were no proactive explanations. When the robot provided proactive explanations, they could gain information about the robot from the explanations. For the experiment conditions of no explanations, they had to turn around to check the robot's status. More data will be needed to confirm the correlation between human monitoring and proactive explanations. In addition to understanding the influence of proactive explanations on human monitoring, we wish to determine the relationship between human monitoring of an autonomous system and human-robot trust. Earlier work in human-automation interaction proposed an inverse correlation between the amount of time to monitor an automaton and human-robot trust [72]. Our results, however, failed to either confirm or refute this previously observed relationship. Therefore, there is no evidence to confirm or refute our second hypothesis.

The third hypothesis was that 'why' proactive explanations would more greatly increase human-robot trust relative to 'what' proactive explanations (H3). Our results provided only anecdotal evidence that the level of human-robot trust was higher when a robot delivered proactive "why" explanations (PE2) than when it delivered only proactive "what" explanations (PE1). Hence, we can either confirm or refute H3 based on our current data analysis, suggesting the need for more data collection. Anecdotally, several participants, who were given a short interview after they finished all task sessions, indicated their preferences for proactive explanations with less information than fully described proactive explanations. Participants reported that PE2 was wordy and not necessary. This extends previous findings by Stange et al. [73] showing that there are no differences between 'why'-explanations and 'what'-explanations on increasing the robot's understandability or desirability. This conclusion can be used to explain our findings.

CHAPTER 6

CONCLUSION

In this work, we conducted a human-subject study to better understand the relationship between human-robot trust and robots' proactive explanations. Our results suggested that proactive explanations lead to increased human-robot trust. Our results failed to confirm or refute, however, any relationship between human monitoring of robots and human-robot trust, or any difference in effectiveness between proactive explanations that explain why a robot is about to perform an action and proactive explanations that only inform the participant about the robot's action itself. Additional data will be needed to either confirm or refute these latter two relationships.

Our results suggest that to increase human-robot trust, robot designers need to design robots that are able to provide appropriate proactive explanations. However, it is not clear what factors may dictate when and how often such explanations should be generated. In future research, it may thus be important to study, for example, the impact of proactive explanations on teammates' mental workload, and to design communication policies for autonomously deciding when to generate proactive explanations on the basis of such factors.

REFERENCES CITED

[1] DR Billings, KE Schaefer, N Llorens, and PA Hancock. What is trust? defining the construct across domains. In *Poster presented at the american psychological association conference, Division*, volume 21, 2012.

[2] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, and Pim Haselager. Do robot performance and behavioral style affect human trust? *International journal of social robotics*, 6(4):519–531, 2014.

[3] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997. doi: 10.1518/001872097778543886. URL `https://doi.org/10.1518/001872097778543886`.

[4] D. R. Billings, K. E. Schaefer, J. Y. C. Chen, and P. A. Hancock. Human-robot interaction: Developing trust in robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–110, March 2012. doi: 10.1145/2157689.2157709.

[5] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. Building appropriate trust in human-robot teams. In *2013 AAAI Spring Symposium Series*, 2013.

[6] Joseph B Lyons. Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*, 2013.

[7] Tove Helldin. *Transparency for Future Semi-Automated Systems : Effects of transparency on operator performance, workload and trust.* PhD thesis, University of SkövdeUniversity of Skövde, School of Informatics, The Informatics Research Centre, 2014.

[8] Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human Factors*, 58(3):401–415, 2016. doi: 10.1177/0018720815621206. URL `https://doi.org/10.1177/0018720815621206`. PMID: 26867556.

[9] Tom McManus, Yair Holtzman, Harold Lazarus, Johan Anderberg, and Ozum Ucok. Transparency, communication and mindfulness. *Journal of Management Development*, 2006.

[10] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. Improving robot transparency: real-time visualisation of robot ai substantially improves understanding in naive observers. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1424–1431. IEEE, 2017.

[11] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. Situated language understanding with human-like and visualization-based transparency. In *Robotics: Science and Systems*, 2016.

[12] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE, 2005.

[13] Mohan Sridharan, Ben Meadows, and Zenon Colaco. A tale of many explanations: towards an explanation generation system for robots. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 260–267, 2016.

[14] Martin Gebser, Tomi Janhunen, Holger Jost, Roland Kaminski, and Torsten Schaub. Asp solving for expanding universes. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 354–367. Springer, 2015.

[15] Ben Leon Meadows, Pat Langley, and Miranda Jane Emery. Seeing beyond shadows: Incremental abductive reasoning for plan understanding. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[16] Nengchao Lyu, Lian Xie, Chaozhong Wu, Qiang Fu, and Chao Deng. Driver's cognitive workload and driving performance under traffic sign information exposure in complex environments: A case study of the highways in china. In *International journal of environmental research and public health*, 2017.

[17] Michael A Goodrich, Alan C Schultz, et al. Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275, 2008.

[18] Vignesh Narayanan, Yu Zhang, Nathaniel Mendoza, and Subbarao Kambhampati. Automated planning for peer-to-peer teaming and its evaluation in remote human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 161–162, 2015.

[19] Akihisa Ohya. Human robot interaction in mobile robot applications. In *Proceedings. 11th IEEE international workshop on robot and human interactive communication*, pages 5–10. IEEE, 2002.

[20] Nhan Tran, Kai Mizuno, Trevor Grant, Thao Phung, Leanne Hirshfield, and Thomas Williams. Exploring mixed reality robot communication under different types of mental workload.

[21] Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480): 679–704, 2007.

[22] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 426–434, 2017.

[23] Alberto Montebelli, Erik Billing, Jessica Lindblom, and Giulia Messina Dahlberg. Reframing hri education: a dialogic reformulation of hri education to promote diverse thinking and scientific progress. *Journal of Human-Robot Interaction*, 6(2):3–26, 2017.

[24] Tiffany Hwu, Jacob Isbell, Nicolas Oros, and Jeffrey Krichmar. A self-driving robot using deep convolutional neural networks on neuromorphic hardware. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 635–641. IEEE, 2017.

[25] Jochen Heinzmann and Alexander Zelinsky. A safe-control paradigm for human–robot interaction. *Journal of Intelligent and robotic systems*, 25(4):295–310, 1999.

[26] Masaki Takahashi, Takafumi Suzuki, Hideo Shitamoto, Toshiki Moriguchi, and Kazuo Yoshida. Developing a mobile robot for transport applications in the hospital domain. *Robotics and Autonomous Systems*, 58(7):889–899, 2010.

[27] J Evans, B Krishnamurthy, B Barrows, T Skewis, and V Lumelsky. Handling real-world motion planning: a hospital transport robot. *IEEE Control Systems Magazine*, 12(1): 15–19, 1992.

[28] Qingxiao Yu, Can Yuan, Zhuang Fu, and Yanzheng Zhao. An autonomous restaurant service robot with high positioning accuracy. *Industrial Robot: An International Journal*, 2012.

[29] Robert L Cahlander, David W Carroll, Robert A Hanson, Al Hollingsworth, and John O Reinertsen. Food preparation robot, May 1 1990. US Patent 4,922,435.

[30] Igor M Verner, Alex Polishuk, and Niv Krayner. Science class with robothespian: using a robot teacher to make science fun and engage students. *IEEE Robotics & Automation Magazine*, 23(2):74–80, 2016.

[31] Iolanda Leite, Carlos Martinho, and Ana Paiva. Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, 5(2):291–308, 2013.

[32] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. Understanding social robots. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 169–174. IEEE, 2009.

[33] Satoru Satake, Takayuki Kanda, Dylan F Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to approach humans? strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 109–116, 2009.

[34] G Litzenberger. World robotics 2008. *International Federation of Robotics, c/o VDMA Robotics+ Automation, Frankfurt aM*, 2008.

[35] Omar Mubin, Catherine J Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, 1(209-0015):13, 2013.

[36] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D Janse. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1613–1622, 2010.

[37] Jeonghye Han and Dongho Kim. r-learning services for elementary school students with a teaching assistant robot. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 255–256. IEEE, 2009.

[38] Joris B Janssen, Chrissy C van der Wal, Mark A Neerincx, and Rosemarijn Looije. Motivating children to learn arithmetic with an adaptive robot game. In *International Conference on Social Robotics*, pages 153–162. Springer, 2011.

[39] Takuya Hashimoto, Hiroshi Kobayashi, Alex Polishuk, and Igor Verner. Elementary science lesson delivered by robot. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 133–134. IEEE, 2013.

[40] William Joseph Church, Tony Ford, Natasha Perova, and Chris Rogers. Physics with robotics—using lego mindstorms in high school education. In *2010 AAAI Spring Symposium Series*, 2010.

[41] Kate Highfield, Joanne Mulligan, and John Hedberg. Early mathematics learning through exploration with programmable toys. In *Proceedings of the Joint Meeting of PME*, volume 32, pages 169–176. Citeseer, 2008.

[42] J. Casper and R. R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3):367–385, June 2003. ISSN 1941-0492. doi: 10.1109/TSMCB.2003.811794.

[43] Robin R Murphy. Activities of the rescue robots at the world trade center from 11-21 september 2001. *IEEE Robotics & Automation Magazine*, 11(3):50–61, 2004.

[44] Ming-Yuan Shieh, JC Hsieh, and CP Cheng. Design of an intelligent hospital service robot and its applications. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 5, pages 4377–4382. IEEE, 2004.

[45] Masaki Takahashi, Takafumi Suzuki, Francesco Cinquegrani, Rosario Sorbello, and Enrico Pagello. A mobile robot for transport applications in hospital domain with safe human detection algorithm. In *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1543–1548. IEEE, 2009.

[46] Sara Ljungblad, Jirina Kotrbova, Mattias Jacobsson, Henriette Cramer, and Karol Niechwiadowicz. Hospital robot at work: something alien or an intelligent colleague? In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 177–186, 2012.

[47] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40, 2006.

[48] David B Kaber, Emrah Onal, and Mica R Endsley. Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human factors and ergonomics in manufacturing & service industries*, 10(4):409–430, 2000.

[49] Mica R Endsley. Measurement of situation awareness in dynamic systems. *Human factors*, 37(1):65–84, 1995.

[50] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[51] Jill L Drury, Jean Scholtz, and Holly A Yanco. Awareness in human-robot interactions. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, volume 1, pages 912–918. IEEE, 2003.

[52] Michael A Goodrich and Dan R Olsen. Seven principles of efficient human robot interaction. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, volume 4, pages 3942–3948. IEEE, 2003.

[53] P. Robinette, A. M. Howard, and A. R. Wagner. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, 47(4):425–436, Aug 2017. ISSN 2168-2305. doi: 10.1109/THMS.2017.2648849.

[54] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 141–148, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2883-8. doi: 10.1145/2696454.2696497. URL `http://doi.acm.org/10.1145/2696454.2696497`.

[55] Rosemarie Yagoda and Douglas Gillan. You want me to trust a robot? the development of a human-robot interaction trust scale. *International Journal of Social Robotics*, 4, 08 2012. doi: 10.1007/s12369-012-0144-0.

[56] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[57] Christine Moorman, Rohit Deshpande, and Gerald Zaltman. Factors affecting trust in market research relationships. *Journal of marketing*, 57(1):81–101, 1993.

[58] David P. Biros, Mark Daly, and Gregg H. Gunsch. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, 13:173–189, 2004.

[59] Sim B Sitkin and Nancy L Roth. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization science*, 4(3):367–392, 1993.

[60] Kristin Schaefer. The perception and measurement of human-robot trust. 2013.

[61] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman. Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems*, pages 106–114, May 2007. doi: 10.1109/CTS.2007.4621745.

[62] Eui Park, Quaneisha Jenkins, and Xiaochun Jiang. Measuring trust of human operators in new generation rescue robots. In *Proceedings of the JFPS International Symposium on Fluid power*, volume 2008, pages 489–492. The Japan Fluid Power System Society, 2008.

[63] George Charalambous, Sarah Fletcher, and Philip Webb. The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, 8(2):193–209, 2016.

[64] Tracy Sanders, Kristin E Oleson, Deborah R Billings, Jessie YC Chen, and Peter A Hancock. A model of human-robot trust: Theoretical model development. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 55, pages 1432–1436. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[65] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[66] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

[67] Matthias Scheutz. Ade: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, 20(2-4): 275–304, 2006.

[68] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392. URL `https://doi.org/10.1518/hfes.46.1.50_30392`. PMID: 15151155.

[69] Lei Gao. Latin squares in experimental design. *Michigan State University*, 2005.

[70] Andrew F Jarosz and Jennifer Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, 7(1):2, 2014.

[71] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[72] BONNIE M. MUIR and NEVILLE MORAY. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996. doi: 10.1080/00140139608964474. URL `https://doi.org/10.1080/00140139608964474`. PMID: 8849495.

[73] Sonja Stange and Stefan Kopp. Effects of a social robot's self-explanations on how humans understand and evaluate its behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 619–627, 2020.