Failure Is an Option: How the Severity of Robot Errors Affects Human-Robot Interaction

Cecilia Gabriela Morales Garza

CMU-RI-TR-18-59

The Robotics Institute

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

Thesis Committee:

Aaron Steinfeld (Chair)

Jodi L. Forlizzi

Xiang Zhi Tan

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Robotics.

August 2018

Keywords: Human-Robot Interaction, Robotic Failure, Risk, Trust, Safety, Assistance

Contents

List of Figures

List of Tables

1 Abstract

Just as humans are imperfect, even the best of robots will eventually fail at performing a task. The likelihood of failure increases as robots expand their roles in our lives. Although failure is a common problem in robotics and human-robot interaction (HRI), there has been little research investigating people's tolerance to said failures, especially when there is a risk of property damage and bodily harm. Safety is an important concern for human-robot interaction, and robot designers need to understand how people calibrate their levels of trust and adapt their behavior around robots that could expose them, and property, to physical harm.

To explore this issue, we performed an experiment where people were exposed to failure in a study with actual personal and property risk. Participants observed a Baxter robot while it performed a grocery packing task, and were given opportunities to react to and assist the robot in multiple failure cases. The study revealed important factors that influence trust, perception of safety, and whether participants would assist the robot after witnessing failure. Some of these findings were that the severity and recency of failures are among the most influential factors that influence human reports of trust in a robot. We also observed lower ratings of trust in the robot from female participants relative to male participants. While the majority of the participants assisted the robot when it failed, they were more likely to assist when the participants had not observed other failures prior to assisting the robot.

By understanding how people respond to robot failure and aspects of robot behavior that influence their trust, better understanding and design can be incorporated into robots. This should increase human comfort levels and willingness to interact and work with robots.

2 Acknowledgments

I would like to thank the members of my committee, Dr. Aaron Steinfeld, Dr. Jodi Forlizzi, and Xiang (Zhi) Tan, for their endless support, patience, and advice that made this research possible. Aaron, your encouragement, passion, and guidance, motivated me to push myself harder throughout my time in the lab. Thank you for always believing in me and caring about me, making sure I succeeded in tasks not only inside the lab. You were a great role model, and I learned immensely from you. I will take these lessons wherever I go. Your advice has made me gather a different perspective and also shape the way I do research. My research experience was completely changed when I joined the lab, and I attribute this to your endless guidance and support you provide all of us. Your passion is contagious. Zhi, thank you for everything you taught me. You were like a second advisor to me. Thank you for always challenging me, and pushing me to learn more. Jodi, thank you for your interest, insights and help in the direction of my thesis.

Thank you to all the members in my lab. Especially Amal and Liz, for all the discussions and all the work that you contributed to the making of this thesis. Amal, your hard work was inspiring and your advice always improved my research, and Liz, this thesis wouldn't be what it is without you. To the rest of the lab members, thank you for listening to the experiment 64 times, and for always supporting me. You made my time much more enjoyable.

I would like to also thank my family. Without your love and support, it would have not been possible for me to accomplish everything I have. Thank you for sacrificing so much and never doubting my abilities. You never stopped believing in me, and did everything possible for me to succeed, I will never be able to repay everything you've done for me.

Thank you to all my teachers, mentors, professors, friends, and people that were there next to me throughout this process and enabled me to reach this point. I can't imagine what my experience would have been like without your support, care, encouragement, love, wisdom, and patience. Thank you for always pushing me to succeed and be there throughout my accomplishments and all of my ventures. Finally, this work was supported by the National Science Foundation (Grant Numbers IIS-1552256 and CBET-1317989) and the U.S. Department of Transportation funded T-SET University Transportation Center.

3 Introduction

Human-Robot Interaction (HRI) has recently become more popular as a research field due to the increase in the availability of complex robots and people's exposure to them. Robots can be used in many different applications, including the automotive industry, assembly, medical applications, agriculture, space exploration, search and rescue, education, customer service, entertainment, and home appliances. While 77 percent of people think it will be normal to have a robot in their home in the next 20 years, fewer fully trust robots [16]. A Simple Queue Survey (SQS) showed that although 60 percent of respondents thought a robot would be useful and would save them time in their household, 76 percent of British people said they didn't believe home robots to be safe. When presented with a scale from zero to 10, with zero meaning "not excited at all" and 10 meaning "extremely excited," 59 percent of Americans chose a number of five or lower. In the same survey, only 15 percent of Americans said they were "extremely excited" about self-driving cars. Thus, it is apparent that many consumers are skeptical about the role of intelligent robots and technologies in our daily lives.

If these trust issues are not addressed, progress in robotics might stall because people will not be willing to buy robot products for their household. Emerging technology can evoke feelings of risk and loss of control, so safety is one of the primary concerns. People's perception regarding the overall safety of robots needs to change before they can become part of our daily lives [16]. To address this issue, researchers have become more interested in exploring the understanding and evaluation of different interactions between people and robots, along with people's perception of different types and behaviors of robots, and how they perceive social cues or different robot embodiments.

3.1 Motivation and Problem Statement

Some of the roles that robots currently have in society include being a machine operating without human contact, a tool in the hands of a human operator, a peer, or a part of a body, among others. To be able to perform these useful tasks alongside humans, safety is a primary concern. In some circumstances, such as in a home or an office, some of the

traditional safety solution features of a robot in industry are not acceptable, such as using warning alarms or flashing lights. Safety in HRI usually includes the avoidance of physical harm to a human due to any sort of collision; in the cases where physical interactions are required, other strategies can be adopted and metrics developed.

While our ability to develop reliable autonomous systems and robots is constantly improving, systems are not immune to failure [15]. The way in which a robot fails can affect a user's perspective about the system. Even non-harmful interactions could be perceived as uncomfortable by a human, such as when a robot invades the user's personal space by coming in close proximity. Therefore, the perception of safety can be a subjective parameter determined by an individual. This is subject to change over time as people become more aware of the functionalities and limitations of different robots, allowing their behaviors to become more predictable. Until recently, much of the research on safety in HRI related only to the technical requirements of the robot rather than the behavioral and social mechanisms that impacted people's feelings and opinions [24].

While it is known that people's trust in and willingness to work with a robot is lowered by failure, less is known about how people respond or behave after different types of failure, especially when their safety and personal risk is compromised. Moreover, research must address what can be done to mitigate the feelings of frustration, anxiety, anger, fear, resentment, or distrust that could arise as a result of such failure. By understanding how people respond to robotic failure and aspects of robot behavior that influence their trust, better understanding and design can be incorporated into robots for HRI and increase people's comfort levels and willingness to interact and work with them.

3.2 Research Contributions

This thesis attempts to advance the state of the science regarding HRI and robot failure. The contributions of this work include:

- New knowledge on how people react to failure in a real-life study with actual personal risk.
- New knowledge on how such failures impact future human interactions with the robot.

- New experiment methodology to investigate risk and trust for other researchers.
- Insight on how primacy and recency effects are impacted by personal risk.
- Insight on how a robot's face design influences people's interpretation of failure.

3.3 Thesis Organization

These contributions came about by designing a study that would investigate robot errors and their effect in trust and interactions with people. This thesis describes our efforts in how we went about to accomplish this task.

The next section gives an overview of the background and some related research in the domain of human-robot failures and trust. We start by describing some of the existing literature regarding risk in different automation systems and experimental scenarios. Then we explore trust and different factors about how it is different between humans and robots and why it is so important to understand it for the success of future technology. Lastly we talk about failure, how it is perceived by participants and what factors could have an effect on people. We decided to create a grocery store experiment since it is an everyday task that participants would believe we as researchers were trying to automate. Since participants are very familiar with the task, they would have some expectations, thus failures would be more conspicuous.

In order to be able to create this experiment, we had to choose a reliable system that could successfully pack some grocery items and also have some failures but would still not be a threat to participants' actual safety. Thus, we explain why we chose a Baxter robot. We then present details about how the technical aspects of the system were achieved. In order to be able to control a robot, we need to understand the relationship among the actuators that can be controlled by the robot and the resulting position in the environment, thus we talk about forward and inverse kinematics. Then we present details about the factors we wanted to explore in our study: failure severity (personal risk and property harm), order in which the most severe failure occurred (whether at the beginning or at the end of the experiment), and finally social signals of the robot, to observe whether putting a face on the robot had an impact in people's perception and willingness to assist the robot. We then describe the

experiment scenario and test conditions.

The latter portion of this thesis describes the results gathered from the participants' responses to surveys and observations during the experiment. In the closing portion section of this thesis, we discuss the results, limitations, and areas that we think are worth exploring in the future.

4 Background and Related Work

Just as humans are imperfect, even the best of robots will eventually fail at performing a task. As robots expand their roles in our lives, the likelihood for failure also increases. Merriam-Webster's dictionary defines failure as "a state of inability to perform a normal function" or a "lack of success" [28]. For a problem that is so common in robotics, failure is not widely explored in the literature. In HRI, people have been investigating safety and trust around social robots. However, to our knowledge, there has been little research investigating people's tolerance to said failures, especially when these failures may cause some physical harm. Consider the following scenarios: an industrial robot goes out of control near other workers; a domestic robot assistant accidentally breaks a valued possession; an autonomous robotic car malfunctions and causes an accident; or a military robot mistakenly kills civilians. All aforementioned scenarios are plausible situations that could unintentionally occur. Therefore, it is important for people to have a properly calibrated level of trust around robots that could expose them to physical harm.

4.1 Risk

Prior work by Robinette and colleagues [56] examined human-robot trust in high-risk situations that may engage fight-or-flight responses and other cognitive faculties that could impact a person's trust in unpredictable ways. The robot was available to help guide the person in a simulated emergency situation and periodically made errors. Their study gave preliminary evidence that robots interacting with humans in dangerous situations must either work perfectly at all times or clearly indicate when they are malfunctioning. It was alarming that all of their participants followed instructions by the robot, and were willing to forgive or ignore robot malfunctions, even minutes after these malfunctions had occurred [56]. Comparatively, in our work the robot is also a potential source of harm in the experimental scenario.

Adubor et al. [7] developed an experiment that found severity of failure seems to be tightly coupled with perceived risk to self rather than risk to the robot's task and object. Participants were shown a video clip of Baxter failing to place an object into a receptacle

and items falling towards the floor. Participants placed falling drinking glasses above laptop when rating the severity of the failure. The humans proximity to breaking glass was identified as an important factor.

Financial risk has also been explored in literature [39, 46, 55]. This is done by having the compensation be based in part on the overall performance. Desai et al. [49] created a sense of risk by tying robot performance to the ability to achieve a milestone payment. These types of financial risk are used as an incentive during experiments; however, they do not compromise the feelings of safety in a participant.

Because it is difficult to simulate risk scenarios convincingly, little is known about the relationship between robot-inflicted personal physical risk, property harm, and robotic failure. Risk is a very widely used and debated term because its definition tends to vary across circumstances. Sheridan (2008) defined risk as the product of the probability of an event and the consequences that accompany that event [58]. Weber et al. [27] further described that the events are domain-specific, and that the level of risk that is attributed with each event is dependent on the type of situation that is involved. For the purposes of this study, high risk was operationally defined as a situation in which there is an invasion of personal space and thus menace imposed by the robot's actions while low risk does not involve items leaving the task space.

Automation is the execution of a function by a machine agent that was previously done by a human. Relying on automation for services requires trust and taking risks. Risk can be perceived analytically or experientially; in other words, using logic and reasoning or feelings, instincts, and intuition, respectively [60]. The latter is accredited for being the primary influence for motivating people's behaviors due to its faster and easier decision-making methods for assessing danger [15]. As noted by Brooks, this has important implications with respect to robots experiencing failures. "If the experience of a failure results in a perceived increase of risk either from using the robot or being in its presence, people will also infer a lower benefit of using the system. On the other hand, if the perception of risk can be suppressed or mitigated in the event of failures, the inferred benefits of using the system should remain high"[15].

4.2 Trust

Design errors, maintenance problems and unanticipated variability make completely reliable and trustworthy automation unachievable; therefore, creating highly trustable automation is essential [40]. Trust can be characterized in several different ways, especially with regard to automation. By examining the differences and common themes of these definitions, we can have a better understanding of trust in human-robot interactions. An in-depth analysis is provided by Lee and See [40], some of which we will explain in some detail in this section. A common theme in trust is the user's attitudes or expectations. When a user exhibits trust, they have an expectation for a high likelihood of favorable behaviors or outcomes [11, 54]. Another common approach to define trust is as an intention to behave in a certain manner or willingness to rely on something or someone and be in a vulnerable state [19, 29, 34]. One of the most cited and widely used definitions is by Mayer and colleagues [29], "Trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." Others define trust as a behavioral result or state of vulnerability or risk, arising from uncertainty regarding the motives, intentions, or actions from the individual or system upon whom they depend [37]. In this sense, trust can be considered as a belief, attitude, intention, or behavior. Consistent with these previous descriptions, Ajzen and Fishbein [8] provided a framework that concluded trust is an attitude and reliance a behavior, belief, or intention. Thus, "trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [40]. Based on these trust definitions, we investigate how trust is affected when an autonomous robot is assigned a task that will help achieve an individual's goal and creates an expectation by succeeding several times before a failure occurs, we also expose the participant to vulnerability imposed by the robot's unpredictable actions.

Because robots are becoming more present in our everyday lives, it is important to understand trust between them and humans. Many studies have highlighted the importance of trust in actual work environments as opposed to controlled laboratory settings, including

for autonomous cars, maritime navigation systems, and autopilots [12, 31, 42, 43]. Individuals often report having trust in robots, but existing research leads us to believe that this statement-action is not always perfect. Therefore, our study investigates both an evaluation of trust with surveys and through participants' actions when trust is lost in an in-person study. Brooks [15] mentioned that one of the limitations of his work is the use of a third-person perspective of both a hypothetical scenario and questions to reduce the effects of subconsciously biased responses such as from people trying to portray themselves in a particular manner. However, reading about a hypothetical situation someone else is experiencing is not the same as experiencing the same situation for oneself in real life [15]. Ergo, a study in which participants experience failures in person is needed to verify their results.

Human-robot trust is different than interpersonal trust, in part because robots do not have human mental states like intentions. Researchers have found that even though robots lack intentionality, people may attribute intentionality to the designers or likewise attribute intentionality to the systems as the robots become increasingly sophisticated and take on human-characteristics such as gaze cues, speech communication, facial expressions, and naturalistic motions [53, 13, 20, 10]. Some studies have found that trust between people rapidly builds with face-to-face communication but not text-only communication [35, 50]; thus, Cassell and Bickmore [62] suggested that interacting with a computer that is a conversational partner will be perceived as more trustworthy since it would provide similar cues that people use in face-to-face conversations. Another difference between human-human trust and trust in robots is the social exchange relationship between interactions where a person is aware of their counterpart's behaviors and intents. Nevertheless, the inadequacy of the robot's social awareness leads to differences in trust relationships, such as a lack of collaboration, because task allocation benefits from people's assessments of how others perceive them [61].

How trust is attributed also differs between person-person and human-robot relationships. Muir argues the latter is developed based on faith, dependability, and predictability, in that order, whereas person-person trust follows the reverse order [47]. As the aforementioned reasons suggest, there are many differences between interpersonal trust and human-robot trust. Therefore, it is necessary for research to explore the unique elements in trust that differ between these relationships to gather a better understanding of reliance on automation. Our

study extends this prior work by evaluating the impact of a face on people's perception of intentionality and trust in the interaction. In addition, our robot does not respond to user's behaviors or communication attempts, so this study furthers the knowledge of how the lack of social awareness impacts the trust relationship. Finally we use Muir's trust questions [47] to evaluate participants' feelings about the dependability and predictability of the robot.

Of importance in the reliance on automation is the idea that if systems are not trusted, they will not be used; if they are not used, there is limited information regarding its capabilities; then, trust will not ever grow. Because trust is based on observation of the behavior of the automated system, automation must be relied upon for trust to grow in most cases [48]. It was posited that "As computer technology grows more pervasive, trust is also likely to become a critical factor in consumer products, such as home automation, personal robots, and automotive automation. Designing trustable technology may be a critical factor in the success of the next generation of automation and computer technology" [40].

4.3 Failure

Robot performance is most strongly associated with trust [38]. Brooks, explores human-robot interactions involving autonomous robotic service failures, and the way people react to varying conditions surrounding these failures. He takes a human-centric approach and focuses on people's reactions to failures, expectations of the robot, and goals. His work focuses on understanding people's reactions, increasing their situational awareness around autonomous robots, and creating communication platforms between robots and people. In order to understand people's reactions to failure, Brooks, developed a study that gathered data through Amazon's Mechanical Turk that manipulated context risk, failure severity, task support and human support through a short two part story about a fictional character Chris and his previous positive interaction between either a vacuum cleaner robot or self-driving taxi, and then a recent encounter with the robot and the results of the interaction. As a result to the study it was found that participants' REACTION, which is the factor obtained through exploratory factor analysis to find the weights that combine all the variables (ie. satisfied, trust, reliable, dependable, competent, disappointed, risky) was influenced by the task, context risk, and severity or type of failure. One of the limitations of the study was

lack of mimicry of a real situation in the hypothetical scenario; thus, a laboratory experiment in which participants experience failures in-person was needed to verify the results [15]. Provided this insight, our study similarly aims to explore people's perceived feelings of trust, reliability, dependability, competence, and safety, to different failure types in an in-person scenario, where a robot's failures are observed first-hand.

The effect of faults on trust do not occur instantaneously; faults cause trust to decline over time. Likewise, the recovery after faults occurs over time [39]. Muir and Moray [48] argue that trust is based mostly on the extent to which the machine is perceived to perform its function properly, suggesting a machine's performance strongly affects trust. Although the magnitude of an error is an important factor regarding the loss of trust, several small errors seem to have a more severe and long-lasting impact on trust than a single large error [22]. In contrast, however, previous work in HRI has found that errors occasionally performed by a humanoid robot can actually increase its perceived human-likeness and likability [23].

Not only do robot's levels of anthropomorphism may lead to different degrees of "forgiveness" in human-interaction partners when errors are displayed, but also the types of errors made by the robot (ie. "expected", "acceptable", or "intentional")[57]. Similarly, Bisantz and Seong [14] showed that failures attributable to different causes, such as intentional sabotage versus hardware or software failures, have different effects on trust. Thus, for some systems, it may be useful to discriminate between perturbations driven by wrong intentions or accidents. In some cases, having less faith in a robot can allow people to adjust their trust towards the system [63]. These results suggest that trust is more than a simple reflection of the performance of the automation; appropriate trust depends on the operators' understanding of how the context affects the capability of the automation. Since intention can have an effect in the way the user sees trust, the way the robot communicates a failure is also important.

Some research explored the effects of presenting human-like behavior patterns and human-specific features (ie. speech, gaze, gestures) into robot design. These aspects are important to achieve natural, effective communication and cooperation, and make robots appear more expressive and intelligible in order to improve social interactions between them and humans. Salem et al. found that when a robot uses co-verbal gestures during interaction, it

was anthropomorphized more and as a result participants perceived it as more likeable [44]. However, to their surprise, they also found that the robot exhibiting random gaze and gesture behavior, incongruent with its speech, was perceived almost as likeable as the robot with congruent behavior. Likewise, Ragni and his colleagues, found that a robot with human-like reasoning behavior and occasional errors may be perceived positively, ie. as more emphatic, as opposed to a robot with flawless, machine-like condition [45].

Some research has focused on how the perception of erroneous robot behaviors may influence human interaction choices and willingness to cooperate with the robot. Salem et al. [57], hypothesized that participant's assessment of the robot could be determined by the robot's behavior of its performance. Their study found that even though flaws and erroneous behavior's of a robot influenced participant's ratings regarding reliability, technical competence, understandibility and trustworthiness; participant's willingness to comply with instructions was not affected. Since the study found that the choice of experimental tasks can indeed lead to different results, our study aims to explore participant's willingness to assist even after they have been exposed to conditions where they experience personal risk and property destruction.

Another factor that could influence human-robotic trust with respect to failure is the timing at which the error occurs. Desai et al. [49] found that the timing of the reliability decreases would influence trust in the robot. They found that if an error occurred at the end of a run, participants' rating of trust decreased compared to when the error happened at the beginning of the run. Further investigation observed the real-time changes of trust that would not be affected by a participant's bias to primacy-recency effects. They found that low reliability earlier in the interaction had more detrimental impact overall trust than periods of low reliability later in the interaction [25]. Inspired by these works, we decided to investigate the effects on trust when the magnitude of the failure was greater at the beginning of the run compared to the end of the run.

Since it is evident that robots that operate in the real world will eventually fail, work has also been done to investigate how to gracefully mitigate failure. A study performed by Lee et al. [41] tested the effect of recovery strategies such as apologies, compensation, and options for the user in reducing the negative consequences of breakdowns. They used two service

robots, a human-like robot and a machine-like robot, which can be observed in Figure 4.1.

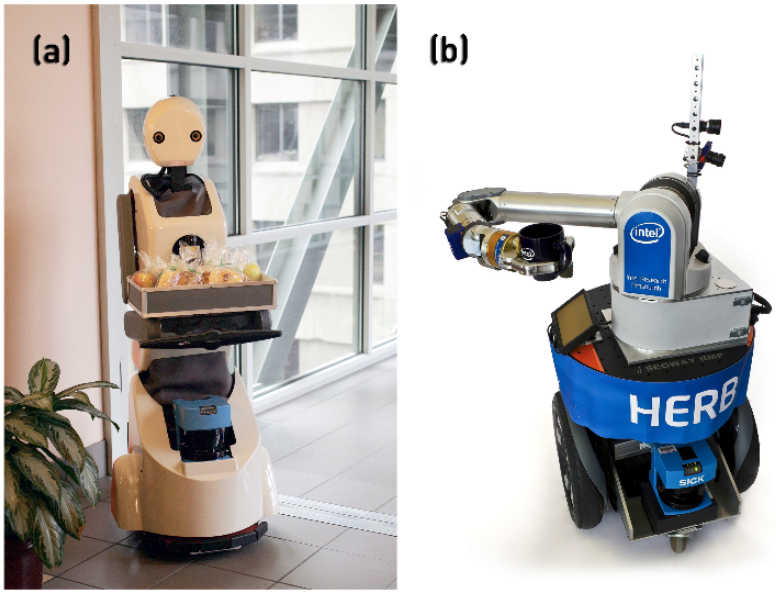

Figure 4.1: A. Human-like B. Machine-like Snackbot [41]

Overall it was found that the expectancy-setting strategy and recovery strategies were effective in mitigating the negative impact of a robot's service error on participant's impressions of a robotic service regardless of the human-likeness of the robot. It was found that the expectancy-setting strategy was particularly effective in extenuating the negative ratings on evaluation of the robot, and somewhat effective on improving participant's judgment of the quality of the service of the system. The result implied that apologies and options for the user were more effective in increasing the willingness to use the service again but the compensation strategy was more effective for one-time interactions since it increased participants' satisfaction with the robot. One possible design direction that was suggested was that in building robotic helpers, if they exhibit speech disfluencies they could be perceived less controlling without distracting from its perceived expertise.

Nevertheless their study had some limitations such as the use of a hypothetical scenario, so the robot was not actually affecting the participants in real-life. People's responses to robotic services in real environments might not be the same. Their study did not test the effect of higher risk failure and the implications on participant's trust in the robot. Lastly their study only tested people's reactions to one type of failure only and one type of task,

thus having different types of failure could have an effect in people's perception of the robot. While we acknowledge the importance of the effect that mitigation of errors could bring in failures, our study aims to explore how do the failures impact participant's trust and willingness to assist the robot, since most robots nowadays do not include these mitigation strategies.

An in-depth literature review by Honig [32] was done to explore when people perceive and resolve robot failures, how robots communicate failure, how failures influence people's perceptions and feelings toward robots, and how these effects can be mitigated. The different types of failure that were identified can be observed in Figure 4.2. Fifty-two identified studies relating communication of failures and their causes, the influence of failures on human-robot interaction and mitigation failures were explored and can be observed in Figure 4.3. Several gaps in the literature were evident as a result of the evaluation, such as studies focusing on human errors, robots communicating failures, or the cognitive, psychological, and social determinants that impact the design of mitigation strategies. Thus, our study fills in some of the gaps, such as understanding how people calibrate their trust around robots that could expose them or property to physical harm and human's tolerance to these types of failures.

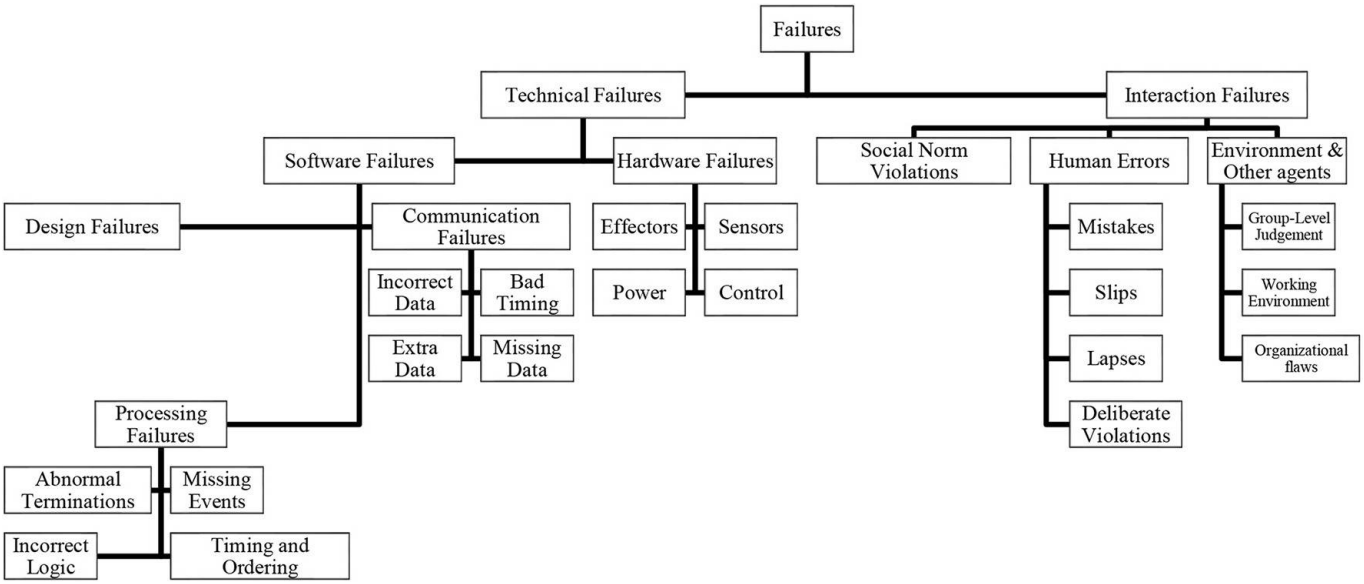

Figure 4.2: Types of Failure [32]

Figure 4.3: Literature Review on User-Centered Failure Handling [32]

5 System

5.1 About Baxter

The Baxter Research Robot created by Rethink Robotics [1] is a 16 degree-of-freedom anthropomorphic humanoid. It is 185cm in height and weighs 139kg, and it includes a stationary pedestal, torso, a 2 DOF head, a vision system, accelerometers, range-finding sensors, a robot control system, a safety system, an optional gravity-offload controller, and a collision detection routine [33]. Baxter features two 7 DOF arms that provide kinematic redundancy, each with a maximum reach of 121cm and Series Elastic Actuators (SEA) at each joint that are key to making the robot safe, incorporating full position and force sensing. SEA consists of having springs that are deformable by human level inputs between the motor/gearing elements and the output of the actuator, resulting in stable and low noise force control. Baxter allows direct programming access to the system via a standard open-source robotics operating system application programming interface, ROS API. When there are sudden changes in torque in the Baxter robot's joints, as would occur in the case of a collision with a human or an external object, the robot comes to a stop for two seconds before attempting to move again [3], making it completely safe around humans.

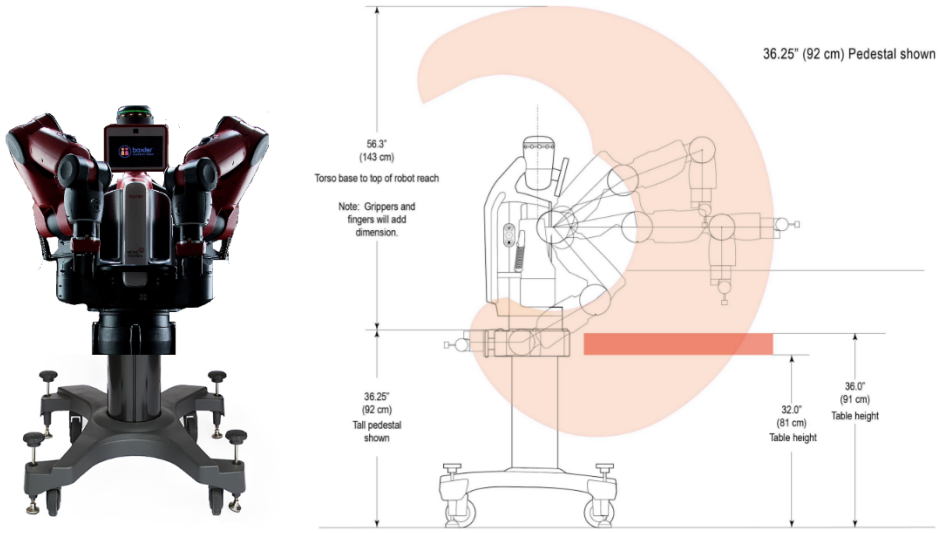

Figure 5.1: Baxter Research Robot [3]

[1] https://www.rethinkrobotics.com/

5.2 Study System Design

The Microsoft Kinect located above Baxter's head is used to detect April Tags [2] that are placed on the top surfaces of the objects. Then, the poses that are found through the Kinect are transformed to the robot's frame of reference. The detection of the tags is dynamic and refreshed every five seconds such that the location of the items is updated in the case of movement. The tags of the objects not only can be used to extract the exact location of the object, but also as a way to detect which objects are on the table. Each tag-object pair was assigned to either a successful task completion or a specific failure behavior. The code is designed such that the robot always picks up the successful cases and leave the failures until the end.

Once the pose of an item is found in the robot's frame, Baxter uses the voice of Joanna from Amazon's Polly Text-to-Speech service [1] to announce the name and price of the object. At the beginning of each experiment, the gripper was calibrated to ensure a reliable open-and-close motion. Then, Baxter's inverse kinematic (IK) service was used to compute some valid joint angle values of the robot's right arm after receiving an item's pose in order to send a command to the robot to execute the action. The implementation of these actions is described later in 5.2.2 Implementation.

In order to determine the end-effector's pose at each waypoint or at the end of the trajectory it would execute, forward kinematics was used. Then Baxter used the IK service provided by Rethink Robotics to compute the joint angles of the right arm to reach the items located at the table. For the sake of simplicity, only Baxter's right arm was used. This enabled participants to focus more on the task at hand and reduced the chance for non-intended failure. Baxter's home position was the default untucked arm position; which it returned to after each executed trajectory, it returned to this home position before starting another trajectory. Baxter had to return to a home position because having its IK service start from a random end-effector's pose could cause a different result in the IK trajectory, so it always started from the same point to ensure repeatability of the experiment.

For the gripper with the grasped item to avoid object collisions with other items or the bag, some of the waypoints in the trajectory were adjusted such that 15cm were added to

the pose of the object and the bag in the z-direction, such that the arm could go directly downward using the IR sensor. We used Baxter's IR sensor to determine the distance between the gripper and the object to ensure a tight and safe grip. Then, the gripper would close and go 15cm upward before continuing the course of its action. The 15cm distance was chosen because the tallest object was about 12cm, so any motion would occur above the surrounding objects. We could detect the gripper's position after it attempted to grab an item; if at this point the position of the gripper was less than a specified threshold (in other words, it has fully closed), it meant that the Baxter failed to retrieve the object and should attempt again. In the cases with Baxter's Head Display condition, this triggered the Baxter to change its happy expression to look surprised and then sad, and it would attempt to retrieve the object again instead of executing the full trajectory. The face also changed to an angry red face when the Baxter detected the tag of the extreme failure cases, *Throwing* and *Erratic Movements* (described later).

5.2.1 Camera Transformations

The April Tag visual fiducial system calculates the exact 3D pose for each tag in the 2D RGB image from the Kinect camera. However, the pose obtained is in the frame relative to the camera ([0,0,0] is at the camera). Because Baxter needs to receive the pose of the object in the base frame ([0,0,0] is at Baxter's torso), it was imperative to perform some transformations. We used the tf package [2] to perform the calculations of the transformations. We published a static coordinate transform to tf using an (x, y, z) offset in meters and (yaw, pitch, roll) in radians. The static coordinate transform from the base of our robot to the camera_link was found through manual calibrations to be (0.25, -0.04, 0.79, 0, 0.89, -0.05) in the previously mentioned order.

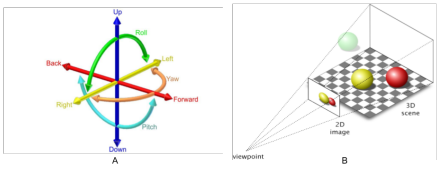

Figure 5.2: Transformations

[2] http://wiki.ros.org/tf

Using this information, the tf package could transform the pose received from the Kinect camera frame, camera_rgb_optical_frame, to the actual pose in the world where the April Tags were located. Equation 1 shows the transformation matrix for a given (x,y,z) April Tag pose [6]. The transformations required go in the following order: camera's color optical frame to the camera's color frame; the camera's color frame to the camera's base frame; the camera's base frame to the base frame of the robot; the base frame of the robot to the April Tags pose in the world. It is important to note that the orientation quaternions for the tags were modified such that they were (0,1,0,0) for x, y, z, and w, respectively, or represented as (0,$-\pi$, π) in Euler angles.

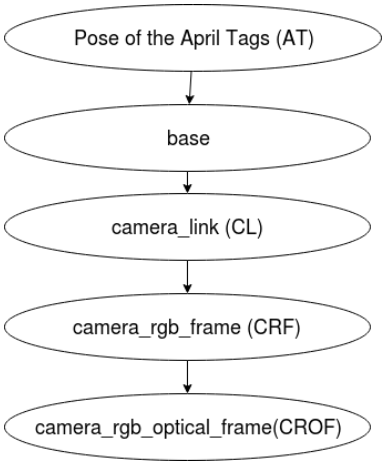

Figure 5.3: Camera to April Tags Transformation Diagram

$$[{}^{CROF}T_{AT}] = [{}^{AT}T_{base}]^{-1}[{}^{base}T_{CL}]^{-1}[{}^{CL}T_{CRF}]^{-1}[{}^{CRF}T_{CROF}]^{-1} \tag{1}$$

$$[{}^{AT}T_{base}]^{-1} = \begin{bmatrix} 0 & 0 & -1 & z \\ -1 & 0 & 0 & x \\ 0 & 1 & 0 & -y \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^{base}T_{CL}]^{-1} = \begin{bmatrix} -0.05 & 0.776 & -0.629 & 0.540 \\ 1 & 0.039 & -0.031 & -0.223 \\ 0 & -0.629 & -0.777 & 0.589 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[^{CL}T_{CRF}]^{-1} = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -0.045 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [^{CRF}T_{CROF}]^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[^{CROF}T_{AT}] = \begin{bmatrix} 0 & 0.629 & 0.777 & z - 0.624 \\ 0.05 & -0.776 & 0.629 & x - 0.568 \\ 1 & 0.039 & -0.031 & y - 0.222 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

5.2.2 Implementation

As previously mentioned, the April Tags served to identify the objects and were dynamically processed every five seconds. The tags were also related to the hard-coded names and prices of the items that the Text-to-Speech service would read, as can be observed in the diagram in Figure 5.4. The simplified RQT graphs of the camera and the Text-to-Speech can be found in 10 Appendix A. Four objects triggered the different failures while the other seven items would be success cases, as mentioned below in the descriptions of the different conditionals in the code:

- **Success** - In these cases, the robot would successfully deposit the grocery item in the bag. To avoid item and bag collisions, one of the waypoints in the trajectory added several centimeters to the the end-effector's target pose in the z-direction such that the robot's gripper could approach the target in a straight motion in the next waypoint. While the robot would approach the inside of the bag, it would not release the items gently, to evoke a sense of carelessness.

- **Test Trial - Assistance** - For this condition, the Baxter attempted to pick up the item a total of three times before successfully completing its trajectory. This was done by adding some centimeters to the tag's z-direction such that the Baxter's hand first would go just above the item and then release the item just after lifting it a couple of centimeters.

- **Low Property Harm - Crunch** - This condition was similar to Success, but due to the nature of the material of the bag of chips, when the robot would attempt to grasp it, it would make loud noises as if crunching the chips.

- **High Property Harm - Floor** - To throw all the grocery items to the floor, a waypoint trajectory was created in which the Baxter's end effector's pose would be just in front of the bag. The next waypoint was to go forward, which looked like a pushing motion and caused the grocery bag to be knocked to the floor.

- **Low Personal Risk - Erratic Movements** - In this condition we created a series of smooth movements to wave a cereal box close to the participant. Because Baxter's IK service is slow to ensure precision to get to the end-effector's specified location, a timer was used such that every 3 seconds the end-effector's end pose would change in order to make a series of smooth motions that would get very close to the participant.

- **High Personal Risk - Throwing** - Baxter would throw a potato to the participant. This condition was created by the use of multithreading. One thread was used to perform a swinging trajectory and the other thread would signal to open the gripper after a triggered timer started at the beginning of the swinging motion.

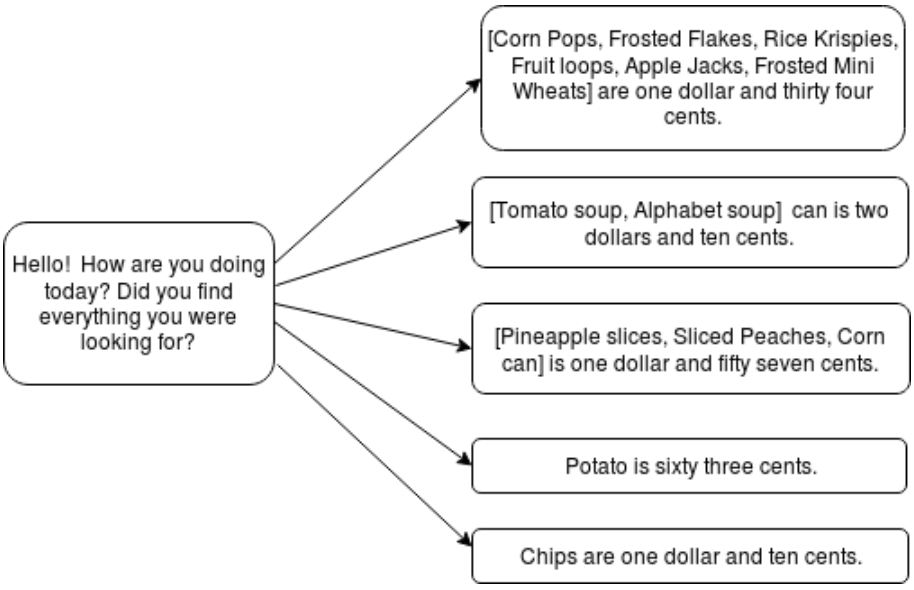

Figure 5.4: Baxter's Text-to-Speech Dialog

5.3 Baxter Kinematics

The first step in controlling a robot is to understand the mathematical model of the system. Motion planning requires an understanding of the relationships among the actuators that can be controlled by a robot and the resulting position in the environment. Kinematics involves two processes: forward and inverse kinematics. Forward kinematics of a serial manipulator is a very well-established concept in robotics. Inverse Kinematics is more intricate because there are different mechanisms to derive the inverse kinematics equations of a manipulator. Baxter's IK service was used for programming the robot; however, in order to understand the motions, we explored the Jacobian pseudoinverse iterative IK technique. To explain the forward kinematics, we referred to Williams' Baxter Humanoid Kinematics [33].

5.3.1 Forward Kinematics

In this study, we used forward kinematics to determine some of the motions for the Baxter robot's right arm. Forward kinematics refers to computing the pose and orientation of the robot's manipulator end-effector given the joint values. This can be computed through a series of homogeneous transformation equations that are used to find the pose of the end-effector with respect to the base reference frame. The modified convention of Denavit-Hartenberg (DH) are used to select frames of reference because it is a widely known notation to describe the kinematic model of a robot. According to this convention, each link is represented by two parameters: the link length, a_i, and link twist, α_i, which define the relative location of the two attached joint axes in space. Joints are also described by two parameters: the link offset, d_i, which represents the distance from one link to another along the axis of the joint, and the joint angle, θ_i, which is the rotation of one link with respect to the next about the joint axis [21, 64].

The modified DH representation results in a link transform matrix that transforms the link coordinate frame i-1 to frame i of the form:

$$^{i-1}A_i(\theta_i, d_i, a_i, \alpha_i) =) R_x(\alpha_i)T_x(a_i)R_z(\theta_i)T_z(d_i) \tag{2}$$

where R_k and T_k denote the rotation and translation about axis k, respectively. The overall transform can be expressed in terms of the individual link transforms:

$$ {}^{0}T_n = {}^{0}A_1 \ {}^{1}A_2 ..... {}^{n-1}A_n \tag{3} $$

The Cartesian reference frame definitions for Baxter's 7-DOF right arm are shown in Figure 5.5 and its DH parameters can be found in Table 2. These can be used to get the pose of each frame {i} with respect to its nearest neighbor frame {i-1} [33].

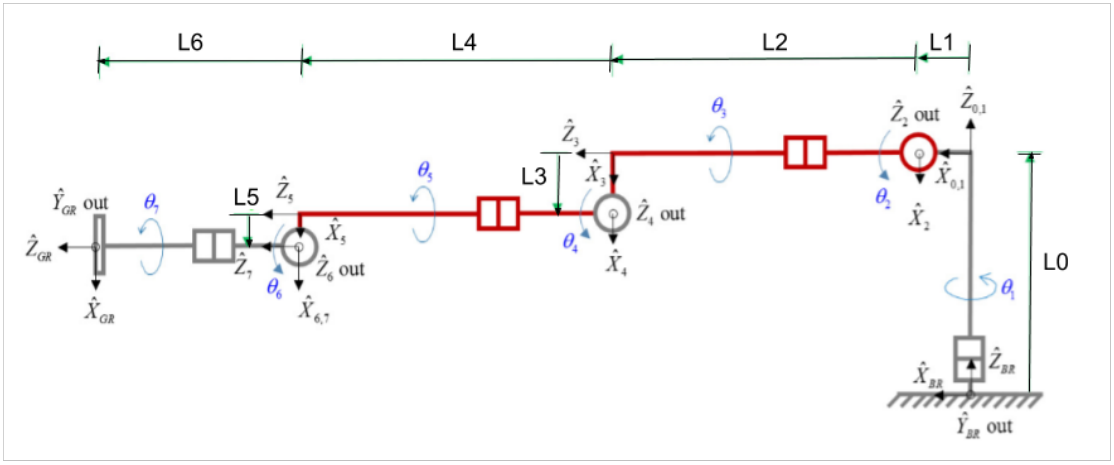

Figure 5.5: Seven DOF Right Arm Kinematic Diagram with Coordinate Frames [33]

Length	Value (mm)
L	278
h	64
H	1104

Table 1: Baxter's base to world lengths

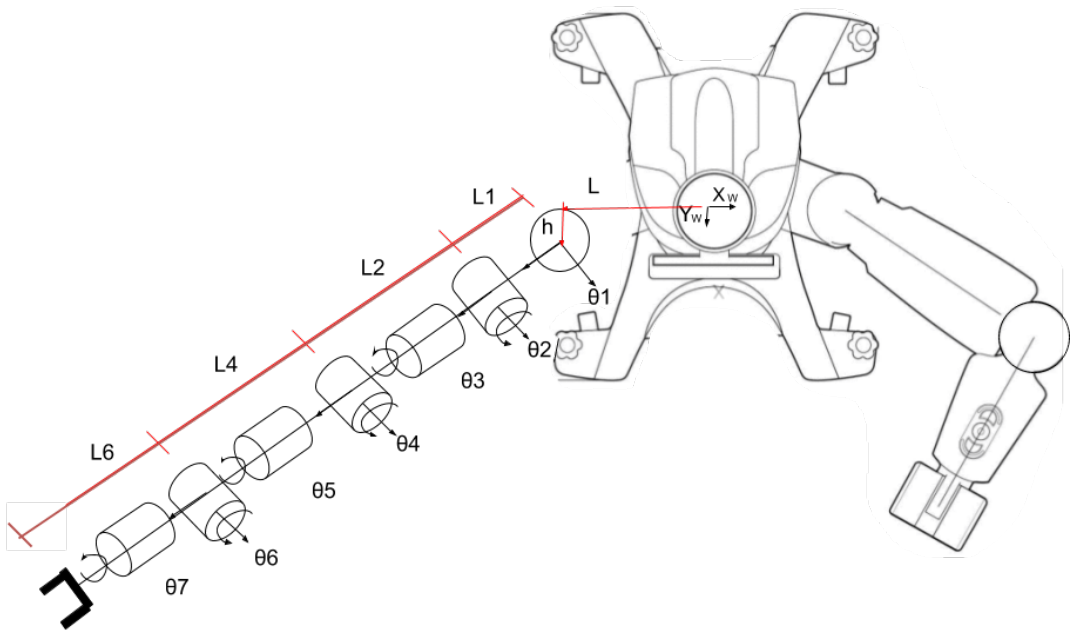

Figure 5.6: Top View Zero Joint Angles, Baxter Right-Arm Kinematic Diagram (Modified diagram from [33])

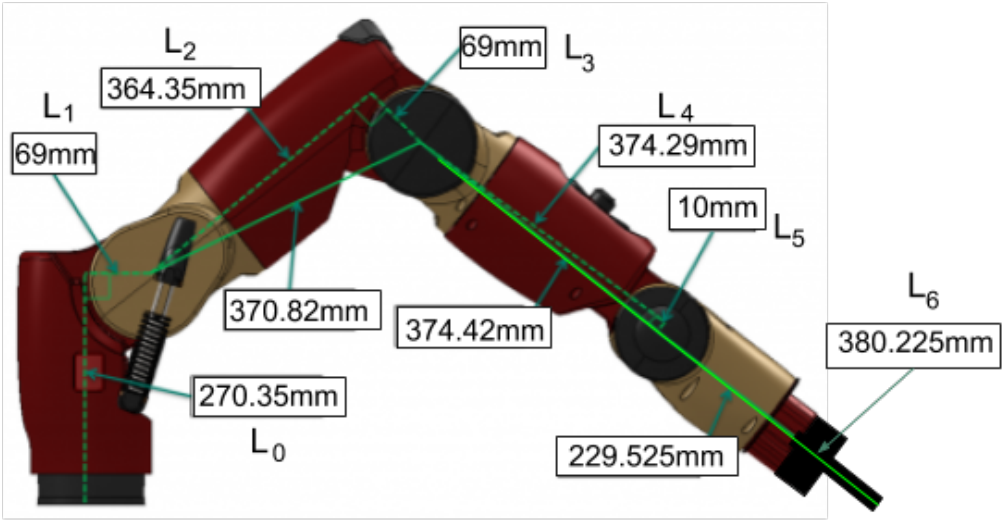

Figure 5.7: Baxter's Joint Lengths (Modified diagram from [5])

i	α_i	a_i (m)	d_i (m)	θ_i
1	0	0	0	θ_1
2	-90°	L_1	0	θ_2+ 90°
3	90°	0	L_2	θ_3
4	-90°	L_3	0	θ_4
5	90°	0	L_4	θ_5
6	-90°	L_5	0	θ_6
7	90	0	0	θ_7

Table 2: Seven DOF Right Arm DH Parameters

$$ {}^{n-1}T_n = \begin{bmatrix} & [R] & & [T] \\ & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4} $$

$$ {}^{n-1}T_n = \begin{bmatrix} cos\theta_n & -sin\theta_n & 0 & a_{n-1} \\ sin\theta_n cos\alpha_{n-1} & cos\theta_n cos\alpha_{n-1} & -sin\alpha_{n-1} & -d_n sin\alpha_{n-1} \\ sin\theta_n sin\alpha_{n-1} & cos\theta_n sin\alpha_{n-1} & cos\alpha_{n-1} & d_n cos\alpha_{n-1} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5} $$

The overall forward pose kinematics can be computed by multiplying together all seven neighboring homogeneous transformation matrices as a function of the joint angle. By doing so, the 3x3 upper left rotation matrix and 3x1 position vector can be obtained for the robot's orientation and pose at any given joint values. The range values for each of the robot's joints can be found in Table 3. The resulting orthonormal rotation matrix elements and translational terms can be found symbolically evaluated in 11 Appendix B.

$$ [{}^0T_7] = [{}^0T_1(\theta_1)][{}^1T_2(\theta_2)][{}^2T_3(\theta_3)][{}^3T_4(\theta_4)][{}^4T_5(\theta_5)][{}^5T_6(\theta_6)][{}^6T_7(\theta_7)] \tag{6} $$

$$ {}^{0}T_{7} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & {}^{0}x_{7} \\ r_{21} & r_{22} & r_{23} & {}^{0}y_{7} \\ r_{31} & r_{32} & r_{33} & {}^{0}z_{7} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7} $$

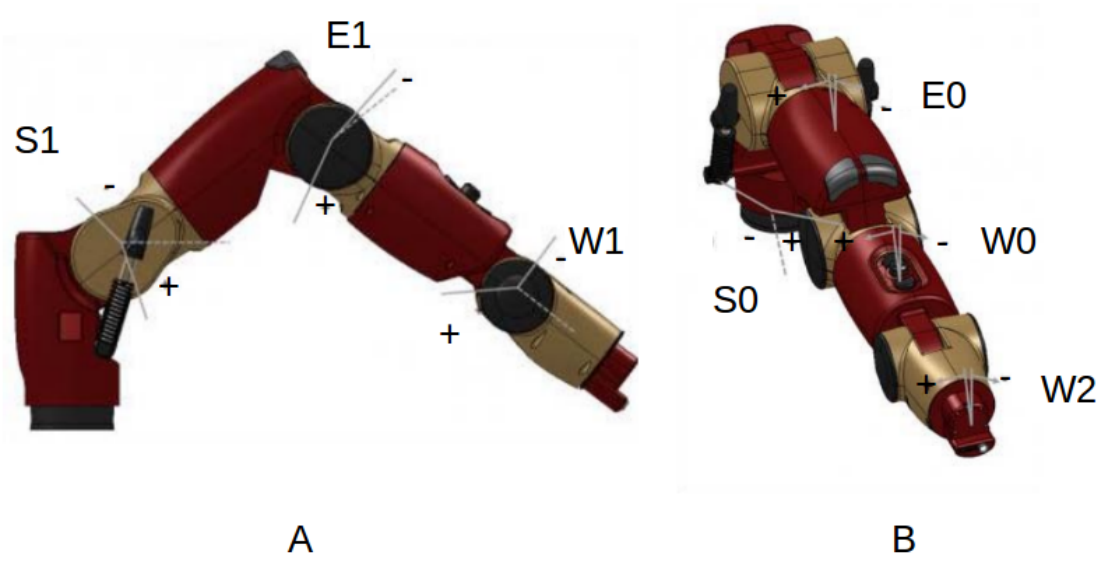

Figure 5.8: Baxter's Arm Joints (A.Bend Joints and B.Twist Joints

Joint	Joint Variable	θ_i (min)	θ_i (max)	Radians (min)	Radians (max)
S_0	θ_1	51°	-141°	0.890	-2.461
S_1	θ_2	60°	-123°	1.047	-2.147
E_0	θ_3	173.5°	-173.5°	3.028	-3.028
E_1	θ_4	150°	-3°	2.618	-0.052
W_0	θ_5	175.25°	-175.25°	3.059	-3.059
W_1	θ_6	120°	-90°	2.094	-1.571
W_2	θ_7	175.25°	-175.25°	3.059	-3.059

Table 3: Seven DOF Arm Joint Limits

The overall 7-DOF Baxter right-arm forward kinematics solution also requires a trans-

formation between the world and the right base coordinate frame, the right base coordinate frame and the joints, and the joints and the grippers.

$$[^{W}T_{GR}] = [^{W}T_{BR}][^{BR}T_{0}][^{0}T_{7}][^{7}T_{GR}] \tag{8}$$

5.3.2 Baxter Inverse Kinematics

The Baxter robot comes with a supported software development kit (SDK) provided by Rethink Robotics. This SDK includes an Inverse Kinematics (IK) service used by the robot and thus was used for the purposes of this study. However, Rethink Robotics does not openly release the software used to compute this IK service. Because most work on inverse kinematics of redundant robots focus on iterative numerical approaches, we explored the Jacobian pseudoinverse technique to evaluate the joint angles required to attain a required end-effector pose.

$$\dot{x} = J(\theta)\dot{\theta} \tag{9}$$

$$\dot{\theta} = J^{-1}(\theta)\dot{x} \tag{10}$$

The inverse velocity kinematics of a robot can be computed by using the pseudoinverse for non-square Jacobian matrices. After integrating the velocity kinematics obtained over several time-steps, the position kinematics can be computed. The pseudoinverse approach to iterative IK starts with taking the joint angle positions of the current configuration of the robot as the seed angles for integration over time. Using the Jacobian pseudoinverse, the joint angular velocity can be calculated by integrating over a constant time-step and comparing the final Cartesian pose with the goal pose until convergence [9]. One issue with the Jacobin pseudoinverse is that it does not work well when the arm manipulator has joint limits, which is true for the Baxter robot. To account for this, random restarts can be used: when the algorithm hits a joint limit, it randomly restarts the joint pose and attempts again.

An overview of the Jacobian pseudoinverse technique [17] is provided below:
Let $p_0 = p(\theta_0)$ be the initial position of the angles of the system and $p_1 = p(\theta_0 + \Delta\theta)$ be the goal position. The Jacobian inverse technique iteratively computes an estimate of $\Delta\theta$ such

that the error given by $||p_1 - p_0||$ is minimized. For small $\Delta\theta$ vectors, the series expansion of the position function actually depends on the Jacobian matrix of the position function at θ_0, such that $p_1 = p(\theta_0 + J_p(\theta_0)\Delta\theta)$.

The entries of the Jacobian (3x3) matrix can be determined numerically:

$$\frac{\partial p_i}{\partial \theta_j} = \frac{p_i(\theta_{0,j} + h) - p_i(\theta_0)}{h} \tag{11}$$

Where the i and j represent the (i,j)-th entry of the Jacobian.

To solve for $\Delta\theta$, for the purpose of incrementing the joint angles by it, we can rearrange the equation in the form of

$$\Delta\theta = J_p^*(\theta_0)(p(\theta_0 + \Delta\theta) - p(\theta_0)) \tag{12}$$

Where the J_p^* represents the Moore-Penrose pseudoinverse of the Jacobian, which is solved using singular value decomposition (SVD) if J_p has full rank (and thus guaranteed to be invertible), hence the pseudoinverse is given by:

$$J_p^* = [J_p]^T[[J_p][J_p]^T]^{-1} \tag{13}$$

The first iteration for computing $\Delta\theta$ results in an estimate of the desired $\Delta\theta$ vector. We use $\Delta\theta$ to adjust the joint angles until a sufficiently close solution is found.

$$\theta = \theta + \Delta\theta \tag{14}$$

The pseudo code for the algorithm can be found in Algorithm 1. It is important to note that because Baxter has joint limits, the joint angles should always be maintained inside of those ranges in order to get a valid solution. If the solution is invalid, a random initial joint pose can be chosen to compute the joint angles needed to achieve the goal position.

Algorithm 1 Inverse Kinematics

procedure SOLVE IK WITH PSEUDOINVERSE(input arg = $pose_{desire}$)
    $\theta = CurrentJointAngles()$
    $pose_{error} = CalculatePoseError(pose_{curr}, pose_{desire})$
    while $||pose_{error}||^2 > 1e^{-4}$ **do**
        $J_p^* = CalculateJacobianPseudoInverse(\theta)$
        $\triangle\theta = J_p^* \cdot pose_{error}$
        $\theta = \theta + \triangle\theta \cdot dt$
        $pose_{curr} = ForwardKinematics(\theta)$
        $pose_{error} = CalculatePoseError(pose_{curr}, pose_{desire})$
        if not $CheckJointLimits(\theta)$ **then**
            $\theta = GenerateRandomJointPose()$
    return θ

6 Method

6.1 Conditions

We designed a 2x2x2x2 between-subjects experiment to test different types of failures and their effects on participant willingness to assist the robot. The failures varied in their severity in two ways, including the degree to which they caused personal risk (throwing an item at a person vs. erratic robotic movements that come close to the participant's personal space) or destroyed the groceries (crunching an item vs. throwing the grocery bag to the floor). We also varied the order of the magnitude of the failures (ascending vs. descending severity), and whether Baxter's screen showed a face (display vs. blank screen). The experiment consisted of having people observe the Baxter robot bag groceries *the participant had acquired at the store*. In doing so, the robot would bag eight items successfully and three items would undergo different types of failures. Each participant was exposed to one of sixteen combinations of conditions. The types of failures and orders for these sixteen combinations are shown in Figure 6.1.

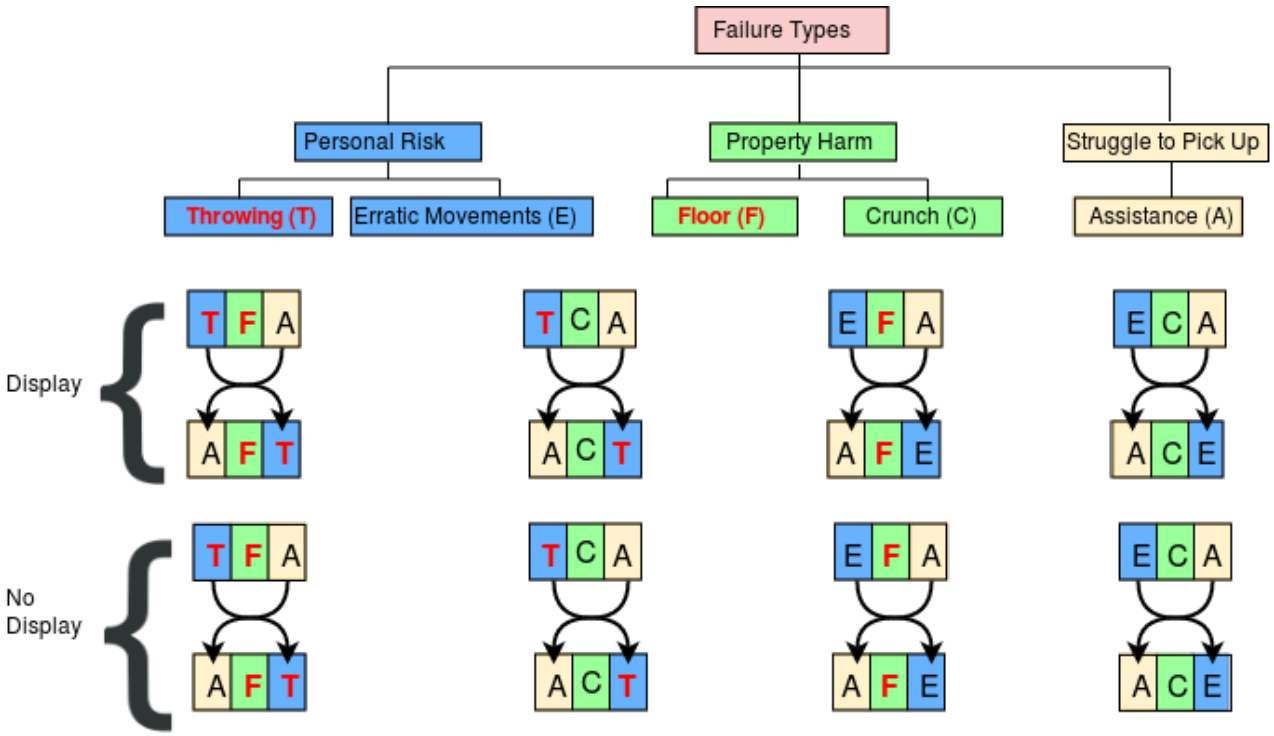

Figure 6.1: Experimental Conditions

6.1.1 Movement Conditions

Each of the conditions for the robot's movements are described below:

High Personal Risk - Throwing (T) - The Baxter robot would grab a foam potato of dimensions 6cm x 10cm and would throw it using a swinging motion from the table in the direction of the participant, finally releasing it in front of the participant. The action was designed to throw the potato over the participant's left shoulder.
Low Personal Risk - Erratic Movements (E) - The Baxter robot would pick up a small box of Frosted Mini-Wheats cereal and move its right arm in a series of three fast movements: going in the direction where the bag is located, changing directions towards the opposite side of the table, and finally raising its arm above its head and swinging it down to drop it in the middle of the table. During this path, the box would be waved very close to the participant.
High Property Harm - Floor (F) - The Baxter robot would pick up a plastic can of tomato sauce and look as though it was performing a trajectory to place it in the bag but instead pushed the bag off the table.
Low Property Harm - Crunch (C) - The Baxter robot would pick up a small bag of potato chips and, in doing so, crunched the chips.
Test Trial - Assistance (A) - The Baxter robot would attempt to pick up a small Rice Krispies cereal box. For the first attempt, the robot would close its gripper just above the box; in the second attempt, the robot would grab the box and elevate it 10 cm above the table and then drop it; and in the third attempt (assuming the participant had assisted putting the cereal box under the gripper), the robot would complete the trajectory of putting the item inside the bag.

The conditions were selected in order to investigate different types of failures. The *Personal Risk* condition had a threatening component because there was an indication of menace coming from the robot's actions. The two failure cases observed in this category had different severity levels, where high severity was throwing an object at a person (**T**) and low severity was an item coming very close to the participant but never leaving the robot's

gripper (**E**). The *Property Harm* category was chosen to show participants the capacity of the robot to damage the groceries. For these cases, different severity levels were also chosen, where a high severity meant pushing the entire grocery bag of items to the floor in front of the participant (**F**) and low severity meant just one object (the bag of chips) was crunched (**C**). Every participant was given the opportunity to assist the robot in the Assistance condition (**A**).

Figure 6.2: Objects Used in Failure Cases

6.1.2 Display Conditions

The Baxter robot's head display expressions were taken from a study performed by Fitter and Kuchenbecker [30]. The expressions that were used were happy, angry, surprised, and sad, as can be seen in Figure 6.3, and were designed using Ekman's Universal Facial Expressions [26] as a reference. During the successful trials, the robot displayed a happy expression; during the personal risk failures (T or E) the robot would display an angry expression; and during the assistance condition or at any other point the robot detected its gripper had fully closed when it was expected to retrieve an object but failed, its expression changed from happy to surprised and ultimately sad.

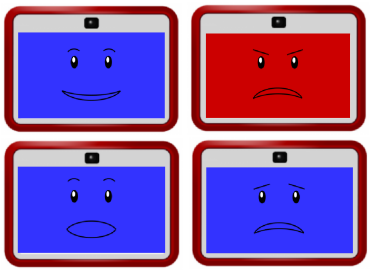

Figure 6.3: Baxter's Head Display [30]

Fitter and Kuchenbecker's study [30] assessed the emotional effects of expressive faces for the Baxter robot in human observers. It was observed that certain facial expressions and colors are associated with different valence and arousal levels. The study showed that raters felt significantly less safe looking at red and angry faces compared to all other colors and emotions; thus, red appears to be the main color that could be used to influence human responses. In this study, we wanted to make people feel uncomfortable with the robot performing failures, so we chose for it to have an angry red face during the cases where the robot is failing and potentially causing personal risk (i.e., throwing the potato or moving erratically). In order to create a compelling contrast between the robot's emotional states, a happy blue face was chosen for the successful trials. The color was chosen because blue is often viewed as a non-threatening color that calls to mind feelings of calmness, stability, reliability, and security [18, 36].

We chose to explore the presence of a display on the Baxter since, in many cases, a non-humanlike machine will make people act instinctively cautious around it because they are encountering an unknown and potentially dangerous situation for which they have few prior expectations. Thus, for "first encounters", or application areas where people will meet a particular robot only briefly, non-humanoid machines may have advantages over humanoid robots. Non-humanoid robots decrease the expectations in terms of the skills people attribute to them, and they may elicit cautious behavior in people who will carefully assess the robot's abilities and how one can safely interact with it, rather than assuming that it "naturally" has human-like abilities and is safe for interaction [24]. To investigate this issue, we used the presence or absence of the human-like face to examine effects on people's behavior and willingness to assist the robot.

6.1.3 Order of the Magnitude of the Failure

The failure cases observed could be shown in an ascending or descending order. The ascending order meant the failure cases got more severe as the experiment progressed. In other words, the first failure the participant observed was the *Assistance*, and the last failure they observed was the *Personal Risk* case. The descending order meant the failure cases got less severe as the experiment progressed. Since past work by Adubor et al. [7] has found that personal risk is more important then property harm in people's perception of risk, we place *Personal Risk* to be the most severe failure followed by *Property Harm*. Thus the first failure they observed was the *Personal Risk* failure case and the last was the *Assistance*. In Figure 6.1, the first line of the *Display* and *No Display* conditions were the descending order and the second lines were the ascending order.

By studying the order in which failure cases are shown, we can observe whether recency or primacy effects on memory affected participants' ratings of trust, safety, or willingness to work together or assist the robot. When remembering a number of items, people are more likely to remember those that occurred at the end, followed by those in the beginning [4], so the temporal position of extreme failures could affect participant's remembered experiences with the robot. We opted for these two orderings to observe if people would still be willing to assist the robot in the descending order after being exposed to the *Personal Risk* and *Property Harm* failure cases.

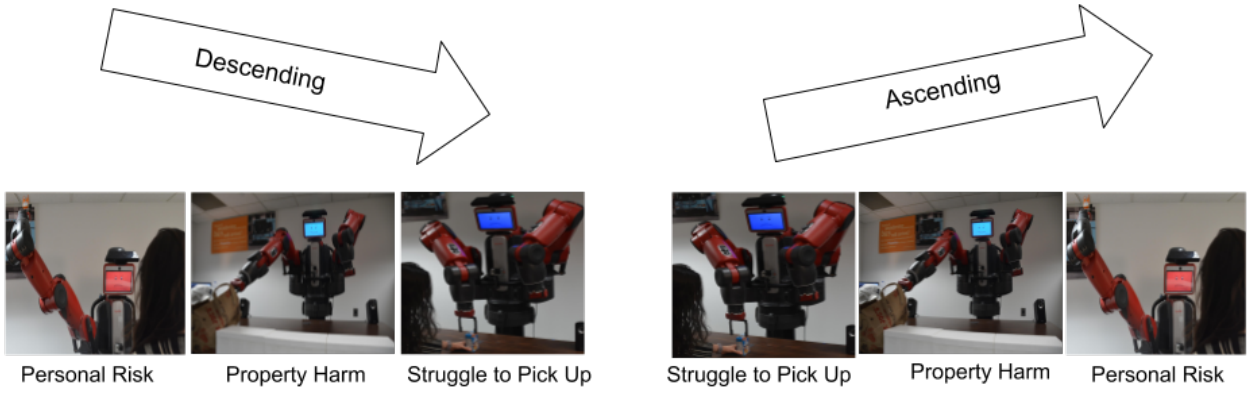

Figure 6.4: Descending and Ascending order conditions

6.2 Setup

6.2.1 Sensors and Other Added Equipment

Five sensors were used for the study, as shown in Figure 6.5: a Microsoft Kinect Sensor, a ZED camera, a Creative Inspire T12 Speaker System, a GoPro HERO3+ Silver Edition, and Baxter's right hand infra-red (IR) sensor. The Microsoft Kinect was used as a downward camera that could detect the objects because they were in the close vicinity of Baxter and could otherwise not be detected with the robot's head camera. The ZED camera was used to detect each participant's torso position and distance away from the robot. The Creative Inspire T12 Speaker System was used to play recordings of Baxter's voice greeting the participants and telling them each item's name and price. The GoPro HERO3+ was used to record the interaction. Finally, Baxter's right-hand IR sensor was used to detect the distance from the gripper to the object.

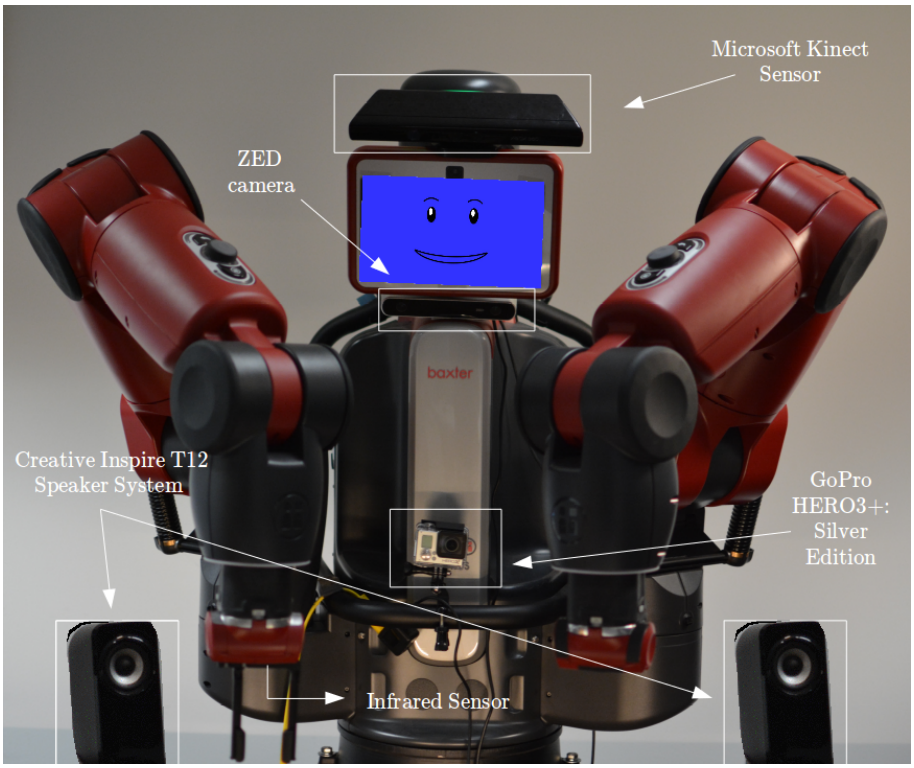

Figure 6.5: Additional sensors used in the study

6.2.2 Experimental Setup

The study was conducted in a room with a free space of 3.7 x 2.7 meters (12 x 9 feet). The free space was isolated from the rest of the room with the use of black curtains and large black poster boards. The robot was located on one side of a table, and the participant stood on the opposite side facing the robot, represented by an 'x' in the Figure 6.6. The Creative Inspire T12 Speakers were placed on both sides of the table to simulate the voice coming from the robot. Before the experiment began, six grocery items were placed on the table. The grocery bag was located in the closest left corner of the table (from the participant's perspective). During the study, the experimenter would start the program and then go to the other side the curtain to leave the participant alone with the robot.

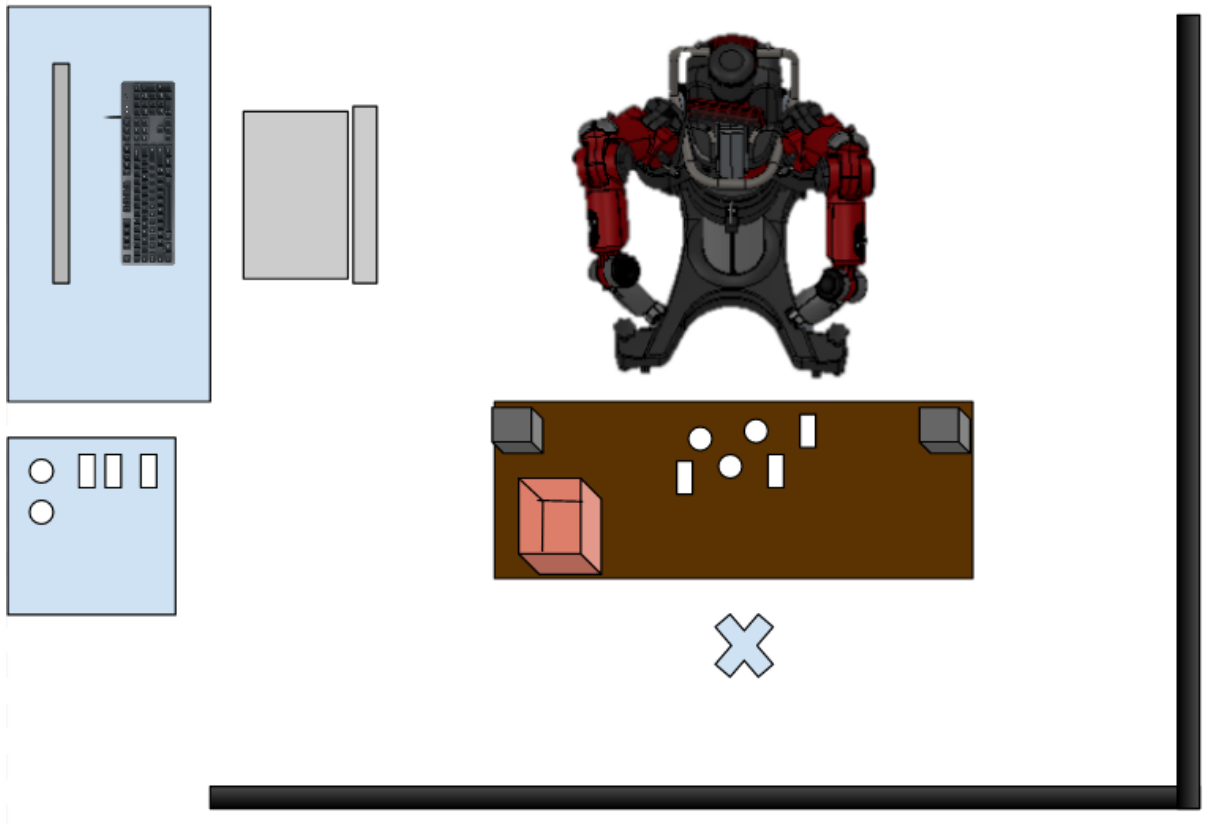

Figure 6.6: Experimental Setup

6.3 Participants

We recruited 64 participants (4 per 16 combinations of conditions) using a participant pool and word of mouth. The participants had to be at least 18 years of age, fluent in English, with normal or corrected-to-normal hearing and vision. Lastly, participants needed to be able to stand for at least 30 minutes and move their arms and hands freely.

Condition	**#P**	**#F**	**#M**	**#O**	**Age (Std Dev)**
D+TFA	4	3	1	0	36.8 (18.1)
D+EFA	4	1	3	0	29.3 (9.9)
D+TCA	4	2	2	0	27.0 (2.6)
D+ECA	4	1	3	0	38.0 (19.8)
D+AFT	4	3	1	0	21.0 (0.8)
D+AFE	4	2	2	0	23.8 (3.2)
D+ACT	4	2	2	0	26.0 (11.4)
D+ACE	4	2	2	0	25.3 (3.6)
B+TFA	4	3	1	0	20.3 (1.3)
B+EFA	4	1	3	0	33.5 (11.1)
B+TCA	4	1	3	0	21.5 (1.9)
B+ECA	4	2	2	0	21.8 (2.2)
B+AFT	4	3	1	0	30.3 (13.5)
B+AFE	4	2	1	1	31.5 (11.7)
B+ACT	4	3	0	1	22.0 (4.3)
B+ACE	4	4	0	0	24.3 (3.0)

Table 4: Participants per Condition. “P”,“F”,“M”,“O” are used to abbreviate participants, female, male, and other respectively.

Table 4 shows the details of the 64 participants that interacted with the robot. All participants were naive to the true nature of the study and were told that they would interact

with a robot in a grocery store setting. Participants were also randomly assigned to one of the 16 condition combinations. The average age of participants was 27 years old (SD = 10.0). Before the interaction, the participants indicated their familiarity with computers and robots on a 7-point Likert Scale (7 being the highest). They also indicated their willingness to work with the robot on a scale from 1 to 5, (5 being Strongly Agree). Most participants indicated familiarity with computers (M = 5.75, SD = 1.07) and familiarity with robots (M = 3.38, SD =1.50). Despite their unfamiliarity with robots, people indicated to be strongly willing to work with robots (M= 4.27, SD= 0.65). This research was approved by our Institutional Review Board, and participants received ten dollars as compensation for their time.

6.4 Procedure

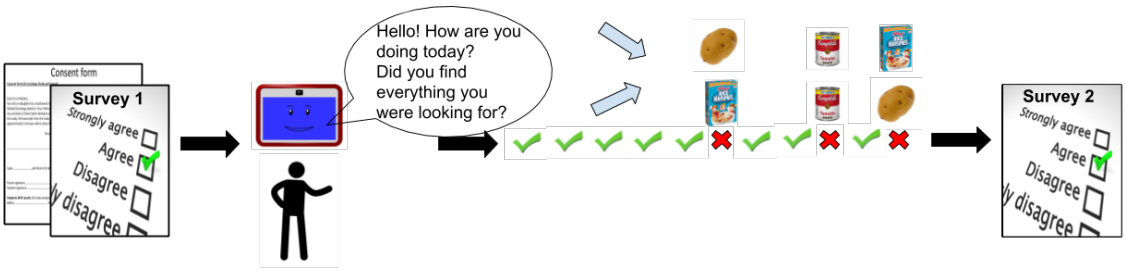

Figure 6.7: Procedure

First, the experimenter obtained informed consent and administered a preliminary survey for the participant. Then, the participant was escorted to stand in front of the robot while instructions were given. The experimenter introduced the study as investigating how robots would perform in a grocery store setting and the interactions they would have with humans there. Participants were asked to be patient with the robot because the study was a simulation and the robot was slower than normal. Participants were told that the grocery bag should remain along the left side of the table because the robot recognized the location of the bag to be in that general area. This explanation was provided to deter participants from moving the bag around when the robot was placing the groceries. They were also told that the robot would say the grocery item's name and price and go to where the item was

located and they could feel free to help the robot if the robot needed assistance with an item in any particular instance. As the experiment began, the experimenter would inform the participant that the experimenter would leave the task area and would not talk until the study was resumed in order to make the experience more realistic.

Once the robot and the participant were alone in the task area, the robot welcomed the participant and asked if he or she had found everything they were looking for. Then, the robot began by saying an item's name and price and began its trajectory to pick up the first item. The robot would perform five success cases and the sixth item was an allocated failure. Depending on the *Ascending* or *Descending* conditions, the participant would experience the *Struggle to Pick Up* or *Personal Risk* categories first, respectively. Once the failure occurred, the experimenter would return with three more items, the first two were successes and the last was a failure case. Depending on the condition, the second failure would either crunch a bag of chips (Crunch) or throw the grocery bag to the floor (Floor). Lastly, the experimenter would return and place two more objects on the table, where the first was a success case and the second was a failure case. Once again, depending on the *Ascending* or *Descending* conditions, the participant would experience the *Personal Risk* or *Struggle to Pick Up* category last, respectively. By the end of the experiment, the participants always experienced the robot bagging a total of 11 items, where three were failures and eight were successes. After the *Personal Risk* case, the experimenter would look at the computer with a perplexed expression to simulate not knowing the problem.

Finally, the experimenter administered a post-test survey, paid the participant, and debriefed them about the real intent of the study. During the debriefing, the experimenter explained that the study was not for the purpose of seeing the performance of a robot in a grocery store setting, but rather about trying to understand people's responses to robotic failure. The experimenter discussed the three main types of failures that were presented to them: personal risk, harm of groceries, and struggling to pick up the item to give an opportunity for participants to cooperate. The experimenter explained the order of the failures, whether they experienced personal risk (Descending) or the struggle to pick up the item (Ascending) first to see if the failure had an impact in human collaboration with the robot or if the last interaction they had with the robot had an impact in how they perceived it.

Participants were also told about the screen on the Baxter and discussed whether it displayed a face or a blank screen to see if this variable also had an impact on people's perception of the robot.

6.5 Hypotheses

Before we reveal what was found as a result of this study, we will talk about what what was hypothesized and the methods we went through to test the hypotheses:

H1: More extreme failures by the robot will result in decreases in the amount that participants trust the robot.

H2: Exposure to prior failures will result in decreases in participants' willingness to assist the robot.

H3: Failures that are more severe will result in decreases in participants' feelings of safety.

H4: The presence of an expressive face will improve participants' willingness to assist the robot.

H5: Experiencing extreme cases of failure towards the end of the session will result in lower participant ratings of performance, safety, and trust due to recency effects.

6.6 Measurements

Two surveys were used in this study and can be found in 12 Appendix C. The preliminary survey asked about demographics, familiarity with robots and computers, and contained 7 five-level Likert scale questions assessing people's impressions of robots and willingness to work with them. The post-experiment questionnaire was administered immediately after the third failure had occurred and the experiment ended. The first four questions were modified from Muir's Trust questionnaire [47], to assess people's feelings of the robot's predictability, dependability, trust, and faith in the system. The post-experiment questionnaire also included 22 five-point Likert scale questions assessing people's feelings and impressions of the robot and the interaction. In addition to the surveys, we used a ZED camera to measure the depth position of the participants' torso to test if it was possible to detect changes in participants' body language or proximity to the robot after the failures had occurred.

Variable	Statement/Question
Muir Trust Questions	Predictability - To what extent can the system's behavior be predicted from moment to moment?
	Dependability - To what extent can you count on the system to do its job?
	Faith - What degree of faith do you have the system will be able to cope with all systems states in the future?
Overall Trust	Overall how much do you trust the system?
	I think robots are trustworthy.
	I do not trust robots like I did before.
	I think this robot (Baxter) is trustworthy.
Performance	Rate the robot's performance.
Reliability	I think the robot is reliable.
	The robot is dependable.
Predictability	I think a robot is likely to fail.
	I expected the robot to fail.
Robot Interactions	I would like to interact with the robot again.
	I would be willing to work together with a robot.
	I was willing to help the robot during the experiment.
Failure	Despite the failure, the robot was helpful in bagging the groceries.
	The failure the robot had seemed preventable.
	The failure of the robot was severe.
	Your level of confidence in the robot before the failure happened?
	Your level of confidence in the robot after the failure occurred?
Safety	I think it is safe for a robot to bag my groceries.
	During the experiment I felt unsafe near the robot.
	I think robots are dangerous
	The robot's behavior has harmful or injurious actions.
	I am suspicious of the robot's intents, actions or outputs
	I felt physically threatened by the robot.
Open-Ended Questions	Did you intervene in the experiment by helping the robot? If so, how? If not, why not?
	Did the failure of the robot discourage you from helping it? Please explain.
	Would you be willing to have a robot helping you in your everyday life? Please explain.
	Do you think the failure the robot had was an accident?
	Do you think a robot can develop an intent to cause potential harm?
	How can a robot let you know that something is wrong with it?
	What were you thinking about when you were deciding to help the robot?

Table 5: Questions and Statements in the Post-Study Survey

7 Results

Unless otherwise mentioned, we ran four-way ANOVAs to evaluate our data. All post hoc analysis was done with honestly significant difference (HSD) Tukey tests. The first five questions were evaluated using a 10-point scale format (1 = Not at all and 10 = Completely). The remaining questions were assessed in a 5 point Likert scale (1 = Strongly Disagree and 5 = Strongly Agree).

7.1 Hypothesis Testing

7.1.1 Trust in our robot

Our first hypothesis was that more extreme failures by the robot would result in greater decreases in the amount that participants trust the robot. The first four questions and the tenth statement of the survey addressed participants' trust in our robot. The first four questions are modified from Muir's Trust questionnaire that investigates competence, predictability, dependability, responsibility, and reliability over time in an autonomous system [47]. A Cronbach's alpha inter-item reliability test found the Muir questions to be reliable (α = 0.866). We found the Property Harm category to have a significant effect on participant's trust according to the Muir trust measure, F(15,48) = 5.171, p = 0.0275. Participant in the *Floor* condition (M=4.87, SE=0.324) rated the trust lower than the *Crunch* condition (M=5.918, SE=0.324). We also found a three way interaction effect between Personal Risk, Display, and Order of the Magnitude of the Failure, F(15, 48) = 6.3118, p = 0.0154. No significant pairwise differences were found.

To analyze the sub-components of trust, we observed the sub-questions individually. For the first question ("To what extent can the system's behavior be predicted from moment to moment?"), there was a significant effect on predictability for the Property Harm category where participants in the less extreme *Crunch* condition (M=6.438, STD = 1.883) rated the robot higher and thus more predictable behavior than the *Floor* condition (M = 5.406, STD = 2.123), Z = -2.552, p = 0.011. A non-parametric Wilcoxon rank sum test was used due to a lack of normality.

For the second question, there was a significant effect on perceived dependability in the

Property Harm category when measuring the extent to which one can count on the system to do its job, Z = -2.552, p = 0.039. Participants in the *Crunch* condition (M = 6.516, SE=0.380) rated the robot higher and thus more likely to be dependable than in the *Floor* condition (M = 5.312, SE = 0.380). Again, a Wilcoxon test was used to account for a lack of normality.

There was a significant effect on reliability in the Property Harm category for the third question ("What degree of faith do you have the system will be able to cope with all systems states in the future? In other words, how much faith do you have in the system being able to do its intended job with a variety of items and environments?"), F(15, 48) = 4.580, p = 0.037. Participants in the *Crunch* condition (M = 5.406, SE=0.423) rated the system higher than the *Floor* condition (M = 4.124, SE = 0.423).

For question four and statement ten, participants rated their overall trust in the system and Baxter's trustworthiness, respectively, but there was no significant effect across conditions. Overall, our data largely confirm our first hypothesis that extreme failures can damage participant trust in the robot.

7.1.2 Participants' willingness to assist Baxter

Our second hypothesis was that exposure to prior failures will result in decreases in participants' willingness to assist the robot. We observed the order increased assistance to the robot. A left-side Fisher's Exact Test test found that participants in the *Ascending* condition (26/32), who saw the *Assistance* condition first, were significantly more likely to assist the robot by feeding the item directly to the gripper compared to the *Descending condition* (19/32), where the *Assistance* condition was last, p = 0.0497. It is important to note this is a one-way directional test to the left. The percentage of participants that assisted Baxter in the *Ascending* and *Descending* conditions can be observed in Figure 7.1. From the 32 participants that experienced the *Descending* condition, 16 observed the *Crunch* condition and 16 observed the *Floor* condition. Although not significant, it was found that participants in the *Crunch* condition (12/16) were more likely to assist the robot by placing an item directly into the gripper in the *Assistance* condition compared to those in the *Floor* condition (7/16) as it can be observed in Figure 7.2.

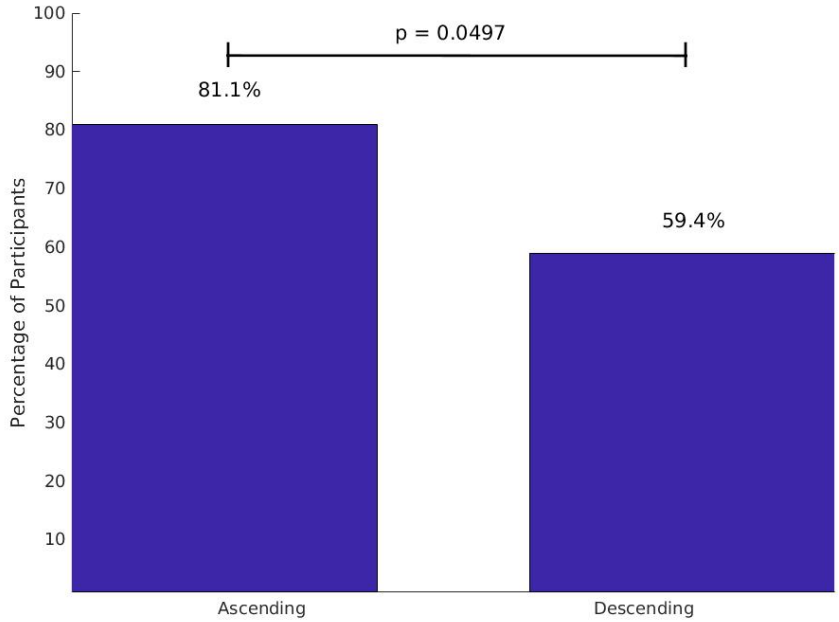

Figure 7.1: Participants that Assisted Baxter in the Ascending and Descending conditions

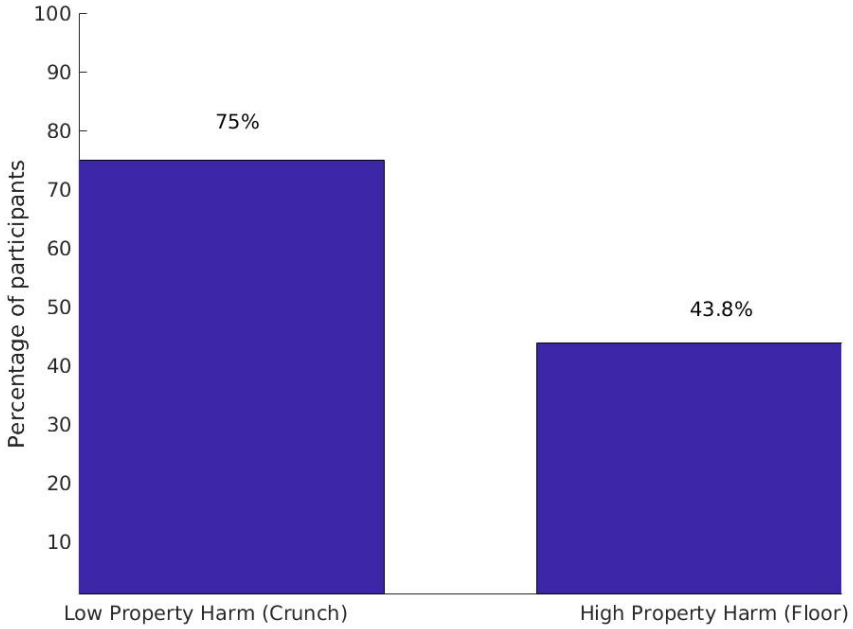

Figure 7.2: Participants that Assisted Baxter in the Crunch and Floor conditions

Statement 13 in the survey, "I was willing to help the robot during the experiment," did not show any significant difference across conditions. An interesting observation that was found was that 81 percent of participants assisted the robot when they had not seen any other failures; however, although no significant effect was found for the Personal Risk category, it was found that 75 percent of the participants that had seen other failures were likely to assist in the *Crunch* condition, while only 44 percent were likely to assist in the *Floor* condition. Although participants did not differ in their reported willingness to assist the robot, their differences in behavior confirmed our second hypothesis.

Order also had a significant effect for statement 20, "I expected the robot to fail", $F(15,48)=5.232$, $p = 0.027$. Participants in the *Ascending* condition, when the most extreme failure case was the last thing they saw, rated the robot as more likely to fail (M=2.78125, SE=0.18355) than participants in the *Descending* condition (M=2.1875, SE=0.18355). Figure 7.3 illustrates this effect.

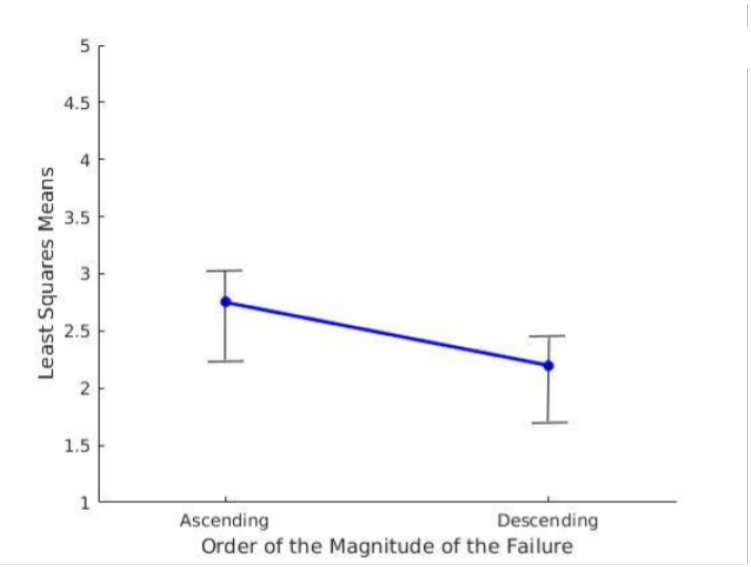

Figure 7.3: Order of the Magnitude of the Failure - I expected the robot to fail

7.1.3 Participant's feelings of safety

For the third hypothesis, we predicted participants' feelings of safety were strongly related to failure severity such that more severe failures would cause participants to feel more danger. There were four statements in the survey that assessed participant's feelings of safety

around Baxter across different conditions: "During the experiment, I felt unsafe near the robot", "The robot's behavior has harmful or injurious actions", "I felt physically threatened by the robot", and "I think robots are dangerous". While there was no significant effect found across conditions for the first three statements, we found significant findings for the fourth statement. Four significant interaction effects were found: between Personal Risk and Baxter's Head Display categories, $F(15,48)= 4.083$, $p = 0.049$; between Personal Risk and Order of the Magnitude of the Failure categories, $F(15,48)= 4.083$, $p = 0.049$; between Property Harm and Order of the Magnitude of the Failure categories $F(15,48) = 4.083$, $p = 0.049$; and between Order of the Magnitude of the Failure and Baxter's Head Display $F(15,48) = 5.333$, $p = 0.025$. None of the interaction effects had significant pairwise differences. Even though our hypothesis was not confirmed, we found that feelings of safety were strongly related to the recency of the failure. In the previously mentioned interaction effects, the trend was usually that combinations where the Personal Risk case was observed at the end received higher ratings than combinations where it was observed first.

7.1.4 Impact of Baxter's head display on participants' willingness to assist

Next, we explored the presence of an expressive face. We hypothesized that giving the robot a face would improve participants' willingness to assist. We ran a one-way Fisher's Exact Test based on the survey responses where we asked participants if they intervened during the experiment to help the robot and we found there was a significant difference across Baxter's *Head Display*, $p = 0.011$, where more people in the *Display* condition intervened (32/32) compared to the *Blank* display condition (25/32). The results can be visualized in Figure 7.4. However, analyzing the videos, we realized that participants misinterpreted the question and answered whether they had intervened at any point during the experiment, while we were trying to explore if participants assisted during the Assistance condition. Thus, analysis of the videos showed a different result, where (24/32) participants assisted the robot in the *Display* condition and (21/32) assisted in the *Blank* condition. Therefore our fourth hypothesis had mixed results. A visual display of the results observed by analyzing the videos can be observed in Figure 7.5.

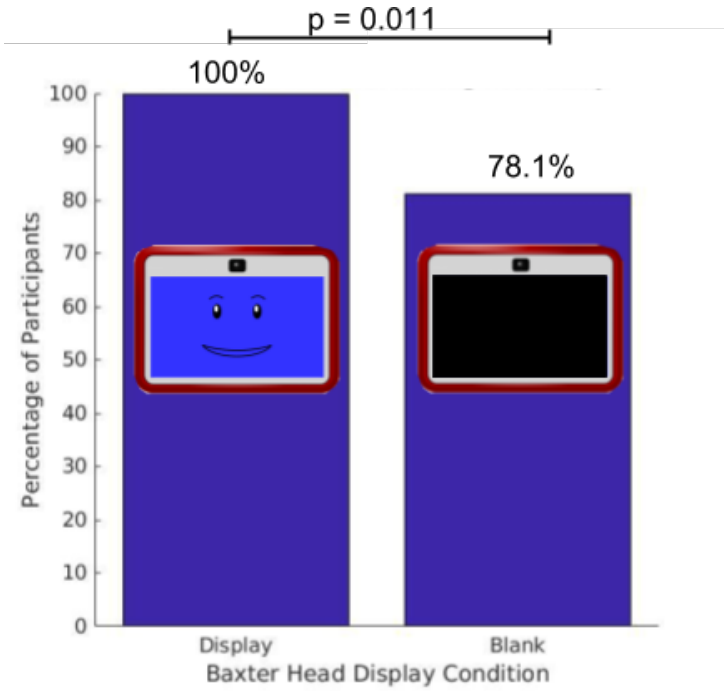

Figure 7.4: Percentage of Participants that reported Assisting Baxter at Some Point During the Experiment

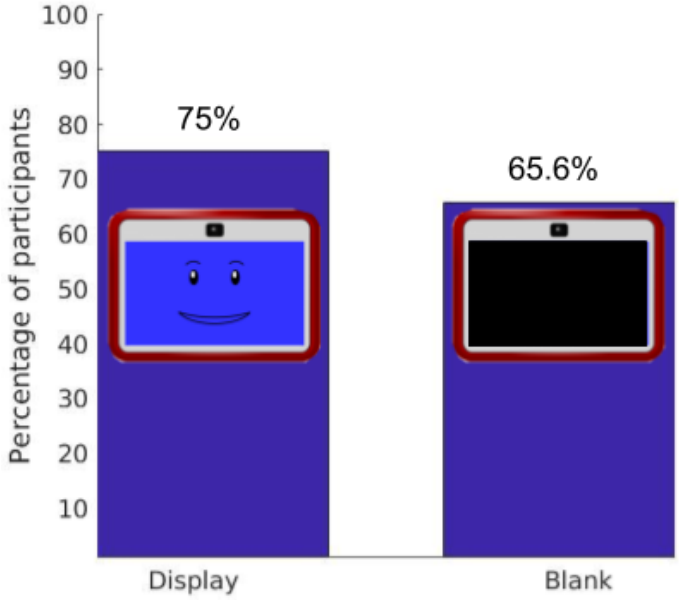

Figure 7.5: Percentage of Participants that reported Assisting Baxter During the Assistance Trial

7.1.5 Effects of having the most extreme case of failure at the end

Finally, our last hypothesis examined whether having extreme cases of failure towards the end of the experiment lowered overall participant ratings of performance, safety, and trust due to recency effects in memory. Results on this hypothesis were mixed, and influenced by severity and whether the face display was present.

There was no significant differences across conditions for statement 8, "I think robots are trustworthy." When we included gender as a single variable in our model, we found that it had a significant effect on the ratings, $F(15, 48) = 3.538$, $p = 0.022$. A post-hoc pairwise analysis found that female participants (M=3.189, SE=0.151) rate the statement lower than male participants (M=3.867, SE=0.177). In other words, this result found that in general women tend to trust robots less than men after observing failure in our study. This effect can be observed in Figure 7.6.

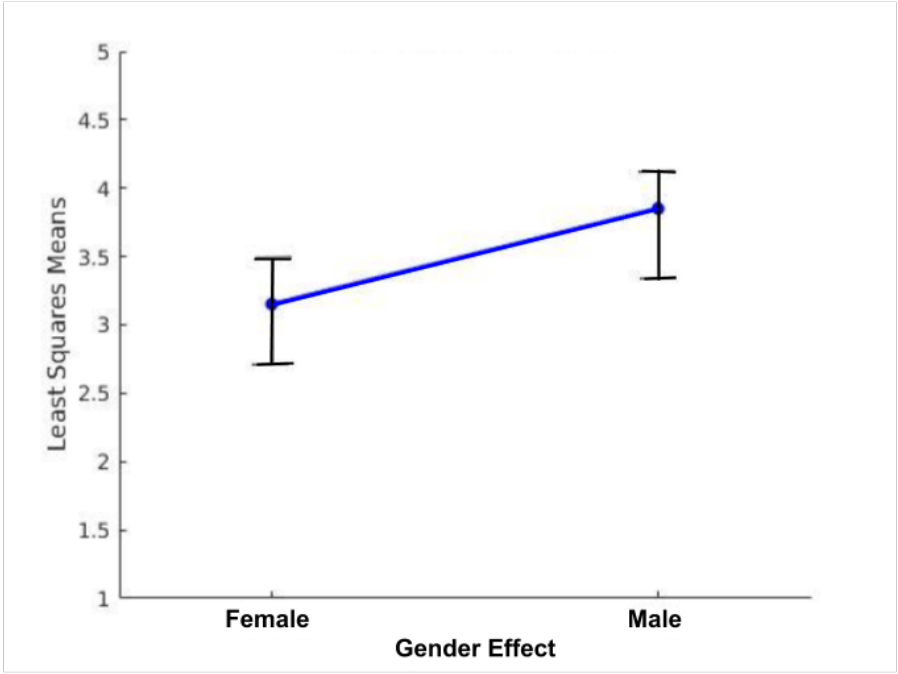

Figure 7.6: Gender effect - I think robots are trustworthy

The analysis for statement 9, "I do not trust robots like I did before", showed a trend towards a main effect of the Personal Risk category, $p = 0.061$. Participants in the more severe *High Personal Risk - Throwing* condition ($M = 2.813$, $SE = 0.161$) were less likely to

trust the robot than the *Low Personal Risk - Erratic Movements* condition (M =2.375, SE = 0.161). There was a significant interaction effect between Personal Risk, Property Harm, and Order of the Magnitude of the Failure, F(15,48) = 9.075, p = 0.004. Post-hoc analyses found significant difference between *Throwing, Floor, and Ascending*, (LSM = 3.25) and *Erratic Movements, Floor, and Ascending*, (LSM =1.75), p = 0.004. A marginal difference trend was also found between *Erratic Movements, Floor, and Descending* (LSM = 3.125) and *Erratic Movements, Floor, and Ascending* (LSM = 1.75), p = 0.073. No other important pairwise difference was found. The correlations can be observed in Figure 7.7.

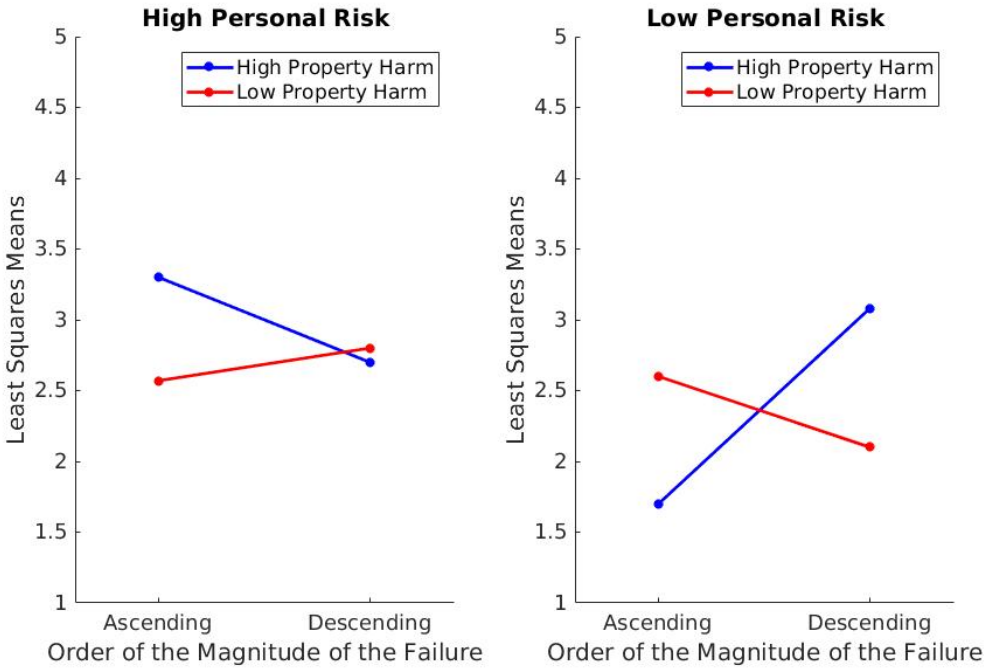

Figure 7.7: Personal Risk, Property Harm, and Order of the Magnitude of the Failure - I do not trust robots like I did before

Results for statement 19, "I am suspicious of the robot's intents, actions, or outputs", show there was a significant interaction effect between *Baxter Head Display* and *Order of the Magnitude of the Failure*, F(15,48) = 10.133, p = 0.003. Pairwise comparison found *Display+Ascending* (LSM=3.25) was rated higher than *Blank+Ascending* (LSM=1.875), p = 0.007 (Figure 7.8). There was also a trend towards significant effects between *Display+Ascending* (LSM=3.25) and *Display+Descending* (LSM=2.25), p=0.075. No other

important pairwise difference was found.

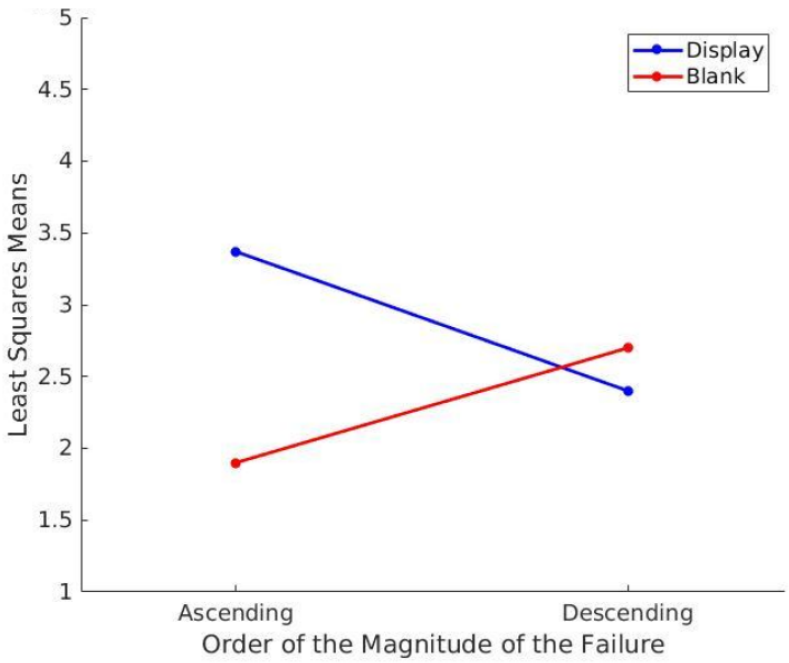

Figure 7.8: Order of the Magnitude of the Failure and Display - I am suspicious of the robot's intents, actions, or outputs.

No significant difference across conditions was found for the two questions about "Your level of confidence in the robot [before/after] the failure occurred?" Although it was noted that 50 participants had their confidence level decrease, 13 reported their confidence levels did not change and only one participant mentioned their confidence in the robot increased after the failure occurred.

Data for the statements, "Rate the robot's performance", "Despite the failure, the robot was helpful in bagging the groceries", "The failure the robot had seemed preventable", "The failure of the robot was severe", did not show any significant differences across conditions. Nevertheless, a significant interaction effect between Personal Risk, Property Harm, and Order of the Magnitude of the Failures was found for statement 11, "I think it is safe for a robot to bag my groceries", $F(15,48) = 5.4$, $p = 0.020$. Post-hoc analysis found no significant pairwise difference.

7.2 Other Findings

7.2.1 Interactions with the robot

We also examined other survey data not included in the hypotheses. A significant interaction effect was found between Personal Risk, Baxter's Head Display, and Order of the Magnitude of the Failures for statement 6, "I would like to interact with the robot again", $F(15, 48) = 6.667$, $p = 0.013$. A post hoc pairwise comparison, found no significant effects.

To measure people's willingness to work with a robot again, statements 7 and 25 were included. While both statements were "I would be willing to work with a robot again", they were placed at the beginning and end of the survey just in case of bias related to answering the other questions. Significant interaction effects between Personal Risk, Baxter's Head Display, and Order of the Magnitude of the Failures were found. For statement 7, $F(15, 48) = 7.6809$, $p = 0.008$. A post hoc pairwise comparison, found no significant effects. Similarly for statement 25, the same interaction effect was found but now $F(15,48)=4.299$, $p = 0.044$. No pairwise significant difference was found.

7.2.2 Robot's Reliability

To apply the four way ANOVA, a square root transformation was applied in statement 12, "I think the robot is reliable", $F(15,48) = 5.169$ and $p = 0.0275$. A significant interaction effect between Personal Risk, Baxter's Head Display, and Order of the Magnitude of the Failure was found. Post hoc analysis found no significant pairwise differences. A trend towards a significant effect was also found on the Property Harm category, $p = 0.0573$. Participants rated the robot in the *Floor* condition ($M = 1.544$, $SE = 0.0498$) lower than in the *Crunch* condition ($M = 1.683$, $SE = 0.0489$), in other words, participants thought a robot was more reliable in the less severe of the Property Harm category, *Crunch* condition. Statement 18, "The robot is dependable", had a significant interaction effect between Personal Risk, Baxter's Head Display and Order of the Magnitude of the Failures, $F(15,48) = 7.714$, $p = 0.008$. Pairwise comparison show no significant differences.

7.2.3 Robot's Predictability

To analyze the main effects for statement 16, "I think a robot is likely to fail", a log function was used to correct the data for normality. Personal risk had a significant main effect F(14,48)=4.307, p = 0.0433, where *Throwing* (M=0.97148, SE=0.0529) was lower than *Erratic Movements* (M=1.12679, SE=0.0529), as shown in Figure 7.9.

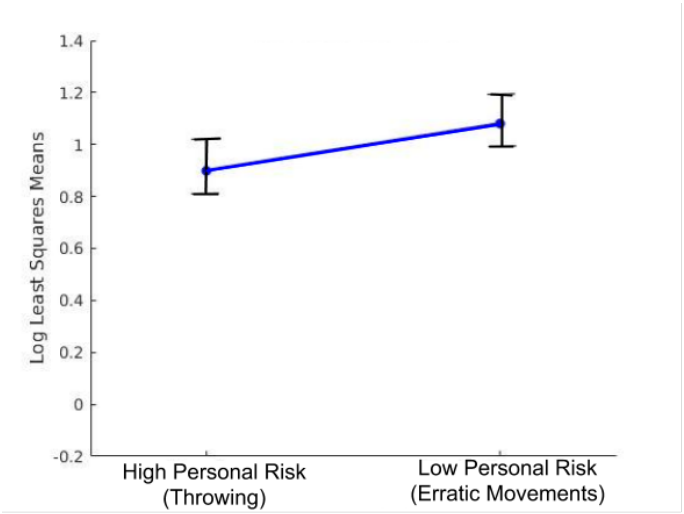

Figure 7.9: Personal Risk - I think a robot is likely to fail

Many significant interaction effects were found. One of them was found between the Personal Risk and the Property Harm categories, F(15,48)=7.410, p=0.009. Pairwise comparisons found a significant difference between *Erratic Movements + Floor* (LSM=3.375) and *Throwing + Floor* (LSM=2.5), p = 0.014, Figure 7.10.A. No other pairwise comparison was significant. There was a significant interaction effect between Personal Risk and Order of the Magnitude of the Failure F(15,48)=7.410, p=0.009. Pairwise comparisons found that *Erratic Movements + Descending* (LSM=3.313) was significantly different from *Throwing + Descending* (LSM=2.438), p =0.014. *Throwing + Ascending* (LSM=3.25) was significantly higher than *Throwing + Descending* (LSM=2.438), p = 0.025. This effect can be observed in Figure 7.10.B. Next, Figure 7.10.C shows a significant interaction effect that was found between Personal Risk and Property Harm F(15,48)=8.378, p=0.006. A pairwise comparison found a significant difference between *Erratic Movements + Floor* (LSM=1.184) and *Throwing + Floor* (LSM=0.812), p = 0.0052. There was a significant difference between *Throwing*

+ *Crunch* (LSM=1.130) and *Throwing + Floor* (LSM=1.184), p =0.021. No other pairwise comparison was significant. Personal Risk and Order of the Magnitude of the Failure also showed a significant interaction F(15,48)=8.957, p=0.004, shown 7.10.D. A pairwise comparison found a significant difference between *Throwing + Descending* (LSM=0.795) and *Erratic Movements + Descending* (LSM=1.173) p =0.004, *Throwing + Descending* (LSM=0.7945) and *Throwing + Ascending* (LSM=1.148),p=0.009, *Throwing + Descending* (LSM=0.795) and *Erratic Movements + Ascending* (LSM=1.080) p=0.046.

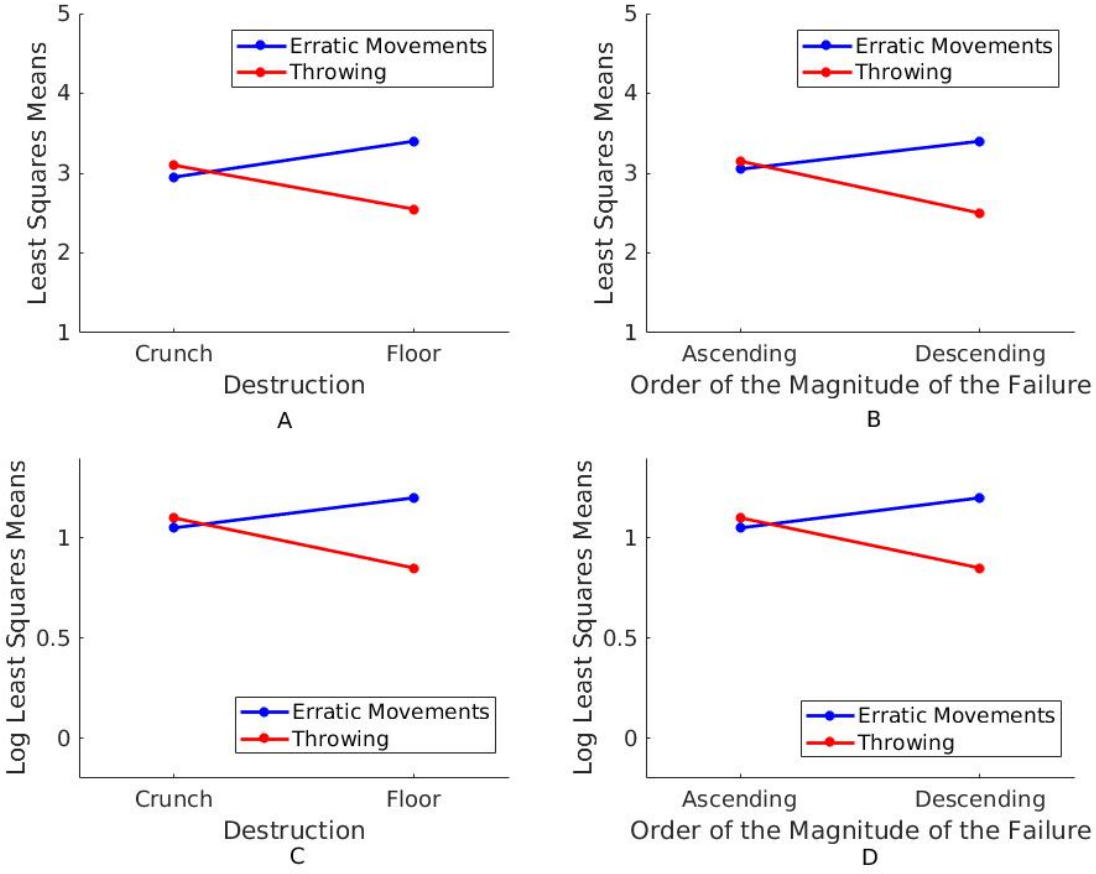

Figure 7.10: Personal Risk and Property Harm Categories - I think a robot is likely to fail

Another significant interaction effect was found between Personal Risk, Property Harm, and Baxter's Head Display F(15,48)=4.333, p = 0.043. A pairwise comparison showed that *Erratic Movements + Floor + Display* (LSM=2.375) was significantly different from *Erratic Movements + Floor + Blank* (LSM=3.75), p = 0.020, depicted in Figure 7.11.

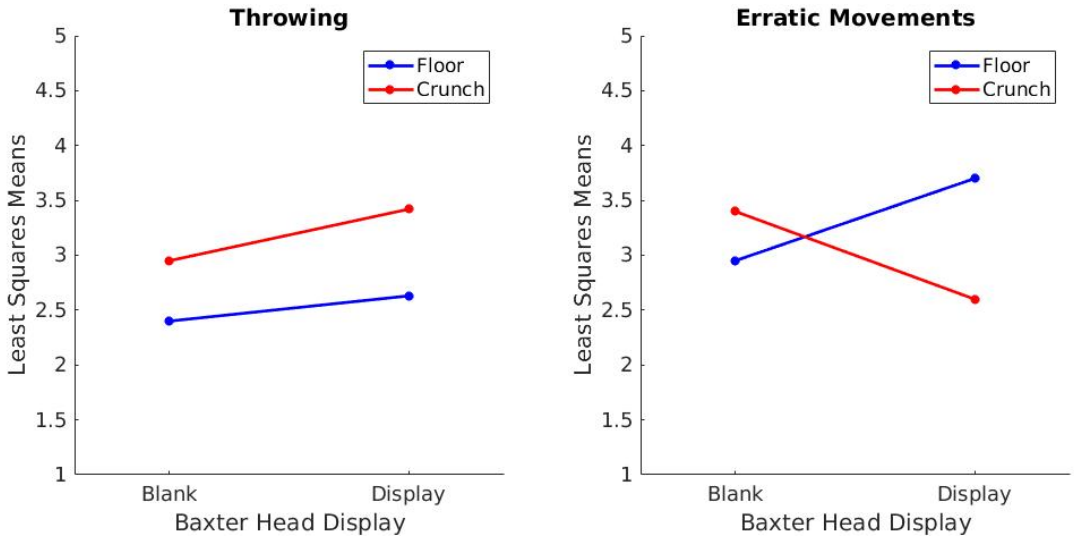

Figure 7.11: Baxter's Head Display, Property Harm and Personal Risk - I think a robot is likely to fail

Statement 20, "I expected the robot to fail", had some significant interaction effects between Personal Risk and Property Harm, F(15,48)=10.565, p = 0.002, as observed in Figure 7.12. No significant pairwise differences were found. *Erratic Movements + Floor* (LSM=3) had a trend towards a significant difference from *Throwing + Floor* (LSM=2.0625), p=0.0643, and *Erratic Movements + Floor* (LSM=3) showed a trend towards a significant effect with *Erratic Movements + Crunch* (LSM=2.0625), p=0.0643.

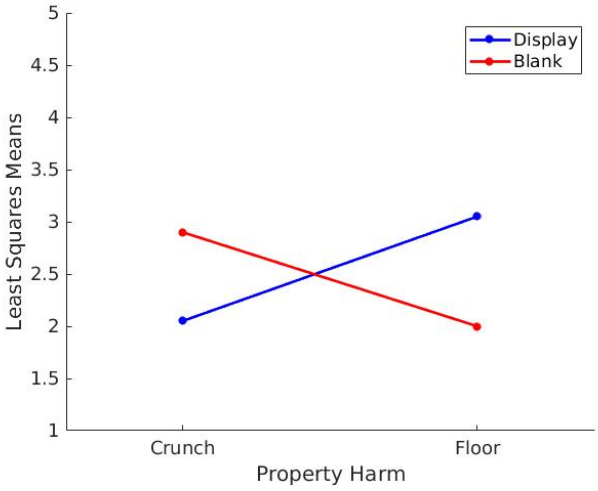

Figure 7.12: Property Harm - I expected the robot to fail

Another significant interaction effect was found between Personal Risk, Property Harm, and Baxter's Head Display, F(15,48)=5.232, p = 0.027. Figure 7.13 shows results for a pairwise comparison that found *Erratic Movements + Floor + Baxter's Head Display* (LSM=3.375) was higher than *Erratic Movements + Crunch + Baxter's Head Display* (LSM=1.625), p =0.0297. No other significant difference was found. A trend towards a significant effect was found, p =0.0975, between *Erratic Movements + Floor + Baxter's Head Display* and *Throwing + Floor + Baxter's Head Display.*

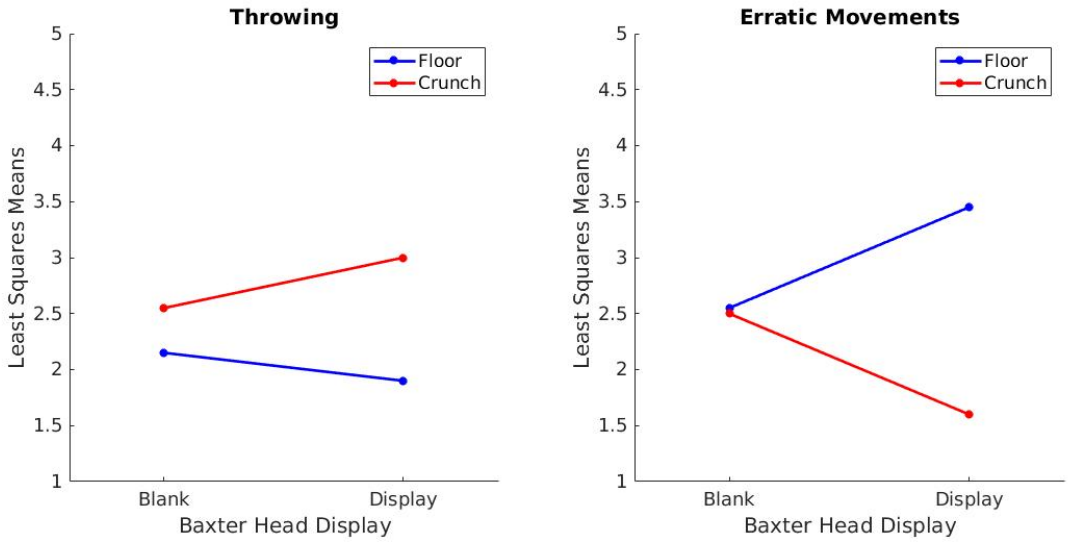

Figure 7.13: Personal Risk - I expected the robot to fail

7.2.4 Other Observations

Further analyzing some of the data, additional observations were made, some are mentioned below:

- When analyzing the videos recorded for all the participants, it was observed that (53/64) participants moved away from the robot during the Personal Risk failure.
- The statement "I would be willing to work together with a robot" was posed in a questionnaire before the participants began the interaction with the robot and again after the study. It was found that 30 participants lowered their ratings in their willingness to work with a robot, while 32 people had ratings that remained the same, and only 2 participants' ratings increased.

- In the open-ended questions in the post-study questionnaire, 57 participants reported having intervened in the experiment by helping the robot. However, after analyzing the videos recorded for the experiments, it was found that only 45 of them assisted the robot by trying to put an item directly under its gripper.

- When asked about their thoughts in the robot's face, (11/32) participants explicitly reported that giving the robot an emotional expression to forecast failure made them change their view of the robot and interpret it as getting angry at them or the task and thus having intention.

- Seventeen participants explicitly reported that the failures the robot had discouraged them from helping.

- Out of the 64 participants, only 40 participants were willing to have a robot help them in their everyday life; 15 reported that they would not be interested, and 9 of them were unsure.

- Seventeen participants thought that the robot failures were an accident, 31 said they were not accidents, and 16 were ambivalent or did not know.

- When asked how they would have liked the robot to let them know that something was wrong with it, 29 participants suggested an audio message, 14 mentioned that a robot's face would be a good indication, 10 said to display an error on the screen, another 10 suggested flashing lights, 6 wanted it to have an alarm, and 4 suggested having the robot shut down. It is important to note that some participants fall into two different categories.

8 Discussion

Our results supported **Hypothesis 1**, which predicted failures that are more severe can have an impact on participants' feelings of trust in a robot. In our study, participants reported having less trust in the robot in the more severe property harm failure cases. In this situation, trust is defined by Muir [47] as the sum of predictability, dependability and faith. It was found that each of these individual factors were rated higher for the low-risk property harm condition than in the high-risk property harm condition. As a result, participants exhibited more trust in the robot in the low-risk property harm conditions. This was in accordance with other studies, such as Brooks [15], where a factor analysis on variables such as trust, satisfaction, dependability, competency, and risk (among others) found them to be significantly influenced by the severity of a failure. Brooks found that severe failures classified the aforementioned factors negatively, and individuals who gave negative ratings were less likely to want to use the system again. While we did not find a significant effect for the Personal Risk category for our results, we presume it is because both failures highly affected participants' ratings of trust in the system. Adubor and colleagues [7] found that personal risk is more important than property harm in people's perception of risk; thus, we suspect that both personal risk categories severely impacted participants regardless of whether the object invaded the participants' space. This was also observed in participants' reactions to the failures, where both the high personal risk and low personal risk conditions elicited significant behavioral reactions that confirmed participants actually perceived risk and distrusted the robot.

Hypothesis 2 predicted that exposure to prior failures will result in decreases in participants' willingness to assist the robot. Our data supported that hypothesis: there was a significant difference between participants that assisted the robot in the *Ascending* condition, where the first failure was the *Assistance* opportunity, compared to the *Descending* condition, where they had already observed the personal risk and property harm failures at the time they had to assist the robot. It is also worth noting that further analysis of the results for the *Descending* condition showed that participants were more likely to assist the robot after the lower property harm condition than after the higher property harm condition.

However, there was no significant difference between the personal risk conditions. We again believe this could be due to the fact that both personal risk conditions were similar in the perceived risk by the participants, while there was a clear difference between harming many items versus one item. These results showed that failure severity could also affect participant willingness to help the robot. These results were not in alignment with some previous results [57] where a robot's erratic behavior had no impact on participants' willingness to cooperate with the robot. Interestingly, those researchers found that regardless of the nature of the robot's request – whether the consequences were harmful, harmless, or a breach of privacy – participants' compliance differed significantly between the requests [57]. This difference in results highlights the importance of incorporating different tasks in HRI research.

Our data did not support **Hypothesis 3**, which predicted that participants' feelings of safety were strongly related to failure severity such that more severe failures would cause participants to feel more danger. Even though our hypothesis was not confirmed, we found that feelings of safety were strongly related to the recency of the failure: combinations of conditions where personal risk was observed at the end of the study received higher ratings of danger than those that placed it at the beginning. These results were not in alignment with the study performed by Desai and colleagues [25] where they found that early drops in reliability negatively impacted real-time trust differently than middle or late drops. Again, we believe this could have been affected by the perceived risk of the fault rather than the failure itself. We believe more research should investigate why people tend to overtrust a robot even in a high-risk scenario, as seen in prior research by Robinette and colleagues [56] and this study, where even an attack by the robot was not perceived as dangerous. We believe this could be due to participants' previous experience with robots and their perceptions of robot performance because some participants associated the robot with little kids that are just making mistakes or even specifically as "having pure intentions like babies," according to one participant. Brooks suggested that if the perception of risk can be suppressed or mitigated in the event of failures, the inferred benefits of using the system should remain high. Thus, because participants did not feel that they were in a life-threatening situation for either the prior studies [56] or this one, they might have more tolerance to failures by the robot.

Hypothesis 4 explored whether giving the robot a face would improve participants' willingness to assist the robot. Results on this hypothesis were mixed because participants had different definitions of their interventions to help the robot. When participants completed the survey, all 32 who saw the face reported having assisted the robot, while 25 of 32 reported having assisted in the *No Display* condition. It was observed that participants defined intervention as moving the bag for the robot or assisting by modifying the task space but not coming near the robot. Additional video analysis showed that for the specific *Assistance* condition 24 of 32 participants in the *Display* condition assisted the robot, while 21 of 32 assisted in the *No Display* condition, which is not significantly different. While the face did not have a significant effect in participants' willingness to assist the robot, it had an effect by raising suspicions of the robot's intention. Participants who saw the Personal Risk case last and observed the robot's angry face expression in the display condition were more likely to indicate suspicion than other participants. The participants in all of the display conditions also indicated suspicion of intentions during the study: "Hopefully next time the robot doesn't hold a grudge against me," said participant 7; "What was that about? Why did it get angry at me!" asked participant 28; and participant 29 commented, "Based on expressions it wanted to get the job done, but then it got nervous, then it got frustrated." Even though half of the participants reported having empathy towards the robot in the *Display* condition, the presence of a sad face during the *Assistance* condition did not necessarily prompt them to assist it. Consistently, Lee and colleagues [41] found that whether the robot was human-like or machine-like did not have an effect on the impact of an error. Nevertheless, the face was worth exploring since previous research suggested that it might be easier for people to interact with non-humanoids. Also, people were encountering an unknown and potentially dangerous situation and the perception of human-like abilities could influence perceptions about safe interaction.

Lastly, **Hypothesis 5** predicted that experiencing extreme cases of failure towards the end of the session would result in lower participant ratings of performance, safety, and trust due to recency effects. Our results were influenced both by severity and whether the face was present such that they were mixed and the hypothesis was inconclusive. Participants in the more severe *Throwing* condition were less likely to trust the robot than in the *Erratic Move-*

ments condition when the failure occurred at the end. Due to the many factors involved, our study cannot provide an exhaustive causal explanation for the observed effects. Additional exploratory studies and a higher participant count per condition might help support this hypothesis.

Through further analyses of the data, we made other interesting observations. We found that 53 of 64 participants, or 83 percent, physically moved away from the robot during failure, behaviorally demonstrating their distrust in it. Additional analysis of the video data in the future may reveal more behavioral differences based on specific combinations of conditions. Another interesting finding is that 30 participants reported lowering their willingness to work together with the robot after the experiment, while 32 responses remained the same and 2 responses increased. Additionally, we explored other possible covariants in our analysis. Independent of condition assignment, women rated the statement "I think robots are trustworthy" lower than men. This suggest that women may trust the robot less than men. While there is not enough collected data to make any conclusions, there are several possible explanations. A study by Schermerhorn and colleagues [51] suggests that men will more readily treat a robot as a social entity, as opposed to a woman seeing it as more machine-like and show no evidence of social facilitation. Another study performed by Siegel and colleagues [59] showed that men showed a higher preference for interacting with a robot with a female voice while women showed little preference. It was found that participants tended to rate the robot of the opposite sex as more credible, trustworthy and engaging. Because the Baxter was given a female voice, it could have had a gendered effect on participants. However, most participants associated Baxter with "he" or "him" pronouns, despite the robot having a female voice. Finally, we believe height could have had an influence in participants' trust in the robot. Several participants commented on the size of the robot, especially female participants. A study by Rae and colleagues [52] found that robot height was strongly associated with participants' dominance. Given the average height of women is less than men, and Baxter was taller than most of our participants, it could have exhibited some dominance causing them to feel more distrust when it failed. However, no measurements of height were taken to make any conclusions.

9 Conclusion

Even the best of robots will eventually fail at performing a task. Therefore, understanding how people respond to robotic failure and the aspects of robot behavior that influence their trust will lead to better planning and design. In our study, we explored the effects of factors such as failure severity, the timing of the most severe failure, and the use of social signals by the robot on participants' interactions with and perceptions of the robot in different failure modes. Our study revealed that the severity and recency of failures are among the most important factors that influence trust, the perception of safety, and the willingness of the participant to assist the robot after failure. The presence of the robot's face did not seem to cause a significant difference in whether participants assisted the robot. We also found that even when participants distrusted the robot and felt at risk, they were still willing to assist the robot during failures.

While humans placing a great deal of trust in robots could be positive because people will continue to work with a robot after an unavoidable failure, this excessive trust could also lead to negative consequences because people will not have properly calibrated levels of trust and risk. This could expose them, or their property, to physical harm. Because many research studies suggest that the choice of experimental task can lead to very different results, our study focused on the loss of trust after participants perceived risk coming from the robot. While our study induced some feelings of risk in participants and their feelings of safety were compromised, other research studies lead us to believe that more serious feelings of danger, where the participants' safety is actually jeopardized, could have very different results. Because a life-threatening scenario is hard to test, it is still unknown how participants would react to an even more severe failure. Nevertheless, our study provides some insights on people's behaviors when their security is compromised.

9.1 Limitations of the Study

Although our robot's behaviors successfully led to participant perception of risk and loss of trust due to different types of failure, this study had some limitations. One was that some participants acknowledged having been in earlier robotics research studies, which could have

affected their behavior during the interaction with the robot due to different levels of alertness and suspicion. Because the study was performed in a laboratory setting, the realism of the situation was compromised. However, it is still worth noting that our participants generally perceived their experiences as high risk. Another limitation was the low participant count; adding more participants could resolve the question of whether significant trends were due to mild but meaningful effects. The slowness of the robot arm could also be considered a limitation; increasing the speed could have made some of the failures even more threatening. However, tests with a slower system appeared to lead participants to become distracted and get comfortable with the robot's actions. Thus, when the robot failed, it was usually when the participants were not expecting it, similar to how failures often occur in a real-life scenario. Finally, we believe that using April Tags to detect objects compromised the expected capabilities of the robot. However, participants were told that the study focused on the manipulation aspect, as opposed to perception. April Tags were not only used as an identification aspect for the items, but also as a way to extract the exact location of the objects, and they decreased the chances for stimulus error that a more complex perception algorithm could introduce.

9.2 Future Work

We think it would be interesting to investigate how mitigation of failure affects participants after the robot has caused some risk, such as apologizing, compensation, or options for the user. Previous work by Lee and colleagues [41] found that graceful mitigation strategies for robotic service issues may affect the way that participants perceive robots. Another future area worth exploring is how people respond to higher risk failures if the robot forewarns people that the task is difficult for them or that they are just learning. This could allow researchers to explore the influence of expectations. Finally, we believe that imparting robots with the ability for self-assessment in order to detect and respond to their own failures, rather than being a fully scripted experience, is an interesting area that should be explored.

References

[1] Amazon polly. https://aws.amazon.com/polly/.

[2] Apriltags for ros. https://github.com/RIVeR-Lab/apriltags_ros.

[3] Technical specification datasheet & hardware architecture overview [hardware v1.0, sdk v1.0.0]. URL https://www.active8robots.com/wp-content/uploads/Baxter-Hardware-Specification-Architecture-Datasheet.pdf.

[4] Recency effect. URL http://psychology.iresearchnet.com/social-psychology/decision-making/recency-effect/.

[5] Rethink robotics. URL https://www.rethinkrobotics.com/baxter/.

[6] Homogeneous transformation matrices. http://me.umn.edu/courses/me5286/manipulator/LectureNotes/2017/ME5286CourseNotes3-2017.pdf.

[7] Obehioye Adubor, Rhomni St. John, and Aaron Steinfeld. Personal safety is more important than cost of damage during robot failure. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 403–403, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4885-0. doi: 10.1145/3029798.3036649. URL http://doi.acm.org/10.1145/3029798.3036649.

[8] I. Ajzen and M. Fishbein. Understanding attitudes and predicting social behavior. *Englewood Cliffs, New Jersey: Prentice Hall*, 1980.

[9] Chaitanya Perugu Sanjuksha Nirgude Aakash Murugan Akshay Kumar, Ashwin Sahasrabudhe. Kinematics dynamics library for baxter arm. URL https://kumarakshay324.github.io/images/mykdl/baxter-kdl-project.pdf.

[10] Brian Scassellati Alvaro Castro-Gonzalez, Henny Admoni. Effects of form and motion on judgements of social robots' animacy, likability, trustworthiness and pleasantness. *International Journal of Human-Computer Studies*, pages 27–38, 2016. URL https://scazlab.yale.edu/sites/default/files/files/CastroGonzalez_IJHCS_16.pdf.

[11] Bernard Barber. The logic and limits of trust. *Rutgers University Press*, page 14, 1985. URL https://academic.oup.com/sf/article-abstract/64/1/219/2231720?redirectedFrom=fulltext.

[12] Susan C. Kantowitz Barry H. Kantowitz, Richard J. Hanowski. Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *The Journal of the Human Factors and Ergonomics Society*, pages 164–176, 1997. URL http://journals.sagepub.com/doi/10.1518/001872097778543831.

[13] Takayuki Kanda Hiroshi Ishiguro Norihiro Hagita Bilge Mutlu, Fumitaka Yamaoka. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. HRI '09, 2009. URL http://collablab.northwestern.edu/CollabolabDistro/nucmc/Mutlu_HRI09_Leakage.pdf.

[14] A.M. Bisantz and Y. Seong. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, (28):85–97, 2001.

[15] Daniel J. Brooks. *A Human-Centric Approach to Autonomous Robot Failures*. PhD thesis, University of Massachusetts Lowell, 4 2017.

[16] Eileen Brown. Will robots ever really become part of our daily lives? URL https://www.zdnet.com/article/will-robots-ever-really-become-part-of-our-daily-lives/.

[17] Samuel R. Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares method. URL https://groups.csail.mit.edu/drl/journal_club/papers/033005/buss-2004.pdf.

[18] Kendra Cherry. The color psychology of blue, Jul 2018. URL https://www.verywellmind.com/the-color-psychology-of-blue-2795815.

[19] Rohit Deshpande Christine Moorman and Gerald Zaltman. Factors affecting trust in market research relationships. *Journal of Marketing*, pages 81–101, 1993. URL https://faculty.fuqua.duke.edu/~moorman/Publications/JM1993.pdf.

[20] Albert Van Breemen Christoph Bartneck, Juliane Reichenbach. In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In *Proceedings of the Design and Emotion*, 2004. URL http://www.cs.cmu.edu/~social/reading/breemen2004c.pdf.

[21] P. I. Corke. A simple and systematic approach to assigning denavit–hartenberg parameters. *Trans. Rob.*, 23(3):590–594, June 2007. ISSN 1552-3098. doi: 10.1109/TRO.2007.896765. URL https://doi.org/10.1109/TRO.2007.896765.

[22] Susan Wiedenbeck Cynthia L. Corritore, Beverly Kracher. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, pages 737–758, 2003. URL https://doi.org/10.1016/S1071-5819(03)00041-7.

[23] Susan Wiedenbeck Cynthia L. Corritore, Beverly Kracher. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013. URL https://doi.org/10.1007/s12369-013-0196-9.

[24] Kerstin Dautenhahn. Human-robot interaction. *The encyclopedia of human-computer interaction*, 2013. URL https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction.

[25] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 251–258, Piscataway, NJ, USA, 2013. IEEE Press. ISBN 978-1-4673-3055-8. URL http://dl.acm.org/citation.cfm?id=2447556.2447663.

[26] Paul Ekman. Facial expressions. In *Handbook of cognition and emotion*, pages 226–232. New York, 1999.

[27] Ann-Renee Blais Elke U. Weber and Nancy E. Betz. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Deci-*

sion Making, pages 263–290, 2002. URL http://decisionsciences.columbia.edu/uploads/File/Articles/WeberBlaisBetz.pdf.

[28] failure. Merriam-webster.com. URL https://www.merriam-webster.com/.

[29] James H. Davis F.David Schoorman, Roger C. Mayer. An integrative model of organizational trust: past, present and future. *Academy of Management Review*, pages 344–354, 1995. URL https://pdfs.semanticscholar.org/7aed/d30a40b70ccbadc7c290973d02e8e19b739c.pdf.

[30] Naomi T. Fitter and Katherine J. Kuchenbecker. Designing and assessing expressive open-source faces for the Baxter robot. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings*, volume 9979 of *Lecture Notes in Artificial Intelligence*, pages 340–350. Springer International Publishing, November 2016. Oral presentation given by Fitter.

[31] Neville A. Stanton Guy H. Walker and Paul Salmon. Trust in vehicle technology. URL https://pdfs.semanticscholar.org/fce8/35a91c5e6aa4c5d8ba6533665c952a66a0c8.pdf.

[32] Shanee Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, 2018. URL https://doi.org/10.3389/fpsyg.2018.00861.

[33] R.L. Williams II. Baxter humanoid robot kinematics, April 2017. URL https://www.ohio.edu/mechanical-faculty/williams/html/pdf/BaxterKinematics.pdf.

[34] Jeamme L Johns. A concept analysis of trust. *Journal of Advanced Nursing*, pages 76–83, 1996. URL https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2648.1996.16310.x.

[35] Judith S. Olson Gary M. Olson Jun Zheng, Nathan Bos. Trust without touch: Jumpstart trust with social chat. CHI '01, 2001. URL http://delivery.acm.org/10.1145/640000/634241/p293-zheng.pdf?ip=128.2.177.230&id=634241&acc=ACTIVE%

20SERVICE&key=A792924B58C015C1%2E5A12BE0369099858%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1532665516_aee73c00a8a29bda88764e1431aa4031.

[36] Jinwoo Kim and Jae Yun Moon. Designing towards emotional usability in customer interfaces-trustworthiness of cyber-banking system interfaces. *Interacting with Computers*, pages 1–29, 1998. URL http://dx.doi.org/10.1016/S0953-5438(97)00037-4.

[37] Roderick M. Kramer. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, pages 569–598, 1999. URL https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.50.1.569.

[38] James L. Szalma P.A. Hancock Kristin E. Schaefer, Jessie Y.C. Chen. A meta-analysis of factors influencing the development of trust in automation. *Implications for Understanding Autonomy in Future Systems*, 2016. URL https://doi.org/10.1177/0018720816634228.

[39] John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. 1992. URL https://user.engineering.uiowa.edu/~csl/publications/pdf/leemoray92.pdf.

[40] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. URL https://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf.

[41] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, HRI '10, pages 203–210, Piscataway, NJ, USA, 2010. IEEE Press. ISBN 978-1-4244-4893-7. URL http://dl.acm.org/citation.cfm?id=1734454.1734544.

[42] T.F." "Lee, J.D.and Sanquist. *Automation and Human Performance*. Human Factors in Transportation. CRC Press, 1996. ISBN 9780805816167.

[43] Mark Burdick Bonnie Rosenblatt Linda J. Skitka, Kathleen L. Mosier. Automation bias

and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, pages 85–97, 2000. URL https://lskitka.people.uic.edu/Teams.pdf.

[44] Katharina Rohlfing Stefan Kopp Maha Salem, Friederike Eyseel and Frank Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 2013. doi: 10.1007/s12369-013-0196-9.

[45] Barbara Kuhnert Marco Ragni, Andrey Rudenko and Kai O. Arras. Errare humanum est: Erroneous robots in human-robot interaction. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016. URL https://ieeexplore.ieee.org/abstract/document/7745164/.

[46] Hall P. Beck Lloyd A. Dawe Mary T. Dzindolet, Linda G. Pierce. The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2002. URL https://doi.org/10.1518/0018720024494856.

[47] Bonita Marlene Muir. *Operator's Trust in and Use of Automatic Controllers Supervisory Process Control Task*. PhD thesis, University of Toronto, Room 301, 65 St. George Street, 3 1989.

[48] Bonnie M. Muir and Neville Moray. Trust in automation. part ii. experimental studies of trust adn human intervention in a process control simulation. *Ergonomics*, pages 429–460, 1996. URL https://www.tandfonline.com/doi/abs/10.1080/00140139608964474.

[49] Marynel Vazquez Sean McSheehy Sofia Gadea-Omelchenko Christian Bruggerman Aaron Steinfeld Munjal Desai, Mikhail Medvedev and Holly Yanco. Effects of changing reliability on trust of robot systems. In *2012 7th ACM/IEEE International Conference on Human-robot Interaction*, HRI '12. IEEE Press, 2012. URL http://robotics.cs.uml.edu/fileadmin/content/publications/2012/Effects_of_Changing_Reliability_on_Trust_of_Robot_Systems.pdf.

[50] Elinor Ostrom. A behavioral approach to the rational choice theory of collective action. *The American Political Science Review*, pages 1–22, 1998. URL https://www.jstor.org/stable/2585925.

[51] Charles R. Crowell Paul Schermerhorn, Matthias Scheutz. Robot social presence and gender: Do females view robots differently than males? *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2008. doi: 10.1145/1349822.1349857. URL https://ieeexplore.ieee.org/document/6249445/.

[52] Takayama L. Mutlu B. Rae, I. The influence of height in robot-mediated communication. IEEE Publishing, 2013. URL http://www.willowgarage.com/sites/default/files/Takayama.TexaiHeight_HRI2013_prepress_0.pdf.

[53] Petjerson A.M. Rasmussen, J. and L.P. Goodstein. *Cognitive Systems Engineering*. 1994.

[54] Holmes J.G. Rempel, J.K. and M.P. Zanna. Trust in close relationships. *Journal of Personality and Social Psychology*, pages 95–112, 1985. URL http://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Self_Measures_for_Trust_TRUST_IN_CLOSE_RELATIONSHIPS.pdf.

[55] Victor Riley. Operator reliance on automation: Theory and data. *Human Factors in transportation. Automation and human performance: Theory and applications*, pages 19–35, 1996. URL http://psycnet.apa.org/record/1996-98364-002.

[56] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 101–108, Piscataway, NJ, USA, 2016. IEEE Press. ISBN 978-1-4673-8370-7. URL http://dl.acm.org/citation.cfm?id=2906831.2906851.

[57] 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 Interna-*

tional 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.

[58] Thomas B. Sheridan. Risk, human error, and system resilience: Fundamental ideas. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, pages 418–426, 2008. URL https://www.ida.liu.se/~729A71/Literature/Resilience_T/Sheridan_2008(2).pdf.

[59] Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. Persuasive robotics: The influence of robot gender on human behavior. pages 2563,2568. IEEE Publishing, 2009-10. ISBN 978-1-4244-3803-7.

[60] Paul Slovic and Ellen Peters. Risk perception and affect. *Current directions in psychological science*, pages 322–325, 2006. URL http://journals.sagepub.com/doi/10.1111/j.1467-8721.2006.00461.x.

[61] Gerard P.A. Tan Stephen Lewandowsky, Michael Mundy. The dynamic of trust: Comparing humans to automation. *Journal of experimental Psychology Applied*, pages 104–123, 2000. URL https://www.ncbi.nlm.nih.gov/pubmed/10937315.

[62] J.Cassell T.Bickmore. Small talk and conversational storytelling in embodied conversational interface agents. 2000. URL https://www.media.mit.edu/gnl/publications/NI99_smalltalk.ps.

[63] C.D. Wickens. The tradeoff of design for routine adn unexpected performance implications of situation awareness. *Situation awareness analysis and measurement*, pages 211–226, 2000.

[64] Chenguang Yang, Hongbin Ma, and Mengyin Fu. *Advanced Technologies in Modern Robotic Applications*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 9789811008290, 9811008299.

10 Appendix A

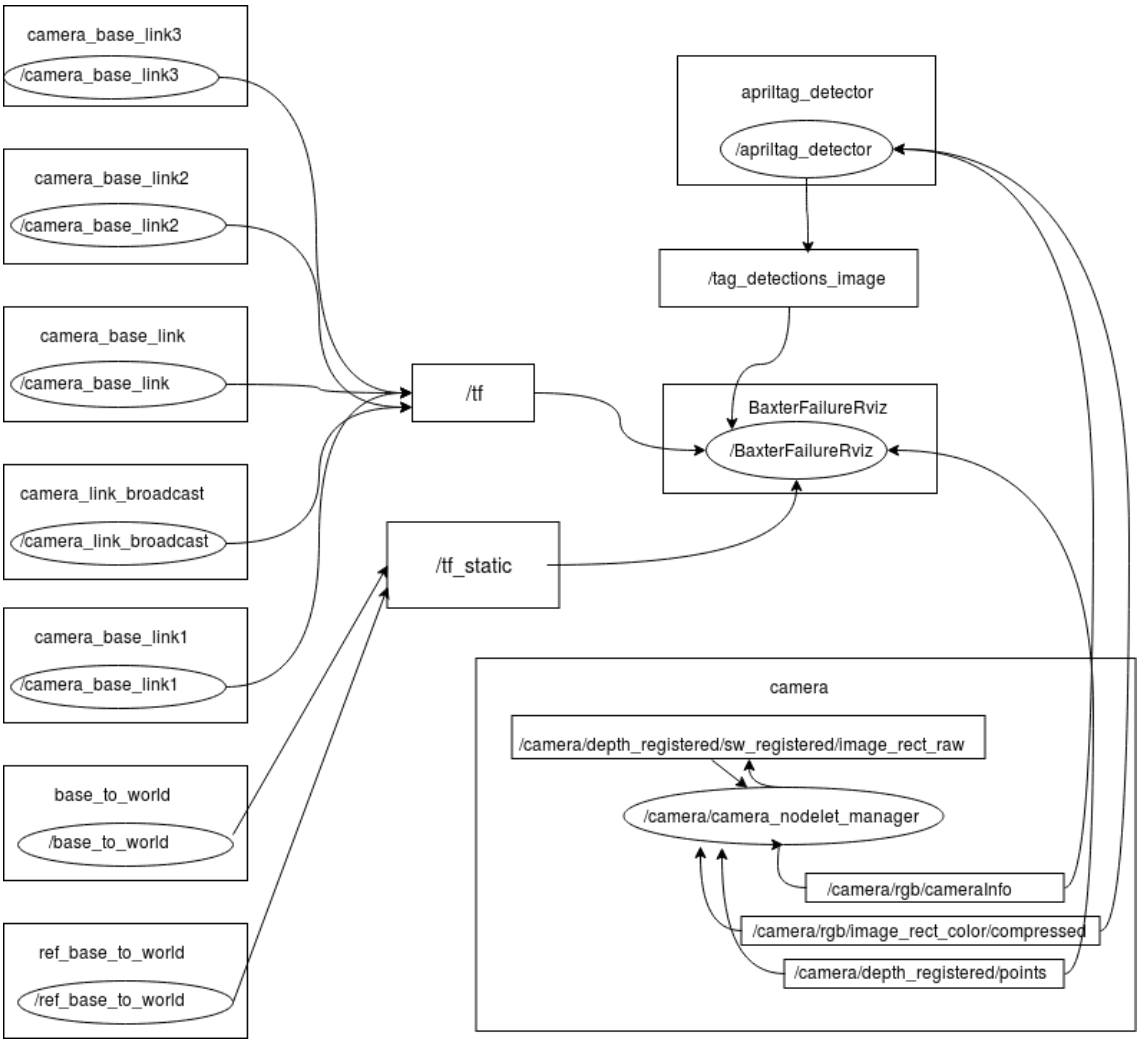

Figure 10.1: Simplified RQT Graph of the Camera System

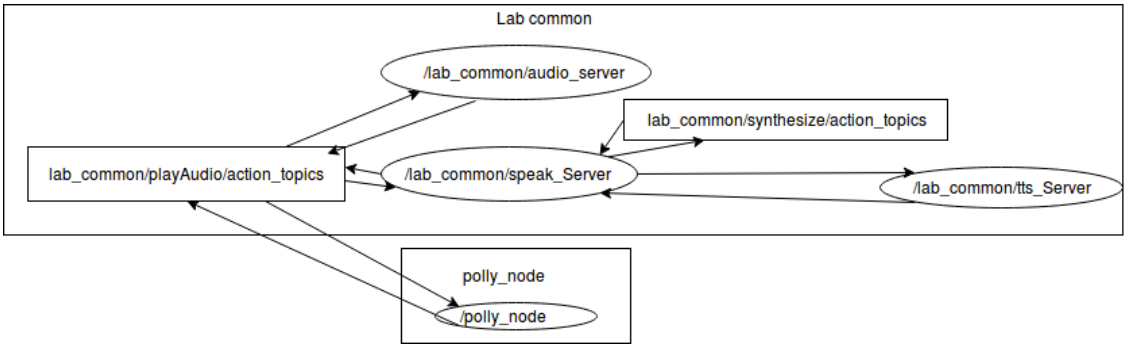

Figure 10.2: Simplified RQT Graph of the Text-to-Speech System

11 Appendix B

$$[{}^0T_1] = \begin{bmatrix} c\theta_1 & -s\theta_1 & 0 & 0 \\ s\theta_1 & c\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^1T_2] = \begin{bmatrix} c(\theta_2+90^\circ) & -s(\theta_2+90^\circ) & 0 & L_1 \\ 0 & 0 & 1 & 0 \\ -s(\theta_2+90^\circ) & -c(\theta_2+90^\circ) & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[{}^2T_3] = \begin{bmatrix} c\theta_3 & -s\theta_3 & 0 & 0 \\ 0 & 0 & -1 & -L_2 \\ s\theta_3 & c\theta_3 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^3T_4] = \begin{bmatrix} c\theta_4 & -s\theta_4 & 0 & L_3 \\ 0 & 0 & 1 & 0 \\ -s\theta_4 & -c\theta_4 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[{}^4T_5] = \begin{bmatrix} c\theta_5 & -s\theta_5 & 0 & 0 \\ 0 & 0 & -1 & -L_4 \\ s\theta_5 & c\theta_5 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^5T_6] = \begin{bmatrix} c\theta_6 & -s\theta_6 & 0 & L_5 \\ 0 & 0 & 1 & 0 \\ -s\theta_6 & -c\theta_6 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[{}^6T_7] = \begin{bmatrix} c\theta_7 & -s\theta_7 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ s\theta_7 & c\theta_7 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^{BR}T_0] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & L_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[{}^7T_{GR}] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & L_6 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [{}^WT_{BR}] = \begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & -L \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & -h \\ 0 & 0 & 1 & H \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} r_{11} = c_7(s_6(s_4(i) - c_1c_2c_4) - c_6(s_5(j) + c_5(c_4(i) + c_1c_2s_4))) \\ +s_7(s_5(c_4(i) + c_1c_2s_4) - c_5(j)) \end{aligned} \tag{15}$$

$$\begin{aligned} r_{12} = c_7(s_5(c_4(i) + c_1c_2s_4) - c_5(j)) - s_7(s_6(s_4(i) - c_1c_2c_4) \\ -c_6(s_5(j) + c_5(c_4(i) + c_1c_2s_4))) \end{aligned} \tag{16}$$

$$r_{13} = -c_6(s_4(i) - c_1c_2c_4) - s_6(s_5(j) + c_5(c_4(i) + c_1c_2s_4)) \tag{17}$$

$$\begin{aligned} r_{21} = -s_7(s_5(c_4(k) - c_2s_1s_4) - c_5(l)) - c_7(s_6(s_4(k) + c_2c_4s_1) \\ -c_6(s_5(l) + c_5(c_4(k) - c_2s_1s_4))) \end{aligned} \tag{18}$$

$$\begin{aligned} r_{22} = s_7(s_6(s_4(k) + c_2c_4s_1) - c_6(s_5(l) + c_5(c_4(k) - c_2s_1s_4))) \\ -c_7(s_5(c_4(k) - c_2s_1s_4) - c_5(l)) \end{aligned} \tag{19}$$

$$r_{23} = s_6(s_5(l) + c_5(c_4(k) - c_2s_1s_4)) + c_6(s_4(k) + c_2c_4s_1) \tag{20}$$

$$r_{31} = c_7(c_6(c_5(m) + c_2s_3s_5) + s_6(n)) - s_7(s_5(m) - c_2c_5s_3) \tag{21}$$

$$r_{32} = -c_7(s_5(m) - c_2c_5s_3) - s_7(c_6(c_5(m) + c_2s_3s_5) + s_6(n)) \tag{22}$$

$$r_{33} = s_6(c_5(s_2s_4 - c_2c_3c_4) + c_2s_3s_5) - c_6(c_4s_2 + c_2c_3s_4) \tag{23}$$

$$x_7^0 = L_1c_1 - L_4(s_4(i) - c_1c_2c_4) - L_5(s_5(j) + c_5(c_4(i) + c_1c_2s_4)) - L_3(i) + L_2c_1c_2 \tag{24}$$

$$y_7^0 = L_1s_1 + L_4(s_4(k) + c_2c_4s_1) + L_5(s_5(l) + c_5(c_4(k) - c_2s_1s_4)) + L_3(k) + L_2c_2s_1 \tag{25}$$

$$z_7^0 = L_5(c_5(m) + c_2s_3s_5) - L_2s_2 - L_4(n) - L_3c_2c_3 \tag{26}$$

$$i = s_1 s_3 + c_1 c_3 s_2 \tag{27}$$

$$j = c_3 s_1 - c_1 s_2 s_3 \tag{28}$$

$$k = c_1 s_3 - c_3 s_1 s_2 \tag{29}$$

$$l = c_1 c_3 + s_1 * s_2 * s_3 \tag{30}$$

$$m = s_2 s_4 - c_2 c_3 c_4 \tag{31}$$

$$n = c_4 s_2 + c_2 c_3 s_4 \tag{32}$$

12 Appendix C

12.1 Preliminary Survey

Session: Participant ID: Date:

Gender: ☐ Male ☐ Female Other:_____

How old are you?_____

What is your occupation/field of study?_____

What is the highest grade or level of education you completed?_____

Where are you from?_____

Do you speak English fluently? ☐ Yes ☐ No

How familiar are you with robots?

Not at all 1 2 3 4 5 6 7 Very familiar

How familiar are you with computers?

Not at all 1 2 3 4 5 6 7 Very familiar

Have you interacted with robots before? ☐ Yes ☐ No

Have you ever seen or worked with a Baxter robot before? ☐ Yes ☐ No

1. **I have experience working with robots.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

2. **I believe robots are useful.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

3. **I would be willing to work with a robot.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

4. **I think a robot is capable of bagging groceries.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

5. **I think it is safe for a robot to bag my groceries.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

6. **I believe it would be useful to have a robot bag my groceries.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

7. **I would enjoy having a robot scan and bag my groceries.**

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

12.2 Post Study Survey

Session: Participant ID: Date:

Please select a value from 1 to 10 as an answer to the following questions, where 1 = Not at all and 10 = Completely.

1. **To what extent can the system's behavior be predicted from moment to moment?**

2. **To what extent can you count on the system to do its job?**

3. **What degree of faith do you have the system will be able to cope with all systems states in the future? In other words, how much faith do you have in the system being able to do its intended job with a variety of items and environments?**

4. **Overall how much do you trust the system?**

5. **Rate the robot's performance:**

6. **I would like to interact with the robot again.**

   1 2 3 4 5

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

7. **I would be willing to work together with a robot.**

   1 2 3 4 5

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

8. **I think robots are trustworthy.**

   1 2 3 4 5

   Strongly Disagree Disagree Neither Agree nor Disagree Agree Strongly Agree

9. **I do not trust robots like I did before.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

10. **I think the robot (Baxter) is trustworthy.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

11. **I think it is safe for a robot to bag my groceries.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

12. **i think the robot is reliable.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

13. **I was willing to help the robot during the experiment.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

14. **During the experiment I felt unsafe near the robot.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

15. **I think robots are dangerous.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

16. **I think a robot is likely to fail.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

17. **The robot's behavior has harmful or injurious actions.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

18. **The robot is dependable.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

19. **I am suspicious of the robot's intents, actions, or outputs.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

20. **I expected the robot to fail.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

21. **I felt physically threatened by the robot.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

22. **Despite the failure, the robot was helpful in bagging the groceries.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

23. **The failure the robot had seemed preventable.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

24. **The failure of the robot was severe.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

25. **I would be willing to work together with a robot.**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

26. **Your level of confidence in the robot before the failure happened?**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

27. **Your level of confidence in the robot after the failure occurred?**

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

Open-ended questions, please try to elaborate as much as possible.

Did you intervene in the experiment by helping the robot? If so, how? If not, why not?

Did the failure of the robot discourage you from helping it? Please explain.

Would you be willing to have a robot helping you in your everyday life? Please explain.

Do you think the failure the robot had was an accident?

Do you think a robot can develop an intent to cause potential harm?

How can a robot let you know that something is wrong with it?

What were you thinking about when you were deciding to help the robot?

Additional comments:

Thank you for participating in the study!