UNIVERSITY OF TWENTE.

**Faculty of Electrical Engineering,
Mathematics & Computer Science**

Mitigating Unpredictable Robot Actions for Fluent Human-Robot Interaction

Leonie Zoe Sophia Prümm

Master Thesis
April 2024

Supervisors:
Dr. Bob Schadenberg
Prof. Dr. Vanessa Evers
Dr. Edwin Dertien

Research group Human Media Interaction (HMI)
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

Welcome to my master thesis report on ”Mitigating Unpredictable Robot Actions for Fluent Human-Robot Interaction”. This is the full report of my research project, which is based on previous scientific literature research in this area carried out during the ”research topics”.

Throughout my studies, I have realized my particular interest in robots. Although I think it is difficult to work with social robots in my home country, Germany, in the near future, I decided to participate in HRI research through my master's project. I made this decision because I think that the introduction of robots into everyday life is a very difficult task that needs to be critically questioned and carefully done. Technologies are developing fast and that is good, but these developments need to be viewed objectively by people who do not have economic benefits in mind: scientific researchers. Otherwise, in my opinion, we risk the future and key elements of society and civilization.

I owe a deep debt of gratitude to my supervisor, Dr. Bob Schadenberg, whose guidance and support have been crucial throughout my journey. His expertise and mentorship provided me with the tools I needed to effectively navigate the complexities of my graduation project, which at times felt like an impossible task. In addition, he was always receptive to my mental and physical well-being issues resulting from a snowboarding accident, and gave me the time I needed to address them. Furthermore, I would like to thank Prof. Dr. Vanessa Evers for her expertise in the critical phase of writing this report. I would also like to thank the Interaction Lab at the University of Twente for providing the Pepper research platform and the project rooms for working on this project. I would like to name my big gratitude to Luc Schoot Uiterkamp who had always a good tip when troubleshooting the robot. A big thanks to all the participants. Your reliability was a gift. Finally, I would like to thank my friend Vanessa Markos for providing accommodation during the experiment execution and testing beforehand. Without you, writing this thesis while living in Cologne would not have been possible.

Summary

This master's thesis explores the mitigation of effects of robot unpredictability on trust from the perspective of ”predictive coding”. The introduction of robots into people's daily lives holds great potential but also presents challenges, as successful interaction and collaboration require trust in robots. An important aspect for forming trust in a robot is its predictability. However, predictability is not always possible. Sometimes robots will act in ways we do not expect, and this can affect how much we trust them. However, there are ways to lessen these unexpected effects.

This report firstly discusses the scientific lense ”predictive coding”, taken in this study. Predictive coding is an approach from neuroscience that describes how people predict the behavior of others. The theory describes the brain as an inference machine. To form expectations, the brain uses a hierarchical forwarding model that compares sensory inputs with what is already known about the situation – internal models. By comparing information from the internal models and the actual sensory input, expectations are formed and rules of behavior are learned. In this report, the theory is for human-robot interactions operationalized as a learning process. In this process, a robot's predictability becomes more important over time after being consistent in its behavior. As a reason for this we identify the observer's high confidence in the robot model, which developed through learning the rules and structures of its behavior.

Reducing this confidence in the model is identified as our main goal. To achieve that, two time slots are defined: Before the interaction with the robot and before the unpredictable action. Per time slot and based on implications from current related research, a mitigation strategy is developed. This results in two applied strategies: foreshadowing movements before the robot acts unpredictably and informing the observer about changes before interacting with the robot.

These strategies are tested in a quantitative study. An experiment is conducted in which participants experience unpredictable robot behavior and evaluate the robot's predictability and trustworthiness afterwards. The statistical analysis does not provide evidence for the effectiveness of the applied strategies. This does not necessarily mean that the approaches should be completely discarded, but rather that changes to the experimental design should be considered. We eventually discuss what we learned from this study and how we can address the problem better in future research.

Contents

List of acronyms

HHI	Human-human interaction
HRI	Human-robot interaction
C-Predictable	Experiment condition 1: only predictable behavior
C-Unpredictable	Experiment condition 2: only unpredictable behavior
C-Text	Experiment condition 3: unpredictable behavior & strategy A
C-Foreshadowing	Experiment condition 4: unpredictable behavior & strategy B
C-Both	Experiment condition 5: unpredictable behavior & a text & foreshadowing
ANOVA	Analysis of variance
MANOVA	Multivariate analysis of variance
MANCOVA	Multivariate analysis of covariance

Chapter 1

Introduction

Today, the potential for utilizing robots in social tasks is being investigated in various sectors. One of these sectors is healthcare. Societies are aging [1]. Thus, the need for medical treatments increases. At the same time, healthcare facilities struggle to satisfy the need with sufficient employees [2]. Covering this need with robots offers hope in the daily work of doctors and nurses [3] [4].

However, using robots for social interactions in healthcare facilities presents challenges. Interacting with robots requires trusting the robot [5] [6] [7], similar to interacting with humans [8]. People need to be sure that a robot is willing and capable of protecting their interests to follow its suggestions or accept given information as valuable [6]. A key component of robot trust is its predictability [5] [6]. Predicting the actions of others is a part of daily life for everyone. People are constantly trying to understand what others are going to do next [9], because understanding what others will do is fundamental for successful (social) interactions [10]. While predictability is an important factor for robot trust, it is not always given. In many situations, unpredictability can be unavoidable [11], for example when a robot has new functionalities. Thus, it is important to look into the effects of unpredictability in human-robot interactions (hereafter "HRI").

Research on designing for predictability goes mostly in the direction of motion predictability [11]. Thus, it requires further scientific intention [11] [12] [13] [14] [15]. There is currently a lack of scientific research on the effects of unpredictable behavior. As a result, the research question for this master project is:

> **RQ:** How can the effects of unpredictable robot behavior on trust in the robot be mitigated?

The research question was further developed and addressed through an experimental study. In the study, participants experienced a task consisting of unpredictable behavior. The expected effects of this behavior on trust were addressed with mitigation strategies. Participants were then asked how they perceived the robot. This report follows a structured format, beginning with an introduction and explanation of the project's motivation in this first chapter. In the second chapter, the background for understanding the executed study is given 2. In chapter 3, the research question is fully developed. This includes deciding on timing and potential strategies to deal with the effects of unpredictability. In chapter 4, the developed study to answer the research questions is described. The results of the study are in chapter 5 presented and in chapter 6 discussed. Eventually, a conclusion is made in chapter 7.

Chapter 2

Background

This chapter develops the background for understanding the research question. First, the relationship between trust and predictability is explained in section 2.1. Here, we decide on a trust measure as well. Second, predictability in HRI is discussed as a complex concept to define the term for this work in section 2.2. In section 2.3, based on the theory ”predictive coding”, it is explained how people predict the actions of others to create an understanding of the scientific perspective taken in this research project. Eventually, predictive coding is operationalized for HRI in section 2.4.

2.1 The Relationship of Trust and Predictability

Many academic fields, including sociology, psychology, philosophy, and neuroscience, have studied the concept of trust. Predictability is critical to trust in human-human interactions (hereafter HHI). In the early stages of scientific trust research, Rempel et al. [8] define trust as an interplay of dependability, predictability – the consistency of someone's behavior over time – and faith. They identify predictability as a key component for developing trust in someone.

Predictability is important not only for developing trust in humans but also for developing trust in robotic technologies. The dynamics of trust in humans and automated trustees are similar [16]. Lee and See [5] identify three components of trust in automation: process, performance, and purpose, with predictability as a key component of the performance attribute. In HRI, predictability is a facet of trust in most concepts as well [17]. In a meta-analysis, Hancock et al. [6] identified predictability as a critical factor for developing trust in robots. Predictability is also related to the robot's performance attributes. Performance factors greatly influence perceived trust in the robot and should be considered when designing robots regardless of context.

How trust evolves from predictability in detail is not yet finally defined. Lewis et al. [16] identify three aspects of trust formation: predictability, dependability, and faith. These aspects exist in three phases of robot trust: formation (i.e., developing trust over time), dissolution (i.e., lowering trust after a violation), and restoration (i.e., developing trust after a violation). They define predictability as a component of early trust development before dependability and faith develop. Thus, predictability predicts trust in a robot in the early development stages.

2.1.1 Evaluating Human-Robot-Interactions

While many researchers measure trust, it is a multi-dimensional concept that is under-theorized in HRI research [18] [19] [20]. Thus, there is no standardized concept and no resulting standardized and validated set of measures for trust. This results in confusion about the term itself (within studies and when comparing them [18] [19] [20]), the goals for which researchers aim (i.e. relation-based or performance-based trust [19] [20]) and the necessary measurement approaches to evaluate these goals (i.e. subjective or objective measures [19] [20]). Some measures are used repetitively, but they are not executed in a standardized way [20]. When trust is defined, measures often do not evaluate the defined aspects of trust, but other [18]. As a result, many studies measure different aspects but call them trust [20]. This demonstrates a need for standardization and alignment to an agreed-upon concept and a resulting model and validated measure.

When comparing robot trust questionnaires today, two seem to be mainly used currently in scientific research: ”The Multi-Dimensional Measure of Trust” (MDMT) questionnaire by Malle & Ullmann [7] and the ”Trust Perception Scale-HRI” by Schäfer [21]. In Malle & Ullmann [7], trust is understood as a multi-dimensional concept that incorporates moral and performance aspects with a specific focus on social tasks. These aspects are explained to be reasonable based on findings in human-automation and human-human trust research. Schäfer [21]'s approach is more general. The goal was to develop a trust scale that addresses all robot domains and tasks while incorporating trust-relevant environmental, human, and robot elements [21]. For the purpose of this study and based on current discussions on the multi-dimensionality of trust, from our perspective, following Malle & Ullmann [7]'s approach is reasonable. In the next section, we discuss predictability, as we aim to mitigate the effects of unpredictability in this project.

2.2 Predictability in Human-Robot Interactions

Before we can discuss predictability, the term first needs to be defined. Defining predictability in HRI is complicated. This is because the topic is discussed in various scientific fields and is used ambiguously [11]. Predictability can be analyzed and optimized using multiple approaches. Thus, there is no standardized concept of predictability in scientific research today. While some researchers define what they mean by predictability in their work, others use the term without discussing its ambiguities.

According to Dragan et al., [13] [12], expectations regarding robot behavior vary greatly based on people's subjective perceptions. They describe two terms: predictability and legibility. Legible motion allows deriving the goal from (a part of) the motion. The quicker this goal can be inferred, the higher the legibility of the motion is. Contrary to that, predictability is defined as the quality of an action to match an expected action. Accordingly, a movement is predictable if it matches the expected movement – the higher the match, the higher the predictability. They argue that legibility and predictability stem – especially in highly ambiguous situations – from opposing directions: predictability follows the concept of "action-to-goal" by deriving the goal from the ongoing action, while legibility follows "goal-to-action" by predicting necessary actions based on the goal. They state that legibility and predictability are "fundamentally different" and can be contrary aspects of robot motion. Higher legibility can result in decreased predictability and vice versa.

Lichtenhähler et al. [22] [15] investigate the concept of predictability as well, naming Dragan et al.'s legibility "goal-predictability" and predictability "trajectory-predictability." They define legible motion as the overall goal. Legible robot behavior is achieved from their perspective if both goal-predictability and trajectory-predictability are given. As a result, robot behavior is legible if an observer can understand intentions and predict the goal with high confidence [22] [15] and accuracy while the executed behavior meets the expectations of the observer [15]. Unlike Dragan et al., they do not perceive legibility and predictability as opposing attributes of robot actions [15].

Schadenberg et al. [11] describe predictability as a dynamic aspect of a robot that is influenced by its perceived novelty and experiences. From their perspective, robot behavior is predictable if the human observer can learn quickly and accurately how to predict the robot's behavior. The level of predictability is determined by the action's structural regularities. Structural regularities in behavior allow the observer to infer rules from behavior that can be used to predict behavior in future interactions. It is not only relevant whether these structural regularities exist by design, but also whether they can be perceived by the observer. Their definition of predictability differs between behavioral and attributed predictability. The difference between the two is that robots that behave predictably are not necessarily perceived as predictable. Robots with high behavioral predictability are programmed to be highly predictable, i.e., to do what is understood to be predictable. Robots with high attributed predictability are perceived as highly predictable by their human observer. They state that both concepts are related but do not necessarily lead to the same design implications. From their perspective, the relationship between behavioral and attributed predictability depends on the interaction context, thus on aspects like familiarity and previous experiences with the robot. For example, somebody who worked with the robot often may perceive unexpected actions more negatively than somebody who is newly introduced to the robot and does not yet know what to expect.

The variety of used terms shows the current challenge of agreeing on a common concept. This missing shared understanding results in different approaches to design for predictability in HRI [11]. To address this issue and with the goal of contributing to a transparent, standardized term in the future, we do not define a new term for this project. For this work, our understanding of predictability follows the definition of Schadenberg et al. [11]. Thus, predictability is a dynamic robot characteristic that depends on the context, the human user, and their earlier experiences with the robot. Predictable robot motion contains structures and rules that can be interpreted by a human observer as typical for that particular robot. Robot behavior can be predictable by attribution through the observer and by behavior design [11].

2.3 How Humans Predict

It is necessary to understand how humans predict the actions of others before strategies to address this process can be developed. Interacting with others relies on understanding what they will do next. This is achieved through brain signals in the entire cortex, using lower cognition areas for low-level processes like vision and higher cognition areas for high-level processes like intention attribution. Various theories try to explain how people predict the behavior of others. An influencing theory from the area of neuroscience is "predictive coding" [23] [24]. The concept of predictive coding is heavily influenced by Rao and Ballard's [23] theory that the brain contains internal statistical models of the world and a hierarchical structure of predictions. Predictive coding is today a dominant concept in cognitive neuroscience [24] and has become more relevant in other scientific fields like philosophy as well [25].

Rao and Ballard [23] describe that the brain compares sensory inputs with internal models (synonymously called 'generative models'). Models exist on different levels of abstraction. On every level, the brain compares (multi-)sensory input with the most probable model and forwards mismatches (called 'prediction errors') to a higher hierarchical level (see Figure 2.1). Every time, the estimate is corrected by using the error signal. As a result, higher-level information influences what is estimated on lower levels, while lower-level models' errors influence how higher-level models are updated [23]. This system is embedded into a higher hierarchical system

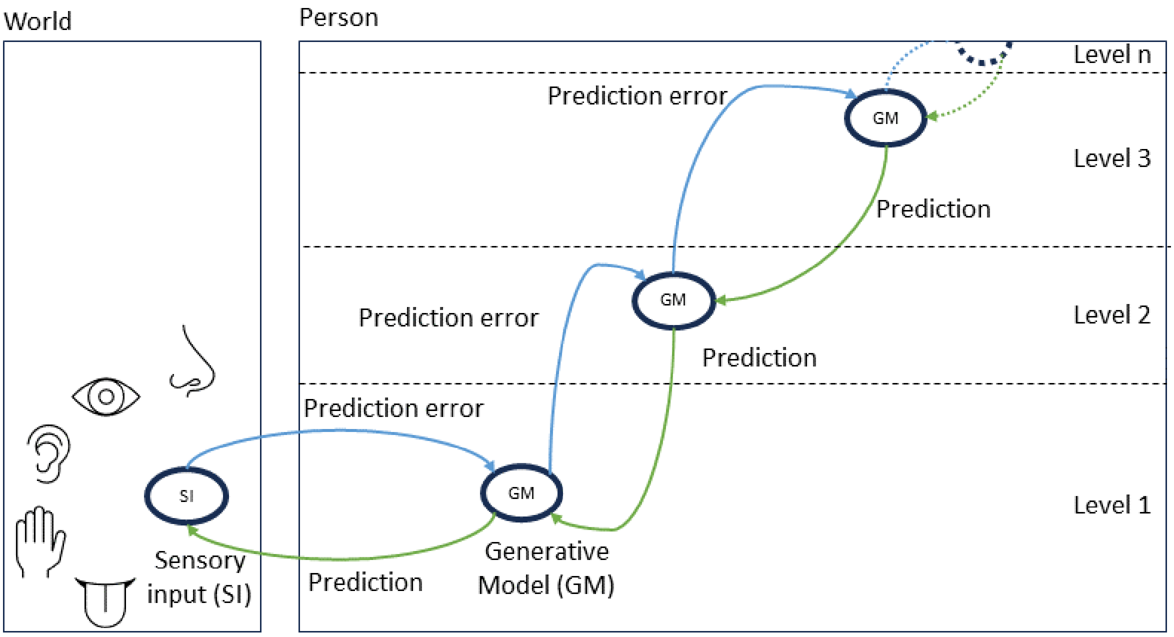

Figure 2.1: Predictive coding

that processes information about the surrounding context. As a result, the sensory input is not only compared to what is known about similar inputs but the processing itself is influenced by the context [26]. Growing evidence shows that social knowledge, like stereotypes and biases, influences how we deal with sensory input in the early processing stages. Thus, even when not knowing something, it can be expected that the brain processes sensory input after perceiving a stimulus by comparing it with prior knowledge, forming a final perception result. This is in contrast to classical models of perception, which assume bottom-up processing of perceived information without any influence from higher cognitive levels [27].

Predictions based on internal models are not necessarily correct, as they can come from incomplete, immature, or downright wrong internal models, for example, resulting from lacking knowledge about the experience. This is especially true when something is experienced for the first time since then, no well-developed model exists. In most cases, prediction errors impact an internal model when the error corrects the model. However, there are prediction errors that do not result in updating the internal model due to the stochastic nature of specific cases. For example, when tossing a fair coin, the experience that the result was other than predicted is still in line with experiences and expectations about the 50-50 chance to toss each side [28]. Thus, prediction errors differ in their impact on internal models. Whether the prediction error influences the model building depends on the error's confidence. How confident somebody is about an error depends on the cause of the uncertainty, so if the uncertainty is reducible (e.g., by increasing knowledge) or irreducible (e.g., because the situation is naturally the way it is, like the fair coin with a 50/50 chance for each side). This is called the precision of the prediction error. The higher the precision, the higher the confidence that the error occurred [28].

Precision does not only exist for errors but for internal models as well. The precision of the prediction describes the confidence about the outcome. As a result, in new situations where one does not yet know what to expect, internal models have low precision [28]. Consequently, unexpected behavior is less impactful than when one is aware of a person's typical behavior over an extended period [11].

Another theory closely related to predictive coding is the free energy principle. They are closely related because the free energy principle says that the brain tries to match the perceived reality to its internal models and reduce free energy as much as possible. Free energy is the surprise and uncertainty resulting from internal models not matching reality. Karl Friston, who proposed the concept, states that humans strive to reduce free energy by actively adapting the world to their expectations, to maximize the probability that the internal model is correct [29].

As learned in this section, people predict their surroundings by comparing what they perceive with the experiences and learnings from previous experiences. This manifests in internal models. The brain always tries to match experiences to the most probable model. With every new experience, internal models update with the help of the error signal and its precision. At the beginning of a learning process, models have low precision – the perceived person is unpredictable. When knowing somebody well and this person behaves in a consistent manner, models become more detailed and have higher precision – the person becomes more predictable. This concept is mapped onto interactions with robots in the following.

2.4 Modelling Predictability from the Perspective of Predictive Coding

When applying predictive coding to interactions with robots, it can be assumed that similar processes happen. In the beginning, when somebody interacts with the robot for the first time, they do not know what to expect. Since related models are vague and the precision that errors are correct is high, the perceived sensory input has a strong influence while the models do not influence the overall experience (see Figure 2.2: Initialization phase). Still, it can be assumed that there are models to which the brain tries to relate, as it always tries to reduce free energy [29]. As a result, it can be assumed that no acquaintance process starts without any referred model, but there is a related experience to what the brain refers to (see Figure 2.2: Related experience).

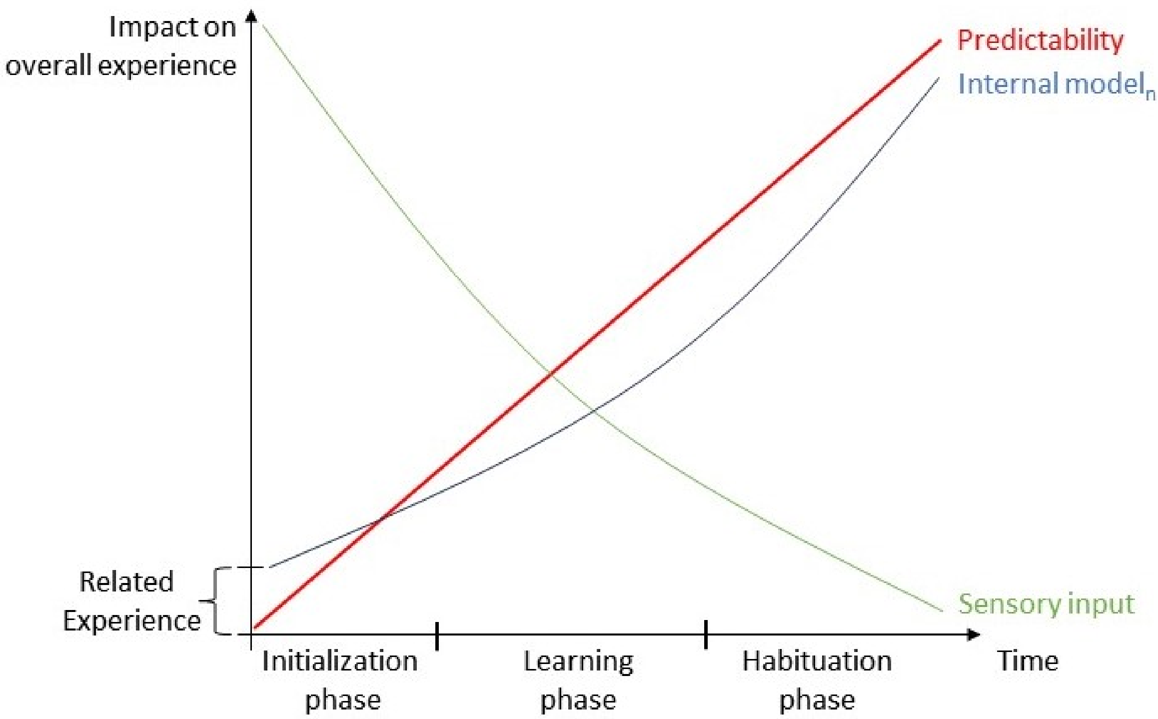

Figure 2.2: Influence of Internal Model, Sensory Input and Predictability of a Robot on the Overall Experience

When interacting further with a robot, assuming that it always behaves consistently, the model of the robot becomes clearer. Here, the influence of the model on the robot currently being built increases with each interaction in which it shows structural regularities [11]. Simultaneously it can be expected that how the robot acts has less impact on the overall experience (see Figure 2.2: Learning phase).

When interacting with a robot for a longer period, assuming that robot behavior is perceived similarly over time, how the robot acts has only low influence since the observer strongly relies on the developed model after experiencing consistent behavior over time. Suppose now, the robot acts differently and is, as a result, unpredictable. In that case, we can assume that this can negatively affect the overall interaction (see 2.2: Habituation phase) and thus trust as well.

Thus, the internal model, which is re-evaluated with each interaction, can be described by

$$model_0 = relatedExperiences_{t<0}$$

$$model_{n+1} = model_n + predictionError_{t>0} * precision$$

where "relatedExperiences" describe the previous experience before the situation, consisting of anything relatable, such as seeing robots in movies. The "prediction-Error" describes the new incoming information with each interaction, starting with the first interaction. Its influence depends on the confidence that it is correct, reflected by ”precision.”

To sum it all up, it can be assumed that the influence of predictability on the overall experience with a robot increases over time when experiencing consistent robot behavior with structural regularities. Since predictability is connected to trust, it can be assumed that trust is lowered when a robot behaves unpredictably after being used to it. Unpredictability is expected to have the most impact when being used to a robot. As a result, it is most reasonable to intervene with strategies for dealing with the adverse effects of unpredictability when unpredictable behavior has the most impact: in the habituation phase. Here, the robot model is detailed, and the precision of this model is high.

To minimize the impact of unpredictability on trust, we aim to set the observer back into a state where we do not expect predictability to matter: the initialization phase. This phase requires low precision of the robot model. Thus, we aim in this project to mitigate the effects of unpredictability by reducing the confidence (i.e., precision) in the expectation about the robot (i.e., the internal model). To the best of our knowledge, no similar approach aims to achieve this goal with interaction and robot design techniques. In chapter 3, we describe how we address this goal.

Chapter 3

The Problem Statement

After explaining the necessary background, this chapter further develops the research question. This includes discussing the most reasonable part of the interaction to intervene with mitigation strategies. In section 3.1, this is accomplished by combining predictive coding with implications from the health care use case and existing research in HRI. After discussing when strategies should be applied, it is discussed where potential mitigation strategies could come from in section 3.2, resulting in three specific sub-research questions.

3.1 Intervening with Mitigation Strategies: The 'When'

As discussed in section 2.4, predictability has the lowest influence in the early stages of model development and the highest when models are well-developed. Thus, it is aimed at intervening in the habituation phase with the goal to let the observer rely less on models, like in the initialization phase. This is expected to be achievable by reducing the robot model's precision.

Before discussing strategies to achieve this goal, a decision must be made on when to apply mitigation strategies in the explicit situation with the robot. Timing is an essential aspect of developing strategies to deal with the consequences of unpredictability since the timing of an intervention can affect how trust the robot is developed [30].

To discuss applicable strategies, we assume two phases surround unpredictable behavior: The action and the interaction layer. Behavior that occurs in the vicinity of unpredictable behavior, i.e., immediately before, during, or immediately after, occurs in the action layer. Behavior that occurs during the interaction but is separated from the unpredictable action happens in the interaction layer (see Figure 3.1).

Addressing every time slot is neither feasible due to this project's scope nor reasonable. Much research has already been conducted, although with slightly different

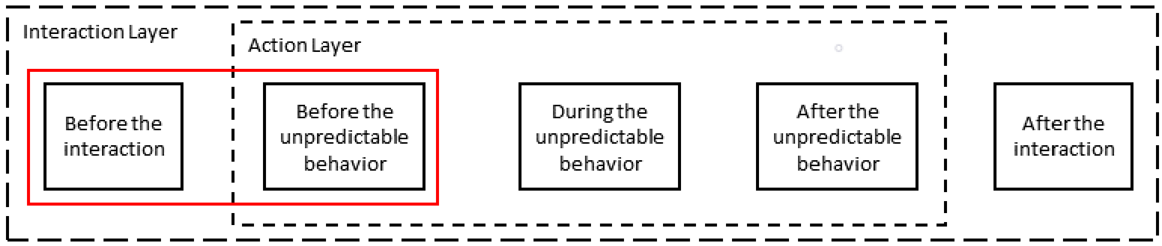

Figure 3.1: (Optimal) Time slots to intervene in with behavior design strategies

objectives, as the definitions of predictability vary. Scientific research addresses the time slot during the unpredictable action by surrounding the action with motions to increase predictability [13] [12] [9]. Intervening after unpredictable behavior has occurred is also a topic in research [31] [32] [33]. Especially the area of trust repair is a broadly investigated scientific field where repair strategies are mostly based on apologies, explanations, denials, promises [30], regretting and expressing reparation [32] [33]. To the best of our knowledge, there is a research gap in preventing the effects of unpredictable behavior in HRI.

The project is being carried out within a hospital use case context. In this use case, users vary from children to seniors, including those who have never seen a robot before and those familiar with robotic technologies in real life and entertainment. It can be assumed that all users have in common the experience of health problems that require medical treatment or visiting close relatives who are receiving medical treatment. These situations can lead to mental stress. Given the specific use case of this project, it is important to anticipate that stress may arise as a mental state during interactions with the robot. Being stressed needs to be considered since when stressed, individuals often withdraw from social interactions and may exhibit irritability and hostility [34]. Therefore, the observer may be less receptive to positive responses, such as robot apologies.

Thus, for this project, we address the two time-slots ”before the interaction” and ”before the unpredictable behavior” with strategies dealing with the effects of unpredictability (see Figure 3.1). Choosing the specific time windows leads to the question of which strategies can be used to address the precision of an existing model. This is discussed in the following sections.

3.2 Intervening with Mitigation Strategies: The ’How’

There are several opportunities to address the issue of defining strategies. This section discusses related work from current HRI research and approaches based on bias research, resulting in two mitigation strategies applied in the following study.

3.2.1 Reducing Unpredictability through Behavior Design

As introduced in chapter 1, research on predictability is mainly from robot motion design research. Especially for isolated, functional motions (e.g., moving an arm towards a box or grasping), using additional motions to signal the robot's actions and improve understanding is a well-established approach [13]. As researchers define predictability differently (see section 2.2), approaches to address it aim at slightly different terms. Takayama et al. [9] aim to increase the readability of actions, defined as more accurate and confident descriptions of subsequent actions, using pre- and post-actions as forethought. Their results show a positive influence of foreshadowing design on actions' readability. Forethought robot behavior consists in their study of the expressive elements engagement, confidence, and timing. Engagement is achieved by taking small steps forward. Confidence is expressed by raising the torso slightly. Timing is used when the expressive part is performed just before the functional movement. Dragan et al. [12] investigated adding and changing robot motions in different studies. They conclude that functional, efficient actions should be accompanied by foreshadowing movements to increase legibility, as this positively influences cooperation with humans and the overall perception of the robot in tasks that require close collaboration. Camblor et al. [35] aim to enhance predictability through situational awareness during collaboration with robots in an industrial work setting. Situational awareness is achieved from their perspective when humans correctly perceive the environment in terms of time and space, understand the current state, and can predict future events based on their own perceptions. The study found that adding elbow tilting to a woodworking task improved predictions about future robot actions.

When choosing an approach, it is essential to consider that different methods are suitable for various use cases. Differences in users, the situation, the task, its context, and the robot's specific abilities must be considered when designing mitigation strategies. Users can strongly differ in their needs (e.g., autistic children [36]), their capabilities to interact with the robot (e.g., deaf and hard of hearing people), their cultural background (e.g., different socially accepted behaviors) and their prior experiences with robotic technologies and this robot in particular. The context of the interaction (e.g., a dangerous factory area, other people, noisiness) can affect the ease of the interaction, the possible robot features, and how the robot is perceived. Moreover, the robot design itself influences which strategies can be applied. A robot might not have a speech module or arms to create specific behaviors.

From the predictive coding perspective, it can be assumed that generative models are associated with specific sensory inputs. The idea at hand for mitigating the effects of unpredictability might be adding more sensory inputs. Adding more sensory inputs could create a discrepancy between the model and reality large enough to rely on another model since, as Friston [29] states, the brain refers to the most probable one. While this might be possible, we assume that actively creating a strong mismatch between model and perception could also cause a high prediction error and a negative user experience. This remains unclear. Thus, we expect adding completely new sensory inputs is not reasonable as a mitigation strategy based on robot behavior design. Instead, we aim to adjust our current capabilities by following the idea of foreshadowing. Foreshadowing movements are – as best to our knowledge – used in scientific research to increase predictability, but not with the goal of decreasing the effect of unpredictable actions. Furthermore, the research focuses on foreshadowing independent, small movements instead of holistic robot capabilities. As a result, a sub-research question of this research project is:

> **RQ.a:** To what extent can a behavior, added in the action phase, influence the negative effects of unpredictability on trust through foreshadowing the ability to move?

3.2.2 Reducing Unpredictability through Applying Anti-Bias Research

Another approach investigated in this research project is applying the goals of anti-bias strategies to the interaction. Anti-bias strategies educate individuals to avoid relying on unconscious bias actively. It seems reasonable to expect a relationship between the formation of unconscious (also known as cognitive) bias and predictive coding since bias can cause predictions [25]. Until today, the relationship between predictive coding and cognitive bias has not been thoroughly researched, but current research indicates that similar processes take place [37] [38] [39]. Biases shape the way we perceive, think, and act. When the brain develops complex models through predictive coding, it takes – since prejudices exist in every society – biased information into account. Thus, training a predictive brain with biased information is similar to training AI algorithms with biased data: The result will be biased, too [39].

The forwarding process in predictive coding can potentially introduce and further develop bias. This can happen if the confidence in the bias (i.e., the precision of the model) is strong because repeated stimuli approve it often. An example of this could be repeatedly reporting on terrorist incidents and implying that all terrorists come from a particular country. This would be a biased and inaccurate assumption. Additionally, models are typically corrected for errors. When the error signal is too weak or its precision too low – e. g., by not attending enough to the situation or due to the influence of emotions such as fear on the error signal –, what actually happens is perceived, but the error is not strong enough to alter future predictions. Thus, the observer relies on the bias, no matter that they perceived contrary information [37]. Biases might not be challenged but self-confirmed with every perception cycle [39].

Especially in work environments, various approaches are used to educate people on their biases. Companies spend billions to minimize the consequences of biased thinking and decision-making [40]. Current research suggests some promising directions for the reduction of bias. One option to reduce bias can be becoming more aware of the existence of this bias [41] [42] [43] [44]. Another option to reduce bias is by informing a person about the advantages of lowering bias [45] [41] [43]. Also, fostering curiosity about new aspects can lead to a reduction of bias [45]. Setting expectations and providing explanations, as done in bias research, is a technique for reducing the impact of robot mistakes on trust in HRI as well. One approach is setting people's expectations of the robot's (in-)capabilities in advance [46] [47]. Lee et al. [32] followed a similar approach by informing participants beforehand about robot (in-)capabilities, resulting in a more positive perception of the robot.

Based on this, for this project, we assume that biases are similar to high precision internal models trained with inaccurate input. Suppose this is the case; this would mean that there is a strong belief in bias, which results in a biased model with high precision. Anti-bias trainings aim to reduce bias. Thus it may be possible that approaches from anti-bias research might be able to trigger change in internal models (see Figure 3.2). Furthermore, current research from HRI [46] [47] [32] indicates that informing about potential changes can be a reasonable. As a result, a sub-research question of this research project is:

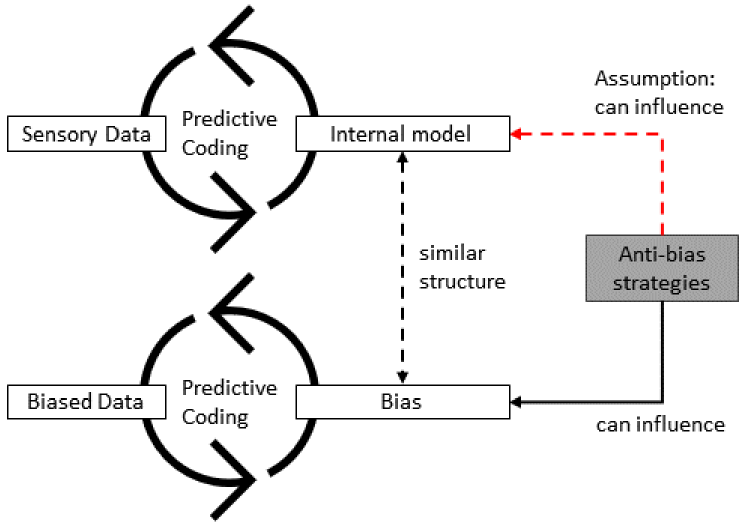

Figure 3.2: Assumed potential influence of anti-bias strategies on internal models

RQ.b: To what extent can an anti-bias-training inspired explanatory text, used in the interaction phase, influence the effects of unpredictability on trust?

Combining an explanatory text with foreshadowing could be useful since the strategies might act complementary when applied in different interaction phases. As a result, the third sub-research question of this research project is:

> **RQ.c:** To what extent can the combination of an explanatory text and a foreshadowing movement influence the effects of unpredictability on trust?

We expect that applying the strategies will positively influence the impact of the unpredictable action. This influence is expected to be less severe, resulting in significantly higher ratings in moral and performance trust. Furthermore, we hypothesize that this happens through increasing robot predictability. Thus, the observer attributes significantly more predictability to the robot than without the text or foreshadowing. Finally, we expect that combining the two approaches will increase this positive influence.

As discussed earlier, we address the time before an unpredictable action happens. It is difficult to determine when the learning phase stops and the habituation phase begins. Thus, while still aiming for the user to be in the habituation phase, participants are observed in an experiment when accustomed to the robot. In the next chapter, it is described which approach is applied to answer these research questions.

Chapter 4

Materials and Methods

This chapter describes the methodology of the executed study. Section 4.1 describes the participant sample and ethical considerations. Materials and methods are explained in section 4.2. This includes a description of the used robot. Then, the interaction design is in section 4.3 described. Here, we explain the experiment task, how we developed it, and which strategies we applied to mitigate the expected effects of unpredictability. In section 4.4, we describe the applied research design. Then, in section 4.5 and 4.6, we describe which setup we used during the experiments and how we executed them. In the last two sections, we describe the measurements (see section 4.7) and the data analysis approach (see section 4.8).

4.1 Participants

The participant sample consisted of 25 people: 15 were male, nine were female, and one identified as diverse. 96% were younger than 34[1] (see Figure 4.1). The majority of the participants was studying[2] and had completed higher education[3]. 52% of them are studying robotics-related fields[4] and 32% study other STEM[5] studies[6]. Their disposition to trust in technologies was high (M = 3.92, SD = 0.58[7]). There is no visible tendency for a negative attitude towards robots, as it is neutral in the

[1] Age 18-24: 56%, 25-34: 40%, 35-44: 4%

[2] Studying: 88%, employed for wages: 8%, looking for work: 4%

[3] Bachelor's degree: 68%, Master's degree: 16%, High school degree: 16%

[4] Interaction Technology 48%, Biorobotics 4%

[5] The natural sciences including chemistry, physics, biology, mathematics and all its derivative disciplines.

[6] Biomedical Engineering 4%, Mechanical Engineering 4%, Business Information Systems 8%, Data Science 4%, Electrical Engineering 4%, Systems, Control & Biomechatronics 4%, Technical Medicine & Psychology 4%

[7] disposition to trust was measured on a scale from 1 (low disposition) to 5 (high disposition).

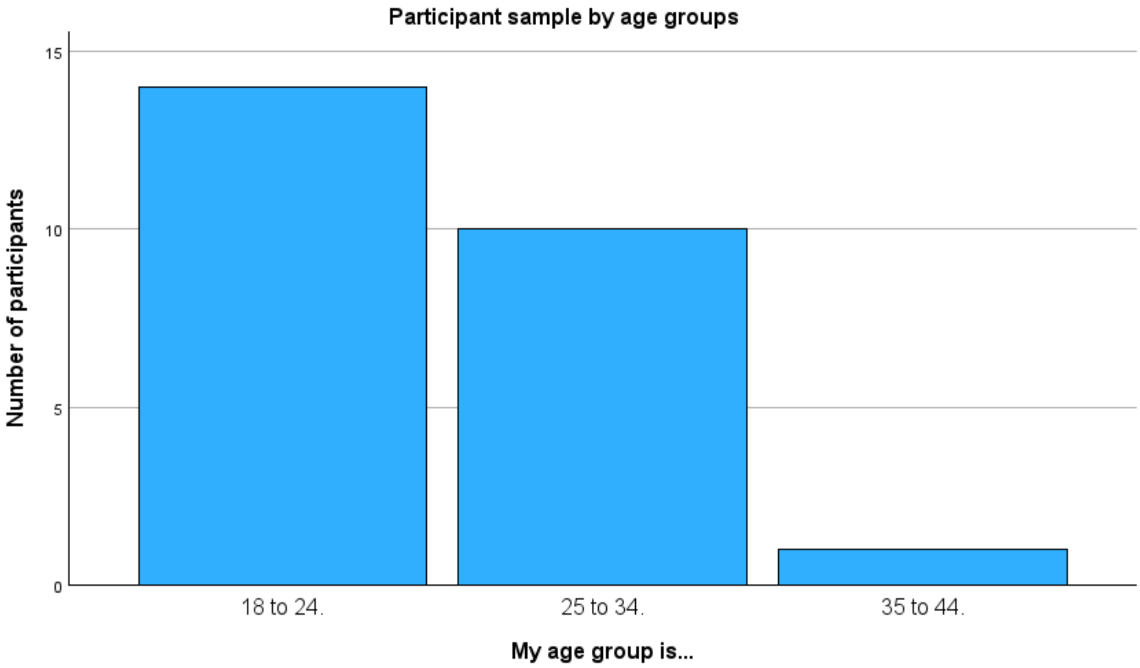

Figure 4.1: Participant ratings by age groups

sample as a whole (M = 2.64, SD = 0.47[8]). Robot experience is also high (M = 4.48, SD = 1.16[9]). Thus, the average participant tends to trust technologies and has already interacted with robotic toys or social robots. This is consistent with the sample group's current jobs, education levels, and age. The participant sample is young, tech-savvy, and well-educated.

The ethical integrity of this research project has been reviewed. This was achieved by submitting the experimental settings to the University Ethics Committee (application number: 240013). The informed consent form used was also submitted for review. Positive advice was given on January 25, 2024, and the experiments were carried out after the ethical review of the procedure. The participants were informed about the study and signed the consent form before the experiments took place.

Participants were recruited following positive ethical advice. First, social media channels were used. Two study WhatsApp groups were used with 58 and 261 people, respectively, resulting in 16 participants. Second, people from the "EEMCS Graduation Support Group" were asked, resulting in two participants. Third, the Canvas course "Recruitment of participants for I-Tech" was used, and the topic was added there, resulting in one participant. As no further participants could be found

[8]Negative attitude towards robots measured on a scale from 1 (positive attitude) to 5 (negative attitude).

[9]Robot experience was measured on a scale of 0 (no experience) to 6 (robot programming experience).

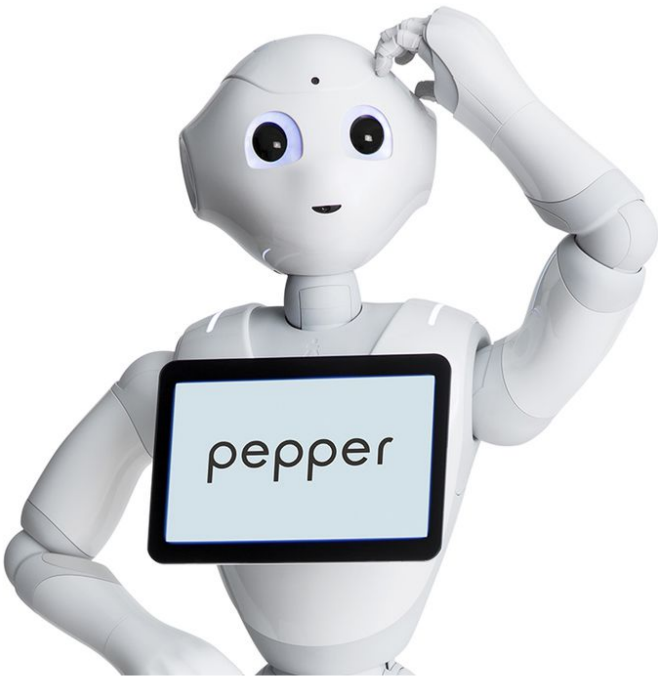

Figure 4.2: Pepper robot

through these channels, six participants were found by asking around the building and campus.

4.2 Materials

For this study, we used the ”Pepper” robot. Pepper (see Figure 4.2) is a humanoid robot developed by SoftBank Robotics. It is 120 cm tall and has 20 degrees of freedom to facilitate movements. The robot has the ability to interact with humans through its touch screen. Furthermore, it can communicate via voice in 15 languages and with gestures. To achieve that, it uses touch sensors, LEDs, and microphones [48].

The robot is used in pilot projects in various health-related scenarios, for example in guiding and informing patients [49], in treating dementia patients through learning their daily schedules, and reminding them [50], making doctor appointments, checking health measures and making contacts with families [51] or in exercising with older adults, to keep them mentally and physically healthy [52]. While the usability of Pepper for these use cases and the use of Pepper in real hospitals in general may be questionable in the long term, for this research project it offers opportunities to be used for the experimental task. Compared to other robots, it is relatively easy to program using the open platform ”Choregraphe”, for which there are also explanations on Github. Furthermore, it is designed to interact socially through different modalities and to be perceived as a social being. Thus, we decided to use Pepper as the research platform to answer the research questions based on the experimental task.

4.2.1 Programming the Robot

The desired task was programmed using the open-source platform Choregraphe. Choregraphe is a modeling tool that allows the design of robot behavior based on libraries. A separate behavior was programmed for each interaction, resulting in seven robot behaviors that differ in task instructions and behavior based on manipulation and applied strategies. The code mainly consists of functions for generating and recognizing speech and for moving the robot's head, arms, and body. Both animated and static speech modules were used for communication. Natural turn-taking was simulated through the speech recognition module. The threshold for correct recognition was set to zero to ensure similar interactions for all participants. As a result, every interaction was hard-coded. The robot was not programmed to react dynamically to the user's input. This deliberate decision aimed to ensure that interactions were as comparable as possible and to avoid participants becoming stuck in interactions with the robot or having more interactions than others. This would be useful in a real-life scenario, but it is not applicable in this case. Additionally, the robot's tablet was used to ensure that no one got stuck while interacting with the robot. The tablet was programmed to display parts of the task as visual support. A website was created to display the information on a full screen, using a module to display apps or web pages on the tablet in Choregraphe. This aimed to ensure that the participant can successfully complete the task.

Figure 4.3: Pick up table with items

4.3 Interaction Design

The goal of the task was to deliver a medical parcel to a specific doctor in time. This was achieved by filling an empty parcel with a particular amount of items and then delivering the parcel to a doctor's post box (see Figure 4.4). Four different item colors existed: blue, green, red, and yellow (see Figure 4.3). The robot initiated the interaction, informing the participant that it could help and asking who the addressee was. The participant then answered with the doctor's name. Pepper told the participant which items to take for this person and where to bring them (see interaction structure in Table 4.1). The participant had to take the same amount of items in every interaction. The number of items was also shown on the tablet screen to ensure the task could be executed fluently without the help of the researcher.

Figure 4.4: Post boxes

4.3.1 Pilot Test of the Experiment Task

We executed a pilot test for the experiment task. In the pilot, the task was tested with three potential participants. The goal was to determine whether the unpredictable action is perceived as unpredictable. The initial task idea differed from the final version in that it aimed to introduce unpredictability by upgrading the robot's capabilities to include real-time location tracking of addresses. To indicate that Pepper was processing information, the initial plan was to show idle movements for approximately ten seconds before making a statement about where to bring the parcel. When testing this first setup, it became clear that the line between an error and unpredictability was thin. The subject immediately asked if he had done something wrong or if Pepper was broken. As a result, the time for the second test setup was reduced from ten to five seconds. While the behavior was now less likely to be perceived as an error, another problem arose. It seemed difficult not to design mitigation strategies as

	Conversation	Robot behavior	Applied Approach
Part 1			Explanatory text
	Robot: ”Hi, I am Pepper. I can help you pack and deliver the parcel you need to deliver. Behind you, you find the resources table with the items. The post boxes or on this floor as well. To which person do you want to deliver the parcel?”	none	
	Participant: ”I have a parcel with important medicine for Doctor Wilson.”	none	
Part 2			Foreshadowing
	Robot: ”Okay, Doctor Wilson’s office is on this floor. You need to deliver one yellow, two red, two green and one blue item. You find the items on the pickup table behind you. After pickup, you need to bring the items to his postbox, right behind the black wall. If you are unsure about the task, please have a look at my tablet. I hope this was helpful, you can start now.”	moving and gesticulating	

Table 4.1: Interaction Design

a researcher automatically. The second participant said the behavior was strange and surprising, but it was still okay because she got an explanation immediately afterward. It seemed that in the second task design, a mitigation strategy was unintentionally applied, as it was explained immediately afterward why the robot acted as it did.

In the following process, the task was completely redesigned and tested in a third pilot trial, resulting in the above-described experiment task. The unpredictable action was made more obvious to elicit reactions better. The complexity of the setup was reduced to ensure comparability. The time per interaction was reduced to one minute because the walking distances in the experiment were shorter with the new setup. In addition, the speech rate was decreased to 90%, and the tablet view was added, as participants in the test reported that it was difficult to understand the robot and needing help from the researcher was not desired. The experimental protocol was also tested, resulting in minor changes and highlighting the most critical aspects. The pilot also revealed some strategies for dealing with frequent hardware problems. The connection to the robot was lost, or the robot or the software crashed. As a result, the robot and Choregraphe were rebooted between each participant, and ample time slots of 60 minutes per participant were scheduled to deal with potential problems.

Unpredictability was eventually designed to be a change in robot capabilities. When showing predictable behavior, the robot is only capable of speaking. When unpredictable, the robot can gesticulate, move around, and move its head. After asking the participant to which person they want to deliver, the robot acts in the task unpredictably by moving in the direction of the participant and lifting its head to make eye contact. Then, it describes the task using gestures, moving and looking toward the mailboxes.

4.3.2 Applied Strategies

Different approaches were considered to influence the impact of robot unpredictability. From the predictive coding perspective and when aiming for a potentially stressed-out user group, intervening with strategies before the unpredictable action happens was considered more valuable than afterward. The applied methods are explained below.

Getting informed via Text about Robot Changes in the Interaction Phase

The first approach, which was applied in the interaction phase before the unpredictable action, consisted of a text that informed the participant. In a real-life scenario, a text could be shown on a screen when entering a hospital or be sent beforehand via email. This approach was chosen and considered relevant based on anti-bias training [41] [42] [43] [44] [45]. In addition, current research suggests that setting expectations about what a robot can and cannot do can have a positive impact on how it is perceived [32] [46] [47]. As described in section 3.2.2, we expected that aiming for anti-bias goals might positively influence the perception of unpredictable behavior. To test this, a text was created based on similar goals. The goal of the text is to make the observer more aware and curious about unexpected behaviors to happen. Furthermore, the benefits of being open to changes and that it is normal to be confused by something new are outlined in the text. The exact new robot feature was not revealed in the text. The goal was to set the observer back into the initial phase, where they expect something completely new and unexpected. The wording was chosen based on Google research for synonyms and related words for the specific goals. To achieve this, ”www.thesaurus.com” and ”www.dictionary.cambridge.org” were used. As input, the words ”awareness”, ”curiosity”, ”benefits” and ”confusion” as well as their verbs and adjectives were used. The following text resulted and was displayed to the participants:

> Dear participant!
>
> Please be aware that the robot has some new features.
>
> It’s normal to be confused by new things, so don’t worry.
>
> You can be sure that each new feature will be of great benefit to you!

Foreshadowing the Ability to Move and Gesticulate

The second approach, applied in the action phase, uses foreshadowing to indicate that the robot is about to move. As described in section 3.2.1, foreshadowing can help increase the predictability of the robot. In contrast to other research like Takayama et al. [9] or Dragan et al. [13], in this project, specific, single movements are not to be foreshadowed. Instead, the holistic capability of the robot to move is foreshadowed. This was done through foreshadowing, moving, and gesticulating. After not being capable of any movements, when foreshadowing was applied, Pepper moved its arms carefully. Then, Pepper lifted the right hand in front of its face, stretched its fingers, and looked at them. Afterward, the unpredictable behavior proceeded. Pepper moved towards the person but stopped at a safe distance of 30 centimeters and used gestures and movements for the remaining interaction. The sensory inputs used in the strategy are the same as in the manipulation: voice and movements.

4.4 Research Design

We executed a quantitative 2x2 between-subjects study with a manipulation check. We expected it to be challenging to recruit a sufficient number of participants. Still, a between-subjects design was chosen because strong learning effects were expected between the conditions due to the task and unpredictable behavior being the same in all conditions. In addition, participants may miss applied strategies from other conditions, leading to a negative experience or perceiving the robot as more unpredictable as they experience more different conditions. We expected learning and transfer effects to be essential to avoid accidentally triggering predictive coding mechanisms. The conditions were randomly assigned to the participants to increase validity. Ten interactions with the robot were performed, divided into two sessions of five interactions each. Between each session, there was at least one night of sleep.

Foreshadowing and the explanatory text served as the independent variables in this study. There were four conditions in the experiment (see Table 4.2 and the videos of the conditions). The robot acted predictably in C-Predictable, so its behavior did not differ from the other interactions. In C-Unpredictable, the robot started gesticulating and moving toward the participant in the second part of the interaction. The robot behaved unpredictably in the other three conditions, but strategies to mitigate this were applied. In C-Text, the participant was informed beforehand about behavior changes. In C-Foreshadowing, the robot executed a foreshadowing movement. In C-Both, both strategies were applied.

		Independent Variable 2 (Foreshadowing)	
		Not applied	Applied
Independent Variable 1 (Explanatory text)	Not applied	C-Unpredictable	C-Foreshadowing
	Applied	C-Text	C-Both

Table 4.2: Experiment Conditions

4.5 Experiment Setup

The experiment was conducted entirely in a meeting room of the Interaction Lab at Citadel (University of Twente). The room was prepared to avoid distractions. During the experiment, the researcher sat at a table on the right side of the room. The participant sat on a chair facing the researcher to answer questions on the researcher's laptop. The participant interacted with the robot placed next to the researcher's table. The items were placed on a table behind the participant. The post boxes were placed behind two black partition walls, so they were not visible directly (see Figure 4.5).

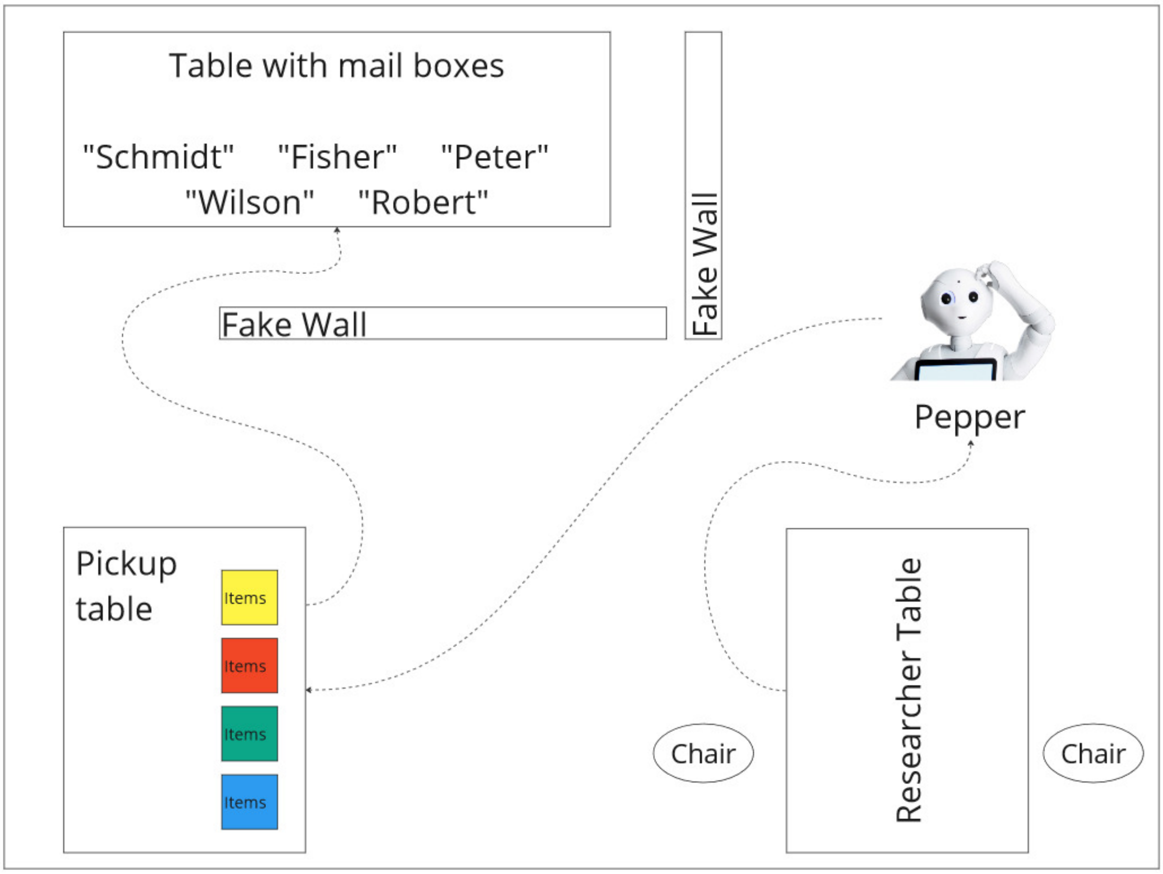

Figure 4.5: Experiment room setup

4.6 Procedure

The experiments took place from 06/02/2024 to 16/02/2024. Every session was a 1-on-1 session. On day 1, the participants entered the room and sat at the researcher's table. First, they were informed about the study, signed the consent form, and completed a questionnaire. Then, the participants were asked to imagine working in a hospital, delivering important packages to doctors. It was described that the package was important. The participant was also told they would be under time pressure, similar to working in a hospital. Initially, the participant was asked to stand in a specific place in the room. The robot was also placed in a particular place to ensure all participants experienced the same distance. The researcher told the participants the goal and handed them an empty parcel. The participant was told to fill the parcel and deliver it somewhere. It was said that the robot would be able to help with this. The participant was told that the robot would initiate the interaction and that if they wanted to respond to a question, they would have to wait for the robot's "beep" sound, which indicates that it is listening. A timer was then set to one minute to deliver the package. The participant took the package and stood in front of the robot. The robot then began to speak, asking the participant to whom the package should be delivered. The participant named the doctor. The robot then told the participant what to pick up and where to give it. The participant walked to the table, picked up the items, and placed them in the desired location. They experienced an entirely predictable robot that did not move or gesticulate. After five of these interactions, the first day of the experiment was done.

On the second day, the participants started with the same four similar interactions as on the first day, constantly experiencing the predictable condition but delivering different items. In the last interaction with the robot, different conditions were experienced. A questionnaire was filled out, and the participants were debriefed. The debriefing included being informed about the actual capabilities of the robot. In addition, the participant was asked not to discuss the aim of the study with anyone until the experiments had been completed to ensure that no one knew that the study was investigating predictability. Finally, it was noted if the participant wanted to read the study results afterward.

4.7 Measures

Moral trust, performance trust, and attributed predictability were measured as dependent variables. Negative attitudes towards robots, robot experience, and disposition to trust were taken as control measures. The measurements were taken in two questionnaires, a pre-experiment (see Appendix A.1) and a post-experiment questionnaire (see Appendix A.2). The pre-experiment questionnaire measured negative attitudes towards robots, robot experience, and disposition to trust. The post-experiment questionnaire measured moral trust, performance trust, attributed predictability, and demographics.

4.7.1 Trust

Moral and performance trust were expected to be influenced negatively by the unpredictable action. To measure these attributes, the "multidimensional measure of trust (MDMT)" by Malle & Ullmann [7] was used. The questionnaire offers two subscales for performance trust, the reliable subscale (items: "reliable", "predictable", "dependable", "consistent") and the competent subscale (items: "competent", "skilled", "capable", "meticulous"). For moral trust, three subscales exist: the ethical (items: "ethical," "principled," "moral," "has integrity"), the transparent (items: "transparent,"

”genuine,” ”sincere,” ”candid”) and the benevolent subscale (items: ”benevolent,” ”kind,” ”considerate,” ”has goodwill”). As suggested by the authors and since it was guaranteed to do the questionnaire on a desktop computer during the experiment [53], an 8-point Likert scale (0 = ”not at all” to 7 = ”Very”) with a ”Does not fit” option was used. Two question blocks were presented, each consisting of ten items. Since the participants were expected to be no English natives, ”I do not know what this word means” was offered to reduce the risk of participants falsifying the data by not knowing a word but still rating it.

	Subscale	Cronbach’s α	Excluded cases due to missing values
Performance trust	”reliable”	.68	0
	”competent”	.73	11
Moral trust	”ethical”	.62	13
	”transparent”	.57	15
	”benevolent”	.84	20

Table 4.3: Reliability analysis of trust measurement tool

The internal consistency was analyzed per scale using Cronbach’s α to ensure reliability. Offering ”Does not fit” and ”I don’t know what this word means” for performance trust and moral trust led to few answers for some scales (see Table 4.3). Thus, the reliability of the scales is difficult to interpret (see Appendix A.5 for in-depth information about missing values). For the reliable subscale, no answers were missing. Internal consistency was questionable (α = .68). For the competent subscale, internal consistency of the items was acceptable (α = .73), but 11 out of 25 cases were excluded from the test due to missing values. For the ethical subscale, the reliability test indicates questionable internal consistency (α = .62), and 13 cases were excluded. Due to missing values for all items, the ethical subscale has only 21 entries instead of 25. Here, a pattern is observable in the data: four participants did not rate any item, and two participants only rated one out of four items. The data show only a few typical similarities between these participants. Their robot experience is slightly lower (M = 3.67 in comparison to $M_{allParticipants}$ = 4.48), their dispositional trust is similar (M = 4.17 in comparison to $M_{allParticipants}$ = 3.92) and their negative attitude towards robots is similar (M = 2.78 in comparison to $M_{allParticipants}$ = 2.64). For the transparent subscale, the reliability test indicates poor internal consistency (α = .57), and 15 cases were excluded. While excluding the item ”transparent” from the scale would mathematically increase α by 0.2 to α = 0.77, this step is not done because ”transparent” is the only item with sufficient answers (24/25) and many cases of the scale were excluded. For the benevolent subscale, the reliability test indicates good internal consistency (α = .84), but 20 cases were excluded due to missing values.

4.7.2 Attributed Predictability

We measured attributed predictability to assess how the participant perceived the robot's predictability. The self-report questionnaire by Schadenberg et al. [11] was used. The scale measures attributed unpredictability through the six items "unpredictable," "irregular," "inconsistent," "random," "variable," and "erratic" based on a 7-point Likert scale (1 = "definitely not associated" to 7 = "definitely associated with" and "I do not know what this word means"), randomly presented. In contrast to Schadenberg et al. [11]'s suggestion, besides the sample were no English natives, no translations were used not to threaten scale validity.

Internal consistency was measured using Cronbach's α. In the attributed predictability questionnaire, answers are complete for the first four items. For attributed predictability, the reliability test indicates excellent internal consistency (α = .93).

4.7.3 Control Measures

As there were many different participants, individual differences may threaten validity. Being used to (robotic) technologies [54] and negative attitudes toward them [55] influence how interactions with them are perceived. To address the issue of confounding influences, we aimed to minimize individual differences by excluding them as control variables. As defined in the working definition in chapter 2.2, it can be expected that previous experience and the learning that results from it will influence how people perceive robots. From the predictive coding perspective, it can also be assumed that previous experiences or a negative attitude lead to (immature) mental models, which then (negatively) influence the interaction with Pepper.

To control whether the robot is perceived as unpredictable or whether the observer does not like robots in general, the observer's negative attitude towards robots negative attitude towards robots was measured using the "Negative Attitude Towards Robots Scale (NARS)" [55]. Robot experience was evaluated through a self-made question that asked about the participant's experience using an order from no experience to programming experience with physical robots. Finally, the disposition to trust in technologies was measured. A high disposition to trust was assumed to falsify the trust measures if people trust technology easily. While dispositional trust is a complex concept, the benefits of not overloading the participant with questions are considered high enough to justify using a short questionnaire for this control measurement. This is done through adapting three questions from

Nissen & Jahn [56].

Internal consistency of the control variables was tested using Cronbach's α. The data was complete for disposition to trust and negative attitude towards robots; no values were missing. The reliability analysis indicated acceptable internal consistency of disposition to trust (α = .74) and negative attitude towards robots (α = .69).

4.8 Data Analysis

As in the hypotheses defined, it was expected that the applied strategies C-Text, C-Foreshadowing, and the combination of both influence perceived trust, measured as performance trust and moral trust. Furthermore, this influence was expected to happen through positively influencing predictability, measured as attributed predictability. To test these relationships, we used IBM's SPSS. In this study, the samples were independent, and three dependent variables (attributed predictability, moral trust, and performance trust) existed.

First, the non-parametric Mann-Whitney U test was executed to compare for statistically significant differences in attributed predictability between the predictable and the unpredictable condition. Since the normality requirement for parametric approaches was not met, we chose a non-parametric test (see Appendix A.3).

Then, a MANOVA was executed. A MANOVA was more reasonable than an ANOVA to reduce the risk of type I errors due to multiple tests and eliminate the risk of overlooking existing effects since combinations of variables are not tested. Furthermore, covariates were measured and included in the analysis through a MANCOVA. All requirements for executing a MAN(C)OVA are given besides linearity (see Appendix A.4). This was ignored because the presence of non-linear relationships among variables does not necessarily invalidate the results of a MANOVA.

4.8.1 Data Preparations

Before analyzing the results, the data was prepared. In the test, the influence of the strategies on attributed predictability, performance trust, and moral trust compared to the unpredictable condition C-Unpredictable is analyzed. As a result, all data samples of condition C-Predictable were excluded from the analysis. For the dependent variables and dispositional trust, the items were averaged. "Does not fit" and "I don't know what this word means" were treated as missing values. For the negative attitude towards robots, the items 3 ("I would feel relaxed talking with robots."), 5 (" If robots had emotions I would be able to make friends with them.") and 6 ("I feel comforted being with robots that have emotions.") were inverted since they are positive statements, while the other statements are negative. Afterward, all 14 items were averaged. Since the robot experience was based on one item, this value represents the user's robot experience.

Chapter 5

Results

In this chapter, we report on the results. First, we describe the results of the manipulation check. Then, we describe the results of the hypothesis testing and observations during the experiments.

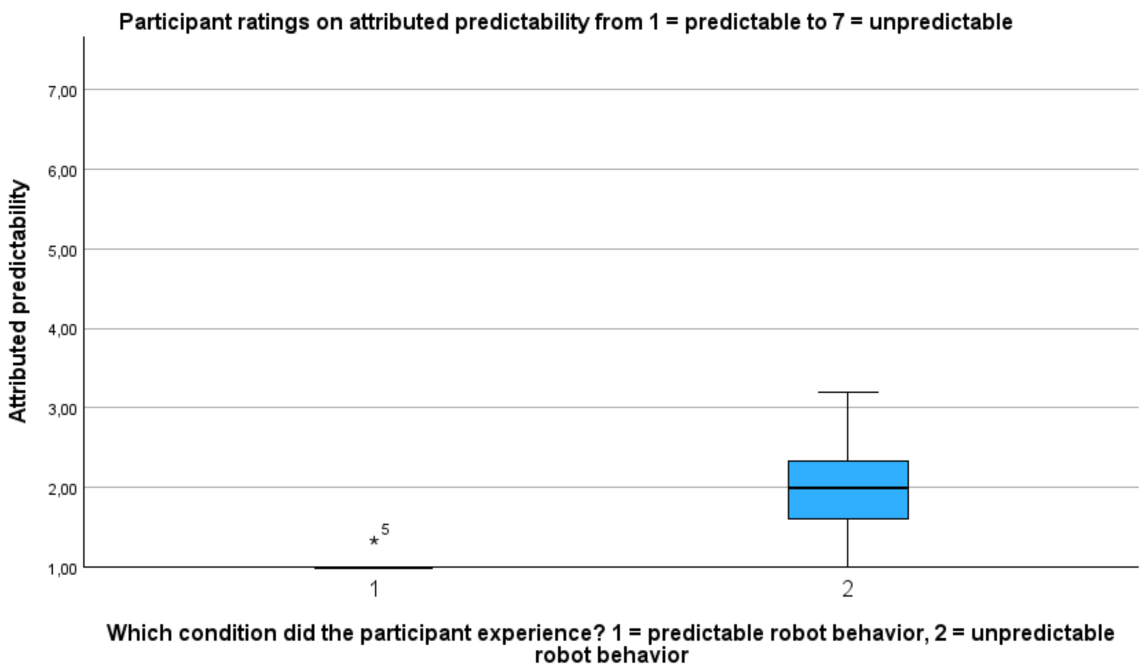

Figure 5.1: Participant ratings for attributed predictability

5.1 Manipulation check

For the manipulation check, we executed a Mann-Whitney U test. This test compares the mean rank of each group by converting scores into ranks. Looking at attributed predictability, lower ranks indicate that the robot was perceived as more predictable. While the mean rank of C-Predictable is 3.60 and the mean rank of C-Unpredictable is 7.40, the results of the Mann-Whitney U test indicate that there was no significant difference between the two conditions, $U = 3$, $z = -2.117$, $p = .056$. The difference was only approaching significance. Since the sample size for the manipulation check was small (n=10), the exact significance was interpreted instead of the asymptotic significance. Looking at the ratings, we see only slight difference between the conditions (see Figure 5.1).

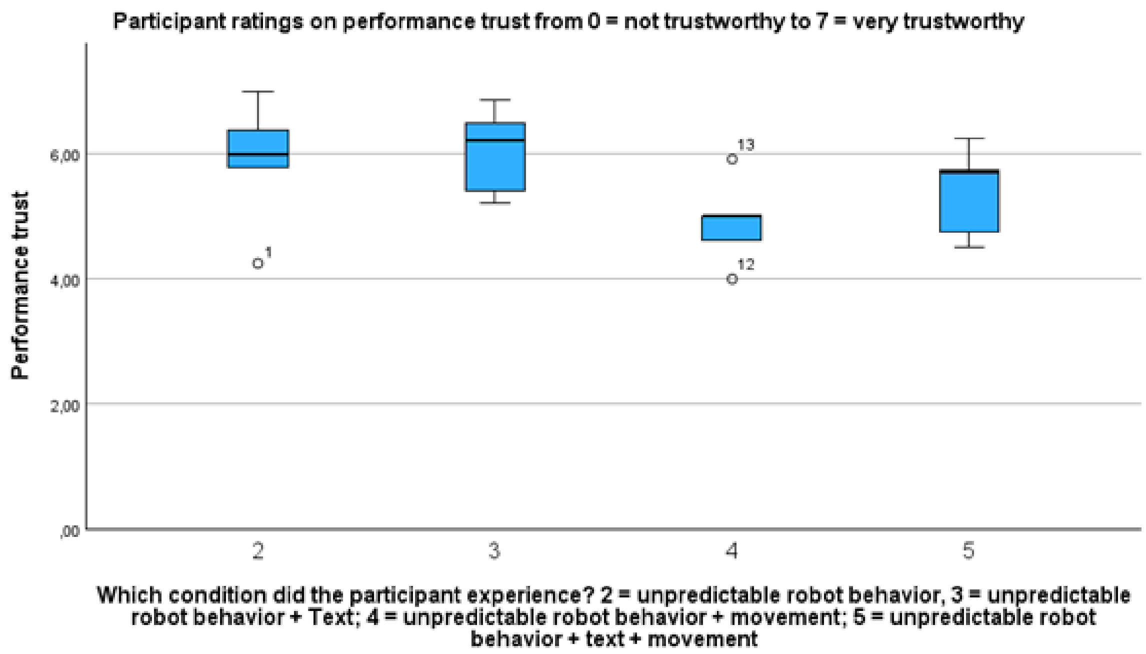

Figure 5.2: Participant ratings for performance trust

5.2 Testing the Hypotheses

The one-way MANOVA showed no statistically significant difference between the conditions on the combined dependent variables, $F(3, 34.223) = 0.908$, $p = .530$. This is reflected in the distributions of answers as well among the conditions (see Figure 5.2 for performance trust and Figure 5.3 for moral trust).

In the next step, we aimed to determine if the control variables' disposition to trust, robot experience, and negative attitude towards robots statistically significantly influenced the dependent variables using MANCOVA. There was no statistically significant difference between the conditions on the combined dependent variables after controlling for disposition to trust, $F(9, 31.789) = 0.789$, $p = .628$, robot experience, $F(9, 31.789) = 0.887$, $p = .547$, or negative attitude towards robots, F(9, 31.789) = .712, *p* = .694. As a result, it can be concluded that the statistical analysis indicates no significant differences between the conditions and no significant influence of the control variables. Excluding moral trust due to many missing values from the study did not result in significant differences based on MANOVA, *F*(6, 30) = 1.1, *p* = .385, or MANCOVA when accounting for the effect of robot experience, *F*(6, 28) = .929, *p* = .490, negative attitude towards robots, *F*(6, 28) = .766, *p* = .603, or disposition to trust, *F*(6, 28) = .897, *p* = .511.

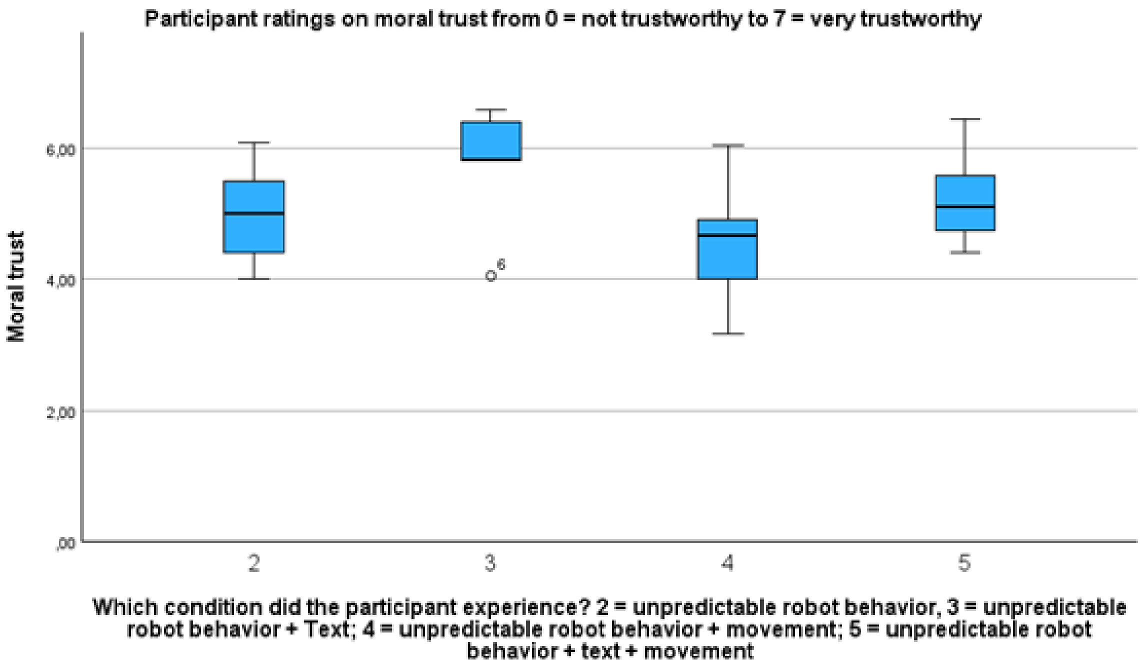

Figure 5.3: Participant ratings for moral trust

5.3 Observations during the Experiments

During the experiment, observations were taken to understand the results better. We have noticed that people use the tablet in very different ways. Some did not look at it, others checked it immediately after interacting with the robots, and others checked it twice (before and after picking up the items). When looking into how participants behaved, in C-Unpredictable, 4 out of 5 participants moved backward and created distance between them and the robot. In C-Predictable, 5 out of 5 participants stayed where they were. In C-Text, 1 out of 5 stayed where they were, and 4 out of 5 moved to make way for the robot. In C-Foreshadowing, 4 out of 5 stayed where they were, and 1 out of 5 moved to make space for the robot. In C-Both, 4 out of 5 participants stayed where they were, and 1 out of 5 stepped aside to make way.

All in all, the task seemed to be appropriate for the desired goal. The participants learned quickly (interaction 3, day 1) what they had to do and remembered it on the second day. At the beginning of each day (interactions 1 and 6), they checked by looking at the pickup table where the items were. No difficulties with the task were observed; it seemed to be clear. 24 out of 25 participants immediately understood where to bring the items. One participant needed a hint but explained beforehand that he was not feeling well.

The following aspects appeared to be not optimal in the experiments: Participants were asked to wait until the interaction with the robot was over before starting the task, but they did not always follow this. It seemed that once they knew what to do, they found it difficult to wait. This meant that sometimes, they were not close enough to read the tablet and had to go back. Also, in at least one session of 21 out of 25 participants[1], the items were not long enough visible on the screen when they wanted to recheck the amount. As a result, participants could not walk back because the tablet no longer showed the items. For one participant, this resulted in a problem since she did not know the amount of items anymore and asked the researcher for help. Others relied on their memory. The visibility time was increased after session 32, but it may still cause confusion or affect trust in the robot and its desire to help. A timer was also set to increase the stress of delivering on time. Except for one person, this did not seem to have any effect, as the participants walked relaxed and took their time picking up the items. It may be that the participants found it counter-intuitive to wait until the interaction was over and to be in a hurry. Furthermore, seven participants[2] experienced the robot crashing. In these cases, the participant was asked to leave the experiment room, and a complete robot reboot was performed. The participant was asked to wait outside because the robot would automatically move during the restart. Finally, although participants were asked to wait until the robot listened and spoke loudly and clearly, some spoke too early or quietly. It was, therefore, necessary to give them a sign during the interaction to repeat what they had said. This happened with eight participants[3]. They needed a hint from the researcher to repeat what they said.

[1] 10 participants experienced one session with this setup, 11 participants both sessions

[2] Condition C-UP: 2, C-PR: 1, C-Text: 2, C-Text: 1, C-Both: 2

[3] Condition C-UP: 1, C-PR: 1, C-Text: 2, B: 3, C-Both: 1; 18 out of 250 interactions with the robot

Chapter 6

Discussion

In this study, we investigated influencing a robot's moral and performance trust by mitigating the effects of its unpredictable behavior. We hypothesized that foreshadowing an unpredictable action can reduce a robot's unpredictability and thus increase trust in its performance and morals. We expected similar effects for informing the participant before the interaction about new robot features and applying both approaches together. We found no statistical evidence for these hypotheses. We looked into the influence of confounding variables as well by measuring the participants' attitude towards robots, previous robot experience, and disposition toward trust. There is no statistical indication that one of the measured confounding variables had impacted performance, trust, or attributed predictability.

We performed a manipulation check to confirm that the intended unpredictable behavior negatively impacts predictability. While there is no statistically significant evidence, the results are approaching significance. We expect this insignificance is due to the limited sample size, which reduces the likelihood of statistically significant effects. Furthermore, we need to consider that, due to a lack of normality in the predictable condition, we executed a non-parametric, less powerful test. Based on the numbers, all participants from the unpredictable condition rated the robot as highly predictable, while for the unpredictable sample, the ratings were more mixed. Looking further into the participants' behavior indicates a difference in how predictable the robot was perceived. In the predictable sample, participants reported feeling at ease with the robot. They stated that they had a positive experience with Pepper, as they knew what to expect throughout the experiment. Furthermore, they stayed the whole interaction at the desired spot without moving away. In contrast, participants from the unpredictable sample group appeared confused. Four out of five participants attempted to distance themselves from the robot when it started to move. Three participants frequently looked toward the researcher, seeking clarification. Two participants reported that the robot had attacked them, while others expressed fear that it might run them over. Comparing these observations indicates that the robot was perceived more unpredictable in the unpredictable condition. But, the statistical results provide no evidence. Thus, we conclude that we are uncertain whether the manipulation worked.

6.1 Scientific Research is More than P-values

Finding no significant effect for the applied strategies does not necessarily mean that there is no effect, but that this experiment could not show effects given the research approach. A quantitative approach was considered reasonable in the planning phase for this research project. This enabled us to objectively and reliably test the complex connections between the approaches, predictability, and trust, as aimed for in this study. A quantitative approach was appropriate because we have a theoretical foundation for the relationship between unpredictability, trust, bias, and foreshadowing, and our goal was to make general statements about how to influence this relationship.

To test the relationship reliably, we lacked sufficient participants in this study. Underpowered tests are a general problem, not only in HRI but in other sciences. Nevertheless, many studies have low power and find effects [57]. To analyze the power of our statistical test, we conducted a post-hoc power analysis using G*Power. The results find that if we had found a large effect in this study, given the sample size of 20 and a significance criterion of $\alpha = .05$, the power of this test would have only been .40. Our goal was to achieve a higher number of participants to increase the statistical power of the study. However, the recruitment process proved to be challenging despite utilizing several channels. The experiments were conducted in the Interaction Lab on campus, and although people near the lab were also approached, only a few participants were found. The reason for this may be, on the one hand, that participants were required to schedule a second session with the researcher. Participants cited time constraints as the reason for not being able to attend both sessions on separate days. On the other hand, the smaller participant sample may be attributed to the timing of the experiments, which took place after the week on campus following exams. This may have decreased individuals' motivation to participate in university-related activities. Thus, many participants came from personal contacts and did not reflect an average person but were tech-savvy and robot-experienced.

But, research is more than p-values. Interpreting a single index [58] should not replace scientific reasoning. While non-significant studies are common but often not published [57], we can still learn from them for future research. In the following sections, we discuss what we learned regarding the approaches, task design, methodology adjustments, and learnings for recruitment and measurements.

6.2 Recommendations for Further Investigation of the Independent Variables

We applied foreshadowing and explanatory text based on anti-bias strategies. The following discusses whether we recommend looking into these approaches again and why.

6.2.1 Applying Foreshadowing

For foreshadowing, we applied the approach of foreshadowing small movements from other research [9] [13] [12] [35] to a broader scope, as the robot foreshadowed moving and gesticulating as a whole. When we reflect on this, we recommend following this approach further. Comparing the participants' behavior in the different conditions indicates that foreshadowing influenced their behavior. In the predictable setting, they did not move and seemed relaxed. In the unpredictable setting, they act confused, creating distance between them and the robot. When using foreshadowing, they reacted similarly to the predictable setting and were not surprised by the robot's behavior. This indicates that foreshadowing did influence how they perceived the robot acting unpredictably afterward. We must be careful when interpreting this since no structured observations were taken. But, besides the statistical results, we do not find any indication of why foreshadowing was unsuccessful. Instead, we expect that problems with the task (see section 6.3) had an impact. Thus, we conclude that these findings are a sufficient indication for further looking into this approach.

6.2.2 Applying Anti-Bias Strategies

Leichtmann et al. [57] state that many research projects in HRI lack a theoretical basis. While this might be the case in other HRI studies, we do not find a lack of theory on the relationship of internal models and bias the problem. Instead, we expect that the anti-bias strategies had an influence, but their implementation was not strong enough. Observing the participants indicates that the approach had some impact. They mostly tried to make way for the robot when using the explanatory text. We carefully interpret that they might have predicted what the robot was planning to do and tried to react since they expected something new to happen based on the text. But reading the explanatory text took only a few seconds. It might be possible that achieving anti-bias goals requires more extended periods and repetition since anti-bias training often takes whole days. Another thought is that some biases cannot be reduced by training [37]. We do not yet know which biases this applies to or does not. Thus, we recommend diving deeper into the question of when anti-bias strategies are successful and when not, and which requirements must be fulfilled. This could be achieved through a literature review of existing anti-bias approaches, resulting in recommendations for successfully designing anti-bias training. We expect a literature review to be reasonable because it is difficult to evaluate the effects of anti-bias training since they mostly rely on self-reports, and it is difficult to measure actual long-term behavior change [42]. Thus, we recommend further investigating the extent to which it is possible to reduce the effects of robot unpredictability on trust using anti-bias training, with the advice to dive deeper into the design of anti-bias training beforehand.

6.3 Recommendations for Task Design

The following discusses what we learned from designing the task for this study and the choice of the robot.

6.3.1 The Task

HRI is highly context-sensitive. Effects might be visible only under certain conditions [57]. We asked participants to imagine working at a hospital while being in a university project room, delivering a parcel to an imaginary doctor. We expect that this influenced the task. Due to the project's scope, it was impossible to create a real situation in a hospital. Thus, the experiment was conducted in an artificial laboratory setting. A laboratory experiment had advantages. It reduced the influence of confounding factors. Moreover, it contributed through the controlled setting to standardization and replicability. These were important for the reliability of our study since we were aiming to make generalized statements through hypothesis testing.

But, participants stated that they struggled to imagine the situation. This was also visible because only one participant hurried, as asked, to finish the task. Observing the participants further indicated that having the same interaction in all trials led to paying less attention to the situation. After executing the third session on day one and the second session on day two, many participants had problems waiting until the robot was finished talking since they already knew what it would say. Some participants turned their backs on the robot early since they already knew what they had to do. From a predictive coding perspective, this appears to be positive. The participants seem to have successfully built a model of the situation since they were able to predict what the robot would say and what they had to do. We wanted to achieve this by showing structural regularities. But, we expect that achieving this would have been possible without creating this unrealistic situation. Thus, when redoing the project, we recommend creating a less artificial task requiring less imagination. We recommend stepping back from the hospital use case and fitting the task into the experiment environment. A similar task could be executed in a university setting. A participant could bring important documents to professors in the building. Designing a more realistic task would reduce repetition as well. Furthermore, we advise that the researcher remain outside of the experiment room to avoid any potential involvement with the participants during the experiment.

Another interesting aspect we would like to mention is that since the robot could still help with the task, it might not be perceived as less trustworthy. In the questionnaires, trust in the robot's performance and the robot's morals was measured. The robot did not lie or fail to help. It might be possible that unpredictable body behavior does not influence trust if it is still capable of helping to achieve the desired goal since the ability to depend on the robot is an aspect of robot trust as well [6]. Overall, we expect it to be reasonable to design a realistic task in which no goal has to be reached. This could be achieved by creating an interaction mainly based on communicating about a topic instead of fulfilling a task.

6.3.2 The Robot

Additionally, we recommend considering another robot. Pepper was chosen because it was easily programmable and available at the university, but it comes with its own (in-) capabilities, which must be taken into account. First, Pepper is a robot that is known by the (interested) public through the news. Especially in the study of interaction technology, Pepper is known, as it is presented by professors at the beginning of studying and used for projects. Many participants came from this study so they may have known Pepper already. Second, choosing another robot would be reasonable due to the unstable hardware- and software, which, for seven participants, led to problems.

From the predictive coding perspective, we expect that the choice of the robot influenced the experiment. Human-like robots can trigger with their appearance expectations of human-like behavior [56] [59] [60]. In this experiment, the unpredictable behavior (i.e., moving and gesticulating) was executed through visible aspects of the robot (i.e., wheels at the bottom of the robot, arms). One could question why a robot has wheels and arms if it cannot move or gesticulate. Thus, only a robot that does not provide any information by visual inspection may not activate related models regarding its capabilities. If the participant had already expected movements and gestures, there would have been nothing to question about the model. It might be possible that we tried to lower precision in a model that is not questioned because there is no prediction error. Thus, we recommend having no visible cues to indicate the unpredictable behavior beforehand for the experiment.

6.4 Recommendations for the Methodology

Furthermore, we recommended doing a complete pilot which also tests the measurement. In this study, only the task itself was tested since we expected most problems with designing the unpredictable action. Designing unpredictability was expected to be difficult because the robot should have only been perceived as unpredictable, not erroneous, to follow our theoretical framework. We anticipated that erroneous behavior would have other effects and thus confound the results. The experiment was tested beforehand with three participants, asking open-ended questions afterward to assess the experiment's quality. In the third pre-test, the participant had few questions, showed no uncertainty about the final task, and did not have questions during it. In the complete experiment then, participants often looked to the researcher for help, or questions were asked. Thus, the pre-testing results did not reflect how the experiment was perceived afterward. The reason for this might be that in the pilot, the participants only experienced five sessions without a night of sleep in between, so it did not represent the actual study. Furthermore, in the pilot tests, no participant experienced a software crash or problems with communicating with Pepper. There were problems, but participants were unaware since we asked them to start later. Lastly, the measurements were not tested in the pilot. In a pilot study, we recommend testing if the participants understand all items (and their translations) well since this was a problem in this study.

6.5 Recommendations for Recruiting

As one of the last aspects, we recommend increasing the power of the study. We conducted an a-priori power analysis using G*Power. The analysis indicates that for this study, 11 participants for each of the four conditions would be necessary to find a large effect considering Cohen [61]'s criteria (small effect: .01, medium effect: .0625, large effect: .14) with a significance criterion of α = .05 and power = .80. Thus we advise to recruit 44 participants. The participants should reflect the general public because factors like employment, income, and educational background can influence the adoption of new technologies [62] and trust in others [63]. Although the data does not demonstrate this influence, we expect that it is probable that the participants' prior experience with robots had an impact. From the perspective of predictive coding, knowing robots already results in predictions on how they will behave. Thus, more requirements for the recruitment process are necessary. First, participants should have no experience interacting with, programming, or developing physical robots to reduce the risk of triggering related models. Second, participants should come from all educational levels. Third, the goal should be to find participants with no Pepper-related internal models, such as participants who have never seen the robot in person.

6.6 Recommendations for Measurements

Lastly, reflecting on the measurement results in the following learnings. The MDMT questionnaire was chosen for trust, which measures performance trust and moral trust but is still in development. Choosing this questionnaire was reasonable since trust is a multi-faceted concept [7]. In the final participant sample, many participants were not English natives and did not know many words. We did not exclude these items from the scale to avoid risking the scale's validity. While we still believe it was the right choice to include these options to avoid a type II error, this influenced the results. Due to a lack of data, the moral trust scale cannot be reliably interpreted.

For future research, we recommend waiting for another publication cycle or choosing another questionnaire. This is because while Malle et al. [7] indicate high internal consistency for the questionnaire, we could not show this. The reliability analysis suggests that the questionnaire items for each scale do not seem to address the same issue well, but many cases were excluded. Thus, we are uncertain about the questionnaire's reliability. When using a revised version of the questionnaire, we still recommend offering "Does not fit" and "I don't know what that means" as answer options to ensure the validity of the results. Depending on the expected participant sample, we recommend translating the scale. Here, the cost-benefit ratio of investing in the translation and validation should be considered.

All in all, depending on the environment in which the study is executed, we recommend either waiting for the next development stage of the questionnaire, considering translations, and validating these translations, or choosing another questionnaire that addresses the multi-facet characteristics of trust.

Chapter 7

Conclusion

In conclusion, this master's thesis explores mitigating the impact of unpredictable robot behavior on trust using the perspective of predictive coding. Recognizing the pivotal role of trust in human-robot interactions, this study acknowledges the challenges posed by unpredictable robot behavior, which can undermine trust formation [16] [6] [17].

The strategies developed in this study were built based on insights from predictive coding, existing HRI, and anti-bias research. These strategies, consisting of foreshadowing movements and information provision, were subjected to empirical testing in a quantitative study. However, the statistical analysis did not yield evidence of their effectiveness in the context of the experimental design employed.

Although the results may indicate that the applied strategies did not achieve the desired effects, it is essential to consider them valuable contributions to the ongoing discourse surrounding human-robot interaction because they provide important information about future research design and how to develop and test for unpredictable behavior. Instead of dismissing the strategies outright, we highlight the necessity for further refinement and reconsideration of experimental methodologies in future research efforts. We recommend looking further into the research questions by executing a quantitative study with a different setup. Participant amounts and requirements during recruiting should be increased. We recommend redesigning the task to be more realistic and aiming to have fewer trigger-related models. Furthermore, we advise reconsidering the measurement tool. We recommend investigating the requirements for successful anti-bias training more profoundly, including the anti-bias approach.

This research contributes through implications for developing experiments and the challenges of designing for and mitigating unpredictable robot behavior. In the future, it is crucial to investigate further and create new approaches in this field to build stronger and more reliable relationships between humans and robots. This will enable us to fully incorporate robots into our everyday activities.

Bibliography

[1] European Commission, "Ageing Europe - statistics on population developments," 2020. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_population_developments

[2] M.-K. Looi, "The European healthcare workforce crisis: how bad is it?" *BMJ*, vol. 384, 2024.

[3] A. A. Scoglio, E. D. Reilly, J. A. Gorman, and C. E. Drebing, "Use of social robots in mental health and well-being research: Systematic review," *Journal of Medical Internet Research*, vol. 21, no. 7, pp. 1–14, 2019.

[4] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, "A Survey of Robots in Healthcare," *Technologies*, vol. 9, no. 8, 2021.

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[7] B. F. Malle and D. Ullman, "A multi-dimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction: Research and applications*. San Diego: Elsevier, 2021, pp. 3–24.

[8] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships." *Journal of Personality and Social Psychology*, vol. 49, no. 1, pp. 95–112, 1985. [Online]. Available: https://psycnet.apa.org/doiLanding?doi=10.1037%2F0022-3514.49.1.95

[9] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: Improving robot readability with animation principles," *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 69–76, 2011.

[10] C. D. Frith, "The social brain?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 671–678, 2007.

[11] B. R. Schadenberg, D. Reidsma, D. K. Heylen, and V. Evers, "I See What You Did There": Understanding People's Social Perception of a Robot and Its Predictability," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 3, 2021.

[12] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 51–58, 2015.

[13] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 301–308, 2013.

[14] M. Koppenborg, P. Nickel, B. Naber, A. Lungfiel, and M. Huelke, "Effects of movement speed and predictability in human–robot collaboration," *Human Factors and Ergonomics In Manufacturing*, vol. 27, no. 4, pp. 197–209, 2017.

[15] C. Lichtenthäler and A. Kirsch, "Goal-predictability vs. trajectory-predictability - Which legibility factor counts," *ACM/IEEE International Conference on Human-Robot Interaction*, no. January, pp. 228–229, 2014.

[16] M. Lewis, K. Sycara, and P. Walker, "The Role of Trust in Human-Robot Interaction," in *Foundations of Trusted Autonomy*, 1st ed., H. A. Abbass, J. Scholz, and D. J. Reid, Eds. Springer Cham, 2018, ch. 8, pp. 135–159. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-64816-3

[17] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 141–148, 2015.

[18] T. Ueno, Y. Sawa, Y. Kim, J. Urakami, H. Oura, and K. Seaborn, "Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods," *Conference on Human Factors in Computing Systems - Proceedings*, 2022. [Online]. Available: https://doi.org/10.1145/3491101.3519772

[19] T. Law and M. Scheutz, "Trust: Recent concepts and evaluations in human-robot interaction," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds. Elsevier Inc, 2020, pp. 27–57.

[20] M. Chita-Tegmark, T. Law, N. Rabb, and M. Scheutz, "Can you trust your trust measure?" *ACM/IEEE International Conference on*

Human-Robot Interaction, pp. 92–100, 2021. [Online]. Available: https://doi.org/10.1145/3434073.3444677

[21] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-HRI"," *Robust Intelligence and Trust in Autonomous Systems*, pp. 191–218, 2016.

[22] C. Lichtenthäler, T. Lorenz, M. Karg, and A. Kirsch, "Increasing perceived value between human and robots - Measuring legibility in human aware navigation," *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO*, pp. 89–94, 2012.

[23] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[24] K. Friston, "Does predictive coding have a future?" *Nature Neuroscience*, vol. 21, no. 8, pp. 1019–1021, 2018.

[25] J. Hohwy, *The predictive mind.* OUP Oxford, 2013.

[26] J. M. Kilner, K. J. Friston, and C. D. Frith, "Predictive coding: An account of the mirror neuron system," *Cognitive Processing*, vol. 8, no. 3, pp. 159–166, 2007.

[27] M. Otten, A. K. Seth, and Y. Pinto, "A social Bayesian brain: How social knowledge can shape visual perception," *Brain and Cognition*, vol. 112, pp. 69–77, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.bandc.2016.05.002

[28] J. Kwisthout, H. Bekkering, and I. van Rooij, "To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions," *Brain and Cognition*, vol. 112, pp. 84–91, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.bandc.2016.02.008

[29] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009.

[30] C. Esterwood and L. P. Robert, "A Literature Review of Trust Repair in HRI," *RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots*, no. July, pp. 1641–1646, 2022.

[31] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing Robot Incapability," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 87–95, 2018. [Online]. Available: https://doi.org/10.1145/3171221.3171276

[32] M. K. Lee, S. Kielser, J. Forlizzi, S. Srinivasa, and P. Rybski, “Gracefully mitigating breakdowns in robotic services,” *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010*, pp. 203–210, 2010.

[33] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, “Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction,” *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 493–500, 2016.

[34] C. Sandi and J. Haller, “Stress and the social brain: Behavioural effects and neurobiological mechanisms,” *Nature Reviews Neuroscience*, vol. 16, no. 5, pp. 290–304, 2015.

[35] B. Camblor, N. Benhabib, D. Daney, V. Padois, and J. M. Salotti, “Task-Consistent Signaling Motions for Improved Understanding in Human-Robot Interaction and Workspace Sharing,” *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2022-March, pp. 275–283, 2022.

[36] B. R. Schadenberg, D. Reidsma, V. Evers, D. P. Davison, J. J. Li, D. K. Heylen, C. Neves, P. Alvito, J. Shen, M. Pantić, B. W. Schuller, N. Cummins, V. Olaru, C. Sminchisescu, S. B. Dimitrijević, S. Petrović, A. Baranger, A. Williams, A. M. Alcorn, and E. Pellicano, “Predictable Robots for Autistic Children-Variance in Robot Behaviour, Idiosyncrasies in Autistic Children’s Characteristics, and Child Robot Engagement,” *ACM Transactions on Computer-Human Interaction*, vol. 28, no. 5, 2021.

[37] T. W. Hung, “Why Human Prejudice is so Persistent: A Predictive Coding Analysis,” *Social Epistemology*, vol. 00, no. 00, pp. 1–19, 2023. [Online]. Available: https://doi.org/10.1080/02691728.2023.2237942

[38] J. Fridman, L. F. Barrett, J. B. Wormwood, and K. S. Quigley, “Applying the Theory of Constructed Emotion to Police Decision Making,” *Frontiers in Psychology*, vol. 10, no. 1957, 2019. [Online]. Available: https://doi.org/10.3389/fpsyg.2019.01946

[39] Z. A. Neemeh, “Bootstrap Hell: Perceptual Racial Biases in a Predictive Processing Framework,” *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*, no. January 2020, pp. 815–821, 2020.

[40] R. Kirkland and I. Bohnet, “Focusing on what works for workplace diversity.” 2017. [Online]. Available: https://www.mckinsey.com/featured-insights/gender-equality/focusing-on-what-works-for-workplace-diversity

[41] S. Barsade, “How Companies Can Develop Anti-bias Strategies that Work,” 2018. [Online]. Available: https://knowledge.wharton.upenn.edu/article/how-companies-can-develop-anti-bias-strategies-that-work/

[42] D. Atewologuh, T. Cornish, and F. Tresh, *Unconscious bias training : An assessment of the evidence for effectiveness*, 2018.

[43] E. R. Carter, I. N. Onyeador, and N. A. Lewis, “Developing & delivering effective anti-bias training: Challenges & recommendations,” *Behavioral Science and Policy*, vol. 6, no. 1, pp. 57–70, 2020. [Online]. Available: https://doi.org/10.1177/237946152000600106

[44] J. Nordell, *The End of Bias: A Beginning: The Science and Practice of Overcoming Unconscious Bias.* Henry Holt and Company, 2021.

[45] F. Gino and K. Coffman, “Unconscious Bias Training That Works,” 2021. [Online]. Available: https://hbr.org/2021/09/unconscious-bias-training-that-works

[46] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, “Effects of changing reliability on trust of robot systems,” *HRI’12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 73–80, 2012.

[47] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, “Robot confidence and trust alignment,” *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 1, pp. 155–156, 2013.

[48] Aldebaran United Robotics Group, “Pepper,” 2023. [Online]. Available: https://www.aldebaran.com/en/pepper

[49] Aldebaran (part of United Robotics Group), “Robots for Healthcare,” 2021. [Online]. Available: https://www.youtube.com/watch?v=2TcmjL6ZQzc&ab_channel=Aldebaran%2CpartofUnitedRoboticsGroup

[50] RobotLAB Inc., “Pepper Robot for Healthcare,” 2021. [Online]. Available: https://www.youtube.com/watch?v=ZFrwk5auOvU&ab_channel=RobotLABInc.

[51] Aldebaran (part of United Robotics Group), “Pepper and NAO robots solutions for healthcare,” 2020. [Online]. Available: https://www.youtube.com/watch?v=PIVR10xLI6c&ab_channel=Aldebaran%2CpartofUnitedRoboticsGroup

[52] FOX 9 Minneapolis-St. Paul, "Robots providing helping hand at nursing home — FOX 9 KMSP," 2022. [Online]. Available: https://www.youtube.com/watch?v=BXpgiT51CoQ&ab_channel=FOX9Minneapolis-St.Paul

[53] B. F. Malle and D. Ullman, "Presentation Format for MDMT questionnaire," 2021. [Online]. Available: https://research.clps.brown.edu/SocCogSci/Publications/Pubs/Malle_&_Ullman_2021_Trust_chap.pdf

[54] M. M. De Graaf and S. Ben Allouch, "Exploring influencing variables for the acceptance of social robots," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1476–1486, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2013.07.007

[55] T. Nomura, S. Tomohiro, K. Takayuki, and K. Kensuke, "Measurement of negative attitudes toward robots," *Interaction Studies*, vol. 7, no. 2, p. 437–454, 2006.

[56] A. Nissen and K. Jahn, "Between Anthropomorphism, Trust, and the Uncanny Valley: A Dual-Processing Perspective on Perceived Trustworthiness and Its Mediating Effects on Use Intentions of Social Robots," in *Proceedings of the 54th Hawaii International Conference on System Sciences*. University of Hawai'i at Manoa, 2021, pp. 360–369.

[57] B. Leichtmann, V. Nitsch, and M. Mara, "Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement," *Frontiers in Robotics and AI*, vol. 9, no. March, pp. 1–15, 2022.

[58] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016. [Online]. Available: http://dx.doi.org/10.1080/00031305.2016.1154108

[59] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, "Brave new world: service robots in the frontline," *Journal of Service Management*, vol. 29, no. 5, pp. 907–931, 2018.

[60] T. Männistö-Funk and T. Sihvonen, "Voices from the Uncanny Valley: How Robots and Artificial Intelligences Talk Back to Us," *Digital Culture & Society*, vol. 4, no. 1, pp. 45–64, 2018.

[61] J. Cohen, *Statistical power analysis for the behavioral sciences.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

[62] A. J. van Deursen and J. A. van Dijk, "The digital divide shifts to differences in usage," *New Media and Society*, vol. 16, no. 3, pp. 507–526, 2014.

[63] J. Delhey and K. Newton, “Who trusts?: The origins of social trust in seven societies,” *European Societies*, vol. 5, no. 2, pp. 93–137, 2003.

Appendix A

Appendix

A.1 Pre-Experiment Questionnaires

Please share your overall thoughts on the robot. There is no right or wrong, it is your subjective perspective that matters. To achieve that, please rate all items on the given scales.

1. Question

   Please rate the following statements based on your personal thoughts on a scale from 1 (I do not agree at all) to 5 (I agree completely). There is no right or wrong.

   - I generally trust technologies.
   - I generally have faith in technologies.
   - I generally trust technologies unless they give me reason not to.

2. Question

   Please rate the following statements based on your personal thoughts on a scale from 1 (I do not agree at all) to 5 (I agree completely) . There is no right or wrong.

   - I would feel uneasy if robots really had emotions.
   - Something bad might happen if robots developed into living beings.
   - I would feel relaxed talking with robots.
   - I would feel uneasy if I was given a job where I had to use robots.
   - If robots had emotions I would be able to make friends with them.
   - I feel comforted being with robots that have emotions.
   - The word “robot” means nothing to me.

- I would feel nervous operating a robot in front of other people.
- I would hate the idea that robots or artificial intelligences were making judgements about things.
- I would feel very nervous just standing in front of a robot.
- I feel that if I depend on robots too much, something bad might happen.
- I would feel paranoid talking with a robot.
- I am concerned that robots would be a bad influence on children.
- I feel that in the future society will be dominated by robots.

3. Question

   This question addresses your previous experience with robots. Please choose the one which describes your experience with robots before today best. Please view the options as an order, starting with ”0 - I have never heard of robots.” as no experience and ”6 - I have previously participated in the development (building, programming, testing) of a physical social robot like Nao, Pepper, or Sophia myself.” as the most experience.

   - 0 - I have never heard of robots.
   - 1 - I know robots only from fictional stories (movies, books, TV-shows).
   - 2 - I know robots only from fictional stories (movies, books, TV-shows) and from real-life formats (TV news, newspaper)
   - 3 - I have already interacted with robot vacuums.
   - 4 - I have already interacted with robot toys such as furby or with animatronics from theme parks.
   - 5 - I have interacted with physical social robots like Nao, Pepper, or Sophia.
   - 6 - I have previously participated in the development (building, programming, testing) of a physical social robot like Nao, Pepper, or Sophia myself.

A.2 Post-Experiment Questionnaires

Please share your overall thoughts on the robot. There is no right or wrong, it is your subjective perspective that matters. To achieve that, please rate all items on the given scales. If you don’t know the word, choose ”I don’t know what this word means”.

1. Question

   Please rate the fit of the robot you just got to know and the following statements on a scale from 0 (Not at all) to 7 (Very). There is no right or wrong:

   The robot I just experienced...

   - is reliable.
   - is competent.
   - is ethical.
   - is sincere.
   - is benevolent.
   - is predictable.
   - is skilled.
   - is principled.
   - is candid.
   - is kind.

2. Question

   Please rate the fit of the robot you just got to know and the following statements on a scale from 0 (Not at all) to 7 (Very). There is no right or wrong:

   The robot I just experienced...

   - is dependable.
   - is capable.
   - is moral.
   - is transparent.
   - is considerate.
   - is consistent.
   - is meticulous.
   - has integrity.
   - is genuine.
   - has goodwill.

3. Question

   Regarding these questions, please think about the last session with the robot.

Please rate the fit of the robot you just got to know and the following statements on a scale from 1 (definitely not associated) to 7 (definitely associated). There is no right or wrong:

The robot I just experienced is...

- unpredictable.
- irregular.
- inconsistent.
- random.
- variable.
- erratic.

4. Demographics

   My gender is...

   - female.
   - male.
   - diverse.

   My age group is...

   - younger than 18.
   - 18 to 24.
   - 25 to 34.
   - 35 to 44.
   - 45 to 54.
   - 55 to 64.
   - 65 or older.

   What is the highest degree or level of education you have completed?

   - No schooling completed.
   - High School.
   - Bachelor's Degree.
   - Master's Degree.
   - Ph.D. or higher.

- Other ... .

I am currently...?

- Employed for wages.
- Out of work and looking for work.
- Out of work but not currently looking for work.
- A homemaker.
- A student.
- In Military.
- Retired.
- Unable to work.
- Other ... .

If you are studying currently, what is your main course?

- Interaction Technology.
- Data Science.
- Business Information Systems.
- Creative Technology.
- I am not studying.
- Other ... .

A.3 Mann Whitney U Test

To see if the manipulation during the experiment worked, it was aimed to execute a one-way analysis of variance (ANOVA) with two groups or a t-test, since for two groups, the results are the same. The requirements for an ANOVA are independent groups and normal distributions of the dependent variable attributed predictability. By research design, the groups are independent. In this study, more than five independent samples exist. Normal distribution is given for four of the five conditions. For the predictable condition, the data was not normally distributed. For testing statistically if the manipulation worked, only the two samples of C-Predictable, the predictable robot behavior, and C-Unpredictable, the unpredictable robot behavior, were relevant. The data sets per group were small, so the risk of false positive results is increased when using an ANOVA with data which is not normally distributed.

As a result, a non parametric test was executed. Non parametric tests are less statistically powerful but allow making statistical inferences without assuming normal distribution. The Mann-Whitney U test is a reasonable non parametric alternative for a test with two independent groups. The requirements for this test are: 1) the dependent variable must be continuous, 2) data is non-normal, 3) data is shaped similarly (if medians are to be analyzed), 4) independent samples and 5) a minimum of five observations per group. 1), 2), 4) and 5) are given by research design. The analysis in SPSS showed that the distributions per group do not have the same variability, so 3) is not given, but medians are not compared, so this is irrelevant.

A.4 MANOVA and MANCOVA

For executing a MANOVA, ten requirements need to be given: 1) independent samples, 2) interval scaled dependent variables, 3) nominal scaled independent variables, 4) per group at least as many cases as dependent variables, 5) linearity between the dependent variables for each group of independent variables, 6) no multicollinearity, 7) homoscedasticity, 8) homogeneity of covariances, 9) multivariate normal distribution and 10) checking for univariate or multivariate outliers. 1-4) are given by research design. The other requirements are described in the following.

In a first step, the data was controlled for univariate outliers. There was one extreme outlier (case 22) in the data. Due to the small sample size and since the answer was plausible, the case not excluded. SPSS does not offer an option to test for multivariate normality of distributions. As a result, the univariate normality was tested. If univariate normality is given, it is assumed that multivariate normality is given as well, since univariate normality is a requirement for multivariate normality. The Shapiro-Wilk test is more suitable for small samples, so it is interpreted regarding normality. All groups were normally distributed across all three dependent variables, as assessed by the Shapiro-Wilk test (α = .05). Another requirement for executing a MANOVA is a no multicollinearity. On the one hand, multicollinearity leads to logical and statistical problems because this indicates that variables are redundant since they measure the same aspect. This weakens the analysis through increasing errors. On the other hand, if multicollinearity is too low, statistical power is reduced as well. The optimal value seems not to exist yet. Research indicates r $<$ .90 as a reasonable threshold. The analysis shows that correlations between dependent variables were low (r $<$ .90), indicating that multicollinearity was not a confounding factor in the analysis. Another requirement for the MANOVA is checking for multivariate outliers. To achieve that, the Mahalanobis distance can be interpreted. The cut-off value for three dependent variables is 16.266. The highest Mahalanobis distance in the data set is 3.19. As a result, no multivariate outliers were found (p

$>$ .001). As a next step it is necessary to check for linearity, since when there is no linear relationship between all pairs of dependent variables, the statistical power is reduced. Linearity could not be observed for the dependent variables through visually analyzing scatter plots. Thus, the statistical power of the MANOVA is reduced. Furthermore, it is necessary to test for homoscedasticity. This was done using the Levene test. There was homogeneity of the error variances for all dependent variables (p $>$ .05). It is necessary as well to test for homogeneity of covariances. This was done using the Box test. There was homogeneity of covariances (p $>$ .05).

For a MANCOVA, another requirement exist: the covariates have to be continuous. This is given by research design.

A.5 Further Explanation of Missing Values for Reliability Testing

In the following it is reported on items of the trust questionnaire and the attributed predictability questionnaire which are lacking more than three entries.

For the ”reliable” subscale of the trust questionnaire, no answers were missing and internal consistency was questionable (Cronbach's $\alpha_{PT-reliable}$ = 0.68). For the ”competent” subscale, the item ”meticulous” showed 6 missing values due to not knowing the word and 4 because of a perceived missing fit. Internal consistency of the items was acceptable (Cronbach's $\alpha_{PT-competent}$ = 0.73), but 11 cases were excluded from the test.

In the ”ethical” subscale data, for ”ethical”, 7 participants chose it does not fit, for ”principled”, 6 chose does not fit, for ”has integrity”, 9 chose ”does not fit” while 3 did not know what the item means and for ”moral”, 9 thought it does not fit. This results in four missing values for the whole subscale, so the ethical subscale has only 21 entries instead of 25. The reliability test indicates questionable internal consistency (Cronbach's $\alpha_{MT-ethical}$ = 0.62) and 13 cases were excluded. For the ”ethical” subscale, a pattern is observable in the data: four participants did not rate any item on the scale of 1-7 and two participants did only rate one out of four items of the scale. There are no typical commonalities about these participants in the data. Their robot experience ranges from 2 (”I know robots only from fictional stories (movies, books, TV-shows) and from real-life formats (TV news, newspaper)”) to 5 (”I have interacted with physical social robots like Nao, Pepper, or Sophia.”) with an average of 3.67, their dispositional trust ranges is relatively high, ranging from 3.33 to 5 with an average of 4.17 and their negative attitude towards robots is neutral, ranging from 2.46 to 3.29 with an average of 2.78.

In the ”transparent” subscale, for ”sincere”, 4 people thought it does not fit. For

”candid”, 5 thought it does not fit and another 5 did not know the word. For ”is genuine”, 5 people thought it does not fit. The reliability test indicates poor internal consistency (Cronbach’s $\alpha_{MT-transparent}$ = 0.57) and 15 cases were excluded. While excluding ”transparent” from the scale would mathematically increase $\alpha_{MT-transparent}$ by 0.2 to $\alpha_{MT-transparent}$ = 0.77, this step is not done because ”transparent” is the only item with sufficient answers (24/25) and many cases of the scale were excluded.

In the ”benevolent” subscale, for ”considerate”, four people thought it does not fit. For ”benevolent, 17 people did not know what it meant. For ”has goodwill”, 8 participants thought it does not fit. The reliability test indicates good internal consistency (Cronbach’s $\alpha_{MT-benevolent}$ = 0.84), but 20 cases were excluded due to missing values.

In the attributed predictability questionnaire, answers are complete for the first four items ”unpredictable”, ”irregular”, ”inconsistent” and ”random”. For ”variable”, two participants did not know what the word meant. For ”erratic”, five participants did not know the meaning. There is no pattern observable for not knowing something in the data. For attributed predictability, the reliability test indicates excellent internal consistency (Cronbach’s α_{AP} = 0.93).