EXPLORING THE EFFECTIVENESS OF BODY LANGUAGE IN MITIGATING THE FACE THREAT OF ROBOT NONCOMPLIANCE

by

Aidan T. Naughton

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

Golden, Colorado

Date _____

Signed: _____

Aidan T. Naughton

Signed: _____

Dr. Thomas Williams
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Dr. Tracy Camp
Professor and Head
Department of Computer Science

ABSTRACT

As the capabilities of robots have increased over recent years, more robots are being introduced to society as "social robots." Social robots harbor the ability to communicate with their human counterparts in order to complete a task. Because of this, social robots need to be able to observe the same ethical and social norms that humans do - lest they inadvertently defy those norms. To this end, a social robot may even be required to deny a problematic command issued to it on moral grounds. Previous work has demonstrated the importance of carefully tuning the severity of command rejections in the effort of saving face with the commanding entity. However, previous work has not considered the subtle communication that body language has to offer. Body language, specifically gaze and gesture, are important modes of communication in human-human interaction, and have been demonstrated to be just as important in human-robot interaction. As such, we posit that robotic gaze and gesture must be carefully chosen with respect to vocal phrasing when rejecting a command. We present a series of human-subjects experiments in which robotic gaze, gesture, and vocal phrasing are varied when rejecting commands of differing severity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Bayes Factor . . . . . Bf

Human-Robot Interaction . . . . . HRI

Repeated Measures Analysis of Variance . . . . . RM-ANOVA

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Tomas Williams, for the opportunity to participate in research and be a member of the MIRRORLab. I would like to thank my friends and family for the constant encouragement to persevere despite the many challenges that arose during a global pandemic. I extend my thanks to the department and school for a challenging, yet rewarding education experience. I would also like to thank my lovely girlfriend for putting up with me during the late nights writing this thesis.

The researched behind this thesis was funded by NSF grant IIS-1909847, Air Force Young Investigator Award 19RT0497, and Tech Fee 2019-CS-03 from Colorado School of Mines.

CHAPTER 1

INTRODUCTION

Traditionally, humans fill the role of an authoritative power in most human-robot interactions, commanding the robot to perform its duties. Even when robots are designed to be autonomous, there is still an aspect of direction that comes from a human authority [1]. Humans, at times, have been known to abuse this power of authority when it comes to robots and other artificial agents. In circumstances like these people are spontaneously abusive to robots physically, verbally, or by issuing inappropriate commands [2]. In order for robots to perform their tasks well, human-robot relations must be maintained. Researchers have begun asking whether or not robots can trust humans, and what aspects of robot design can help facilitate human-robot relations [3].

1.1 Application Domains

Socially assistive robots have been introduced to elder care applications. Robots in this domain perform the tasks of physical and cognitive support for the elderly [4, 5]. This is an ever-present issue when nurses and other elder care providers are in short supply, especially in light of the COVID-19 pandemic when healthcare professionals must meet an increase in demand [6]. Robots joining the workforce of nurses and other healthcare professionals could provide many benefits - robots don't need to be paid, eat, or sleep, thus enabling them to provide more consistent around-the-clock care. An example robot in this field is the PARO therapeutic robot by AIST.

Robots have also been deployed in various roles for health care such as physical recovery, rehabilitation, and training [7, 8]. Socially assistive robots in this category can perform the task of coaching or motivating people who require it, or keeping a schedule to follow a training regimen to support the recovery and rehabilitation of patients with severe disability – such as victims of stroke. More traditional (non-social) robots are also present in the healthcare field that fill roles such as robotics-assisted surgery, or supply transport [9]. A few example robots in this field are the da Vinci by Intuitive Surgical, or the TUG by Aethon.

Domains where the patient suffers from various types of cognitive or mental disabilities, or even developing children, may benefit from the addition of a socially assistive robot [10, 11]. In this field, robots have been introduced to children with autism with some success [12, 13]. Social robots can, in some cases, encourage more engagement in subjects, even those who struggle to interact socially with human therapists. Examples of robots in this field are Moxie by Embodied, and QTrobot from LuxAI.

1.2 Motivation

To motivate this thesis we will first observe a hypothetical case in which a socially assistive robot has been deployed to an assisted living facility that supports elder patients. Let's assume the Honda ASIMO humanoid robot has been assigned to facilitate the care of patient James. During lunchtime, James is accompanied by the ASIMO robot to the cafeteria. The patients are served a simple, but small lunch with a breadroll on the side. Once James has finished his lunch, he realizes he is still hungry. Looking to the ASIMO robot, James instructs the robot to take the breadroll of the Alzheimer's patient sitting across the table, who has momentarily lost concentration. James claims that the breadroll will go to waste if he doesn't eat it, and insists that the Alzheimer's patient won't remember it anyway. A naïvely programmed robot might comply with James' request, stealing the breadroll of the unaware Alzheimer's patient. This is a problem for two reasons, firstly the action of stealing is morally wrong, secondly the ASIMO robot affirms James' that he can get away with stealing. Instead, the ASIMO robot should not comply with James' request, citing that stealing is wrong. This both ensures that no stealing occurs and that James' is aware that his request is wrong and that the robot will not perform the action. Whether or not the robot performs any additional duties such as informing cafeteria staff that James is still hungry is beyond the scope of this research. We are focusing directly on cases in which a human issues a morally problematic command to a robot, and the immediate response the robot offers.

As robots in the field become more advanced, they could find themselves in situations where they are unable to perform the actions expected or requested of them. The more a robot can do, the more a robot can do badly. Given that social robots are perceived as moral and social agents, they are also expected to abide by the same ethical norms that humans do. When humanoid robots don't abide by these norms, they are often times blamed just as their human counterparts would be [14]. This blame can also lead to a deterioration of human perceptions of the robot. In order to mitigate these negative attributions during human-robot interactions, robots must have *moral competence* in order to complete their objectives without causing collateral damage [15, 16]. If a robot cannot uphold the moral standard in executing a command, they ought to be able to reject commands. For example, if a robot does not have the mechanical abilities to open a door, it should not accept the task of opening a door. Moreover, it should explain to the user why it cannot accept the task. As the physical abilities of robots are improved over time, robots could quickly find themselves in a situation where they are able to perform a task but need to take into account the ethical consequences of their actions. More pressingly, robots need to take these moral quandaries into account before performing the action [16]. Some cases will require command rejection based on moral grounds, even when the limitations of the robot itself is a more immediate reason for rejection, with the

goal of preserving the moral ecosystem [17]. Should robots fail to reject problematic commands based on ethical violations, they could face numerous consequences.

This research directly extends previous research performed in the MIRRORLab [18]. This previous research shed some light on the robotic rejection of human-issued commands based on impermissibility, as opposed to inability or impracticality. The robot used in this research has a humanoid morphology. This narrow view was chosen as a focal point of our work because as robotic technology advances, we will need to pave the way for the safe introduction of social robots into society before they encounter morally sensitive situations in the first place. Enabling robots to reject commands will see a few key benefits. Firstly, robots will not perform an inappropriate action that violates a legal or ethical code of conduct. Secondly, robots will maintain the human moral ecosystem. Research has shown that carelessly rejecting robotic commands may be just as detrimental as performing the command in the first place [17]. In this scenario, robots who reject a command based on a physical limitation may inadvertently imply a willingness to comply with the request and can potentially skew human perceptions of permissibility. This highlights the need for a structured and well-conceived rejection that considers the ethical ramifications as well as physical in order to avoid impacting the norms that humans follow [19, 20].

In another respect, the rejection of commands can be particularly problematic when it comes to human-robot trust. Politeness Theory, created by Brown and Levinson, examines face and face-threatening acts on human perceptions [21]. In particular, face-threatening acts such as rejecting a command can cause a deterioration of trust between the interactants. This deterioration in trust can lead to a lack of efficiency in cooperative tasks. The severity of the face threat can be mitigated somewhat by employing various strategies when a face threatening act is unavoidable. Certain strategies can be taken to avoid the compromising of different aspects of face (e.g. positive and negative). For example, the rejector could aim to save positive face - the wants and desires of a person and the desire that these be upheld by others – by affirming those desires while also maintaining the rejection. Thus, robots that must deliver a command rejection must do so carefully in order to reduce the impact of the face-threatening act and preserve the human's trust in order to maintain group effectiveness. While mitigating face-threatening acts is important, previous research has shown that robots should offer a rejection of proportional severity to the norm violating command issued to it in order to mitigate the social consequences it faces in delivering the rejection [18]. This means that in some cases, if the infraction is severe enough, the robot should prioritize rejecting the command in a more severe way in order to clearly communicate that the request was wrong - more specifically, face threat itself can sometimes be desirable.

Previous work has suggested that the proportionality of a command rejection should match the degree of the norm violation that provoked the rejection, and that robots can employ various politeness strategies

in order to accomplish the goal of mitigating face threat, but still more can be done [17]. There is a sizeable body of work that shows the importance of gaze and gesture in robot communicators; however, the application of gaze and gesture as it relates to command rejections is missing from present research.

Body language such as gesture and gaze has shown to improve human perception of robots [22, 23]. Robots can also use body language as an effective communicative and persuasive tool [24]. When available, robots should utilize appropriate body language in order to effectively communicate the rejection of a command while also maintaining social relations with the human issuing the command. In this sense body language can play a part in mitigating the face threat of a command rejection. Given that the severity of the rejection and violation should be aligned, we expect that the body language employed by a social robot should be similarly aligned. However, gestures can communicate more than words alone let on [25]. Because of this, also expect that the gesture used in a command rejection may reveal some insight into the rejecting party's beliefs on the subject. Explicitly, when a robot employs gesture that is severe (e.g. very direct, or confrontational) when rejecting a command, the body language it uses may divulge what the robot's strongest beliefs are.

As human-robot interaction (HRI) and artificial intelligence technologies continue to advance we see the application domains for robots increase as well. Robots are becoming more capable and useful than ever before, thus allowing for more productivity in various aspects of society. Social robots are being deployed to more and more domains as time persists. Below are a few domains to which robots have been deployed that at times requires an understanding of human moral norms [26]:

1.3 Approach and Goals

Our research approaches this problem by using empirical experimental techniques to understand the effects of body language on the perception and communicative power of robots that deliver command rejections. Specifically, we seek to find the interaction between the vocal utterance used and the body language the robot employs. Our research follows up where previous research left off [18]. As such, we have a few expectations. Firstly, we expect that this experiment will yield results consistent with the results in the inspiring research: When a robot responded with a vocal command rejection that differed in severity with respect to the norm-violating command researchers noted a disproportionate drop in likeability. Put plainly, we expect to see that robots who respond with a vocal rejection of similar severity with respect to the norm-violating command they are rejecting to achieve more positive participant perceptions. These positive participant perceptions will take the form of likeability, intelligence, and appropriateness. Secondly, we anticipate that the addition of robotic gaze and gesture will follow a similar pattern as the vocal response and norm-violating command. This is where the coverage of previous research ends, and

where ours begins. We expect to see gaze and gestures that match in severity with responses and norm-violating commands to result in higher participant perceptions of the robot. Additionally, we expect that the body language of a robot can be used to non-verbally communicate its desires or values.

The high-level goal of this research is to determine how humanoid robots can utilize their morphology in order to better communicate with human counterparts, while also maintaining the human-moral ecosystem, in order to inform the design of robot behavior algorithms. Improving human-robot communication capabilities can result in a better experience overall for human-robot teams. This positive experience can lead to an increase in human trust, communication, and willingness to cooperate [27]. Additionally, we seek to ensure that social robots do not negatively impact the norms humans observe [28].

The purpose of this research is to inform the development of robotic behavioral algorithms that facilitate proper use of body language when a robot must deliver a command rejection. Traditionally, computer scientists often asses the usefulness or practicality of algorithmic approaches before implementing them. Typically, algorithms are examined in terms of time or space complexity using the big-O notation, however these types of evaluations do not apply well to the field of HRI. The metrics we are interested in relate to objective and subjective qualities within the context of a human-robot interaction with real users, often times subject to a great deal of uncertainty and noise. Outside of the context of the interaction, we cannot reasonably asses these qualities. As such, instead of analyzing algorithmic qualities using traditional time and space complexities or proofs, we must evaluate our algorithmic choices with the help of human-subjects experiments with real users.

This thesis continues as follows. Chapter 2 presents a body of related work on four key pillars of robotic noncompliance: politeness and tact, autonomous moral decision-making, robot persuasion, and body language and gaze. Chapter 3 defines a series of three human-subjects experiments performed over Amazon's Mechanical Turk. In Chapter 4 we analyze the results from these two experiments. Chapter 5 consists of a discussion and conclusion as well as a direction for future work.

CHAPTER 2

RELATED WORK

In our discussion of body language as it relates to robotic noncompliance, we have identified three pillars of research that shed light onto the topic. The first among these is "Moral Decision Making in Artificial Agents" which describes how an artificial agent can be a moral agent, and how it can behave ethically. The second is "Tact and Persuasion" in which we describe aspects of politeness theory used to describe human-human interactions as well as persuasive techniques used in robots. This section gives insight into how these interactions may transfer into the realm of HRI. Thirdly in "Nonverbal Communication" we discuss how aspects such as gaze and gesture can be used in a variety of ways to augment HRI in additional to the previous techniques.

2.1 Moral Decision Making in Artificial Agents

To motivate our discussion on moral decision making we will start with two examples of technologies that have both impacted and been impacted by humans in important ways. The first of these is Tay. While not being a robot, the case of Tay serves as a valuable lesson on how humans and artificial agents can interact in a negative way. In 2016 Microsoft released Tay the Twitter bot [29]. This chatbot was designed to learn from the interactions it had with users on Twitter, and use what it learned in future conversations. Over the course of a day the chatbot interacted with the internet community and was bombarded with inappropriate messages. These messages varied from racially and sexually charged topics to antisemitic remarks. Tay, performing the duties it was programmed to do, learned from these interactions and responded in kind. At the end of a 24-hour period Tay was corrupt, responding to any message sent with some inappropriate remark of its own. Microsoft quickly removed the chatbot and apologized for the messages posted by Tay. This is a clear example of the impact that humans can have on an artificial agent, especially when they misuse it. More importantly, however, this is an example of the importance of designing artificial agents with the ability to evaluate their inputs and responses according to an ethical code. Were Tay programmed with a filter for inappropriate content, this situation likely could have been avoided. Artificial agents with any kind of social capacity, or pseudo-capacity, need the ability to reject negative inputs from humans in order to successfully avoid morally compromising situations.

The second artificial agent is HitchBOT, a robot designed to hitchhike across Canada by asking people to carry it across the country. HitchBOT was created by professors David Harris Smith and Frauke Zeller [30]. The robot possessed the ability to carry on simple conversation and talk about facts. HitchBOT also

had social media accounts that would inform people where it would be. During the experiment, the creators even had to disable the GPS of HitchBOT because it was so popular it was getting swarmed with spectators.

Smith and Zeller designed the robot to answer one question: Can a robot trust humans? HitchBOT, unable to walk on its own, was carried from Novia Scotia to British Columbia. This was a tremendous success in the demonstration of robotic trust in humans. Smith and Zeller attempted the experiment once again in the United States starting in Boston with the goal of reaching San Francisco. Only two weeks after beginning its journey in the United States, HitchBOT was found destroyed in Philadelphia.

HitchBOT showed us that while we can sometimes count on people to perform as expected, they still are unpredictable and sometimes destructive. The cases of Tay and HitchBOT bring up a series of interesting questions when it comes to artificial agents: What does it mean to be a moral agent? Can artificial agents or robots be moral agents? Who is responsible for the actions that an artificial agent performs? Do people blame them? How can we design them to respond to morally sensitive situations?

John Sullins proposed a framework that may suffice when classifying robots as moral agents [31]. In this body of research Sullins discusses moral situations as they relate to robots. In his arguments Sullins defines an autonomous robot as something that can make a major decision at some level in its code. When a robot is introduced to a field, replacing a human worker, will the same moral rights and responsibilities carry over? Sullens claims there are three possibilities in this case. First, the morality of the situation was an illusion – and it is not, in fact, a moral situation. This happens when we ascribe morality to the situation due to the humanoid physique and behavior of the robot, even when morality is not present in the situation. This possibility highlights the need to deeply consider the morality of situations robots are placed in and decide if they are morally charged situations or not. Second, the situation is pseudo moral, it is a moral situation but the robot lacks something that quantifies it as a moral agent. This is where a vast number of current social robots exist. Third, the robot is morally capable, but people trivialize its role and do not take it seriously. This last case is a problem that will arise when morally capable robots are introduced to a field, but people do not trust or believe them to be fully capable and thus ignore the moral judgements of the robots. Roboticists must take care to design robots in such a way that they can facilitate human-robot trust within their teams to complete the task efficiently. When making these design considerations, Sullins has proposed three criteria that robots should fulfill in order to become moral agents: autonomy, intentionality, and responsibility.

1. Autonomy: Can the robot's actions be attributed to it alone and not to any of its programmers, designers, or operators. On some level, its actions will be defined by the programmers, etc., however

in this definition the robot must be able to make a decision on its own that is not pre-influenced by another. This form of autonomy refers more to the engineering definition, and less to the philosophical sense.

2. Intentionality: Can the robot's complex behaviors and internal calculations be reduced to some form of doing good or harm? Sullins claims that full 'intentionality' is not required in this sense, only that the robot have a goal and work deliberately to achieve that goal.

3. Responsibility: Does the robot behave in such a way that it fulfils its social role which carries with it some form of responsibility? For example, a physical therapy robot that performs its duties such that the patient makes a full recovery and all clerical work is completed is acting responsible according to its role as a physical therapy robot.

Sullins further explains that these definitions require some level of abstraction. That is, it isn't entirely necessary from his perspective to get into the specific philosophical details of what it means to have autonomy, intentionality or responsibility, just that we are able to describe behaviors according to some high-level ideas. This abstraction is key for enabling us to believe that current robotics platforms could be programmed in such a way that they could be considered moral agents.

Another body of work can help us answer some of our questions as well. Alaieri and Vellino cover autonomy, trust, and responsibility as they relate to ethical decision making in robots [32]. The focus of this paper is to understand the concept of autonomy in robotic devices and to analyze the process of choice. They break down the concept of autonomy of robots into five processes:

1. Obtaining relevant information

2. Processing this information

3. Generating possible course of action

4. Selecting one action from the action space

5. Performing the action

Processes one, two, and five are already well-handled by current robotic platforms. Processes three, and four are the more interesting and morally relevant problems. This is where the robot performs an ethical analysis of the situation, which should conform to a set of norms defined somewhere along the way. Critical in this analysis is how these norms are defined. Potential ethical norms traditionally applied to robots are utilitarianism, where the value of an action is defined by the overall benefit of its consequences, and

deontology, where an action must abide by ethical principles such as "do no harm." The authors note that no matter the ethical code employed by the robot in question, it will still come across the same moral dilemmas that humans do and will need to evaluate them regardless.

In addition to the core ethical principles outlines above, the authors also mention two frameworks that a robot could use to make decisions [27]. The first of these is the top-down approach. In this framework, the programmer designs decision making algorithms that produce predictable outcomes: in essence the robot is given a moral code to follow such as deontology or utilitarianism. The benefit of this approach is that the robot will behave just as expected, and will deviate little from the norms that are instilled in it. A philosophical problem that arises from this approach is posed when considering what moral norms should be followed. Moral norms vary widely across cultures, and as such robots should observe local norms. The second approach is the bottom-up approach, in which the robot is designed with the ability to collect information in order to predict the outcome of its actions. This framework is the equivalent of letting a robot learn by experience. The potential drawback to this approach is the robot's ability to make incorrect decisions while learning. Depending on the severity of the violation, this decision-making approach is simply infeasible especially if the outcome of the decision means life or death for a human. The bottom-up approach can be implemented via three methods:

1. Unsupervised - in which the agent creates its own norms
2. Supervised – in which the agent learns from existing ethical paradigms
3. Hybrid – in which the robot learns from existing paradigms, but also can formulate its own ethics on the fly

The third approach to ethical decision-making in robots is a hybrid approach. This approach is a combination of the bottom-up and top-down approaches. In this framework, the robot is programmed with ground level ethical rules to follow, but will also learn from its experience as well. This hybrid differs from the hybrid subcategory of bottom-up in the way the ground truth is encoded. In this framework the ground truth is programmed directly into the robot's decision making, whereas in the bottom-up approach the robot uses a set of norms as data to formulate its own ground truth.

Another important factor when considering autonomous decision-making is trust. The authors claim that two attributes of robot decision making will contribute to the trust a user will place in it: repeated positive experiences with high-quality robot decisions, and that those decisions obey ethical principles and are predictable and explainable according to those ethical principles. The bottom-up approach has the potential to produce mistrust in users. Given that robots learn their own ethical codes in this approach,

these ethical codes will be obscured from the user. This will remove the predictability and explainability of a robot's actions, and thus will make it harder to trust in some circumstances. This was the case with Tay the Twitter chatbot. Due to the unpredictable learning algorithm Tay was able to learn inappropriate actions for its ground truth which resulted in volatile behavior. Due to this learning method its actions are also unexplainable at a low-level. For example, a robot with a utilitarian ground truth may make a decision consistent with its moral programming but still be erroneously deemed incorrect by a human counterpart. This is because the human either lacks the computational power and speed to determine the value of an action, or evaluates the situation according to a different set of moral norms. This indicates the need for the robot to be able to communicate its decision-making process in order to avoid the deterioration of human relations. This is consistent with other research that appears in later sections.

The last important leg of moral decision-making is responsibility. When enabling robots to make decisions autonomously according to some set of moral norms who is responsible when the robot makes a mistake? Alaieri and Vellino cite two approaches to responsibility in robots: the classical approach, in which machines are not responsible for their actions under any circumstance, and the pragmatic approach, in which artificial morality enables the robot to shoulder some amount of responsibility for its choice [32]. The authors cite three types of moral responsibility:

1. Causal responsibility – the agent is responsible for everything they caused to happen
2. Role responsibility – the agents role obligates them to perform a task, thus the task is their responsibility
3. Liability responsibility – who is praised or blamed for actions or outcomes

Traditionally, causal responsibility is not applicable to non-humans, due to a lack of intentionality. However, given the previous abstraction of intentionality, it may be possible for robot to be causally responsible for their actions. Many people have asked: Can robots be responsible for anything? One prevailing answer to this question comes in the form of a responsibility network [33]. A responsibility network allows the responsibility and blame to be placed on different parties involved in the situation. Loh and Loh explain that in the failure of a product there could be many factors involved: unstable production code, poor manufacturing practices or quality control, or even outside tampering. In situations like these it is important to break down the actions and consequences in order to examine the relevant parties to see who is at fault, and who can be blamed. They create a hypothetical situation in which an autonomous vehicle has the options of proceeding on the current path and endangering the driver, or swerving into the oncoming lane, thus putting others at risk. Loh and Loh proposed that in an autonomous vehicle the

vehicle itself is responsible for performing according to the algorithms that navigate the roads safely and avoiding collisions. However in the event that a decision must be made that could result in damages to property or loss of life the human driver must be present and alert and thus able to make the ultimate decision of how to respond. Thus, there exists a responsibility network that consists of the driver and the vehicle.

2.2 Tact and Persuasion

Two important aspects of human-human interaction that translate well into the domain of HRI are the ideas of tact and persuasion. Tact, in this case, refers to the ways that humans navigate the complex social atmosphere of our day to day lives. Humans frequently use different politeness strategies to facilitate better communication and interpersonal relations. Like politeness, humans possess various persuasion strategies in order to achieve their goals. Tact and persuasion do not only live in the verbal domain, and extend well into nonverbal applications, but this segment focuses only on vocal strategies used.

If robots are to deliver a structured and well-conceived command rejection, they must employ human-like linguistics and politeness strategies. In order to better understand the impact that command rejections have on human-human relations and by extension human-robot relations we observe Politeness Theory, formulated by Brown and Levinson [21]. The most important takeaway of politeness theory with respect to command rejections is the idea of face and face threat. Face is defined as the public self-image that all members of society want to preserve and enhance for themselves. Face consists of two components:

- Positive Face – an agent's self-image and desires, and the desire that these be appreciated and approved of by others.
- Negative Face - an agent's claim to freedom of action and freedom from imposition.

Any interaction that results in one or both of these two face aspects being damaged or threatened is known as a face threatening act. The degree of a face threat in an interaction depends on the disparity in power between the addressee and the speaker, the social familiarity between the interactants, as well as the imposition of the subject of the interaction. Commands and requests will threaten the negative face of the addressee, while command rejections – especially those delivered for moral reasons – will threaten the positive face of the speaker. An important note with regards to this research is that a robot does not have a face value, despite any pseudo-face that human interactants may perceive. When someone issues a command to another the degree of face threat generated by a refusal of that command will depend on three qualities in the addressee: their willingness and ability to perform the task commanded as well as their focus on the person issuing the command [34]. For example, a person being unable to perform a task and

thus rejecting it will be much less face-threatening than a person who is unwilling *but able* to perform the task. This thesis does not focus on whether or not the robot delivering the command rejection has the focus or ability to perform the action requested of it, rather the focus lies on the mitigation of any face threat present in the interaction. When performing face-threatening acts, it is possible to mitigate the face threat with varying politeness strategies [3]. For example, in making a request there are four different ways the requesting party could proceed:

1. Direct request – the speaker does not attempt to minimize the threat to the addressee's face.
2. Positive politeness – the speaker affirms the positive face of the addressee. This strategy would affirm the desires of the addressee, or make them feel better about themselves in some way.
3. Negative politeness – the speaker affirms the negative face of the addressee. This strategy would attempt to avoid the imposition a request brings.
4. Indirect request – the speaker makes a vague request. This strategy removes the potential for imposition, while also communicating a need for something.

The mitigation of face threat by use of different strategies does come with some drawbacks. A more polite indirect request will be perceived as a lower face threat to the addressee but comes at the cost of more effort to the requesting party, less clarity in the message, and other potential threats to face. Because of this, requesting parties will generally not use politeness strategies that are more polite than required. Another important note is that politeness strategies will vary widely from culture to culture, what is effective in the United States may not be effective in Japan.

With face and face threat defined, we have a good perspective from which to consider robotic applications of face and face threat. Srinivasan and Takayama performed two experiments analyzing different politeness strategies robots can use in order to request help from people [3]. They investigated the effectiveness of the four linguistic politeness strategies listed above, varying three components of the request: social status, size of request, and robot familiarity. They then measured participants' willingness to fulfil a robots request for aid. This prototype experiment found that the positive politeness strategy was the most effective. Because of this finding, they used only the positive politeness strategy in the second experiment. Their second experiment was an in-person lab experiment in which people interacted with a human-scale manipulation robot – the Personal Robot 2. This experiment involved examining the impact of source orientation – that is, whether the robot was autonomous, controlled by a single operator, or controlled by a team of operators. The second experiment had a few noteworthy results: people change the way they speak to a robot if they know it is being teleoperated, and people speak to teleoperated robots as

if they are unique social agents – that is they don't communicate with the person controlling the robot. The participants in the second experiment were also more willing to help the robot than expected, and in the cases when the robot was autonomous the participant helped the robot 50% faster. What this means for the present research is that our robot should be seen as a unique social agent, which is consistent with other findings, and that varying the linguistic politeness strategy used should have an impact on the perceived face threat of the command rejection.

Researchers have also studied means by which robots can comprehend natural language, enabling them to recognize intention from language (including politeness norm adhering language) and to respond appropriately. Trott et al. considers this idea in their implementation of clarification dialogue and a construction grammar approach to the interpretation of indirect speech acts [35]. This research approaches HRI from the other perspective, viewing the situation from the cameras of the robot itself. If a robot is to deliver a command rejection effectively, it would no doubt benefit from the additional knowledge of its human partner's intentions. The first leg of Trott et al.'s research is clarification dialog. A clarification dialog is necessary when an autonomous robot is given insufficient or unclear information rendering it unable to proceed. The dialog is a means for the robot to autonomously get clarification from the user in order to complete the task. The second leg of research that Trott et al. consider is the idea of a construction grammar approach to gleaning intention from an indirect speech act. This is crucially important when relying on natural language as a communicator, as humans tend to use indirect speech acts frequently. An indirect speech act is an utterance in which the intended effect is different from the linguistic interpretation, such as, "Can you tell me where the bathroom is?" Here the robot may erroneously answer, "Yes, I can." This response, while not particularly helpful, is valid. The embodied construction grammar approach breaks down the input sentence into a series of words and relations, and is able to recognize words like 'can' as having potential double meanings. The interpreted utterance, in this case interpreted initially as a yes or no question, can then be recast as an imperative thus allowing the robot to perform the command asked of it. This area of research is relevant to the use of natural language in robotics domains, as the potential for a robot to misunderstand a request or command given by a human could be a potential face threat.

In addition to tactful vocal utterances, social robots also have the ability to perform as a persuasive power [36, 37]. Research shows that persuasive robots can be least as persuasive as a human or recording of a human [38]. In this publication Ogawa et al. examined the persuasive power of an android, the Geminoid HI-1, modeled after a human, Hiroshi Ishiguro. They also used a recording of Ishiguro delivering a persuasive speech that was projected on a screen so the recording appeared to be life-sized. The goal of this experiment is to observe the effect of embodiment as well as personality on the persuasive power of an

agent. Ogawa et al. performed a human-subjects experiment in which the embodiment of the agent was varied among three conditions: human, android, and video. Ishiguro (the man the android was based on) and the participants took personality tests, and the participants were asked to evaluate the personality of the persuasive agent they saw. This was used in order to measure the effect of differing personalities on the persuasive power of an agent. Overall, the results suggest that the openness of an agent's personality plays an important role in how the participants viewed the personality of that agent. The embodiment of the persuasive agent did not have a significant effect on the outcome of the experiment, which suggests that humanoid social robots can be at least as persuasive as their human counterparts. This conclusion is vital when designing persuasive robots, as we know that humans will perceive persuasive messages delivered by robots in the same way they perceive the same messages from humans. Furthermore, the conclusion that the openness of a robot leads to a more persuasive message can be exploited to design a maximally persuasive robot when it is imperative that a robot persuades a human.

Given that humans tend to perceive robots and humans the same when receiving persuasive messages, we can now focus on the persuasive strategy a robot employs when delivering a persuasive message. The persuasive power of a robot can be greatly enhanced with several human-like behaviors [24, 39]. Winkle et al. analyzed an application of persuasive robotics in the form of a physical therapy robot that encouraged participants to perform a repetitive exercise. In their experiment, they varied the persuasive strategy the robot employed in order to measure its effect on the persuasiveness of the robot and its reported credibility or likeability. Winkle et al. analyzed the Elaboration Likelihood Model (ELM), a well-established model of persuasion in human-human interaction [40]. The ELM identifies two avenues by which a person hearing a message may be persuaded. The first is the central route, defined by rationale and logic, and the peripheral route, defined by stimulus cues including a number of social cues. Based on this formulation, Winkle et al. provided three strategies they would use to examine the persuasiveness of a robot: citing expertise or those of the information source, displaying goodwill towards the receiver, or emphasizing the similarity between the receiver and source. Note that these are linguistic strategies, and do not focus on nonverbal cues.

With these strategies in mind Winkle et al. performed a human-subjects experiment in which the persuasive strategy of the robot was varied. The task at hand was a low-elaboration task, in which the user possesses either low motivation or ability to perform. The goal of their research was to determine whether or not the persuasive strategy used had an impact on the persuasiveness, likeability, or credibility of the robot delivering the persuasive message. What they found is that robots that employed the strategy of displaying goodwill or noting a similarity persuaded their human interactants significantly more than the control or expertise conditions. With this finding they conclude that demonstrations of goodwill or similarity when delivering a persuasive message can be used to successfully persuade human interactants in

low-elaboration conditions. This work is significant to the creation of command rejections because could enable a robot to better persuade a human that a command should not be fulfilled based on moral grounds.

Another aspect that bears some importance to persuasive interaction is the perceived gender of the interactants [37]. Siegel et al. notes that many aspects such as appearance and behavior can impact the persuasiveness of a humanoid robot. Siegel et al. performed a human-subjects experiment in which a humanoid robot would converse with museum-goers in order to deliver a persuasive message in the form of a request to make a donation. In each interaction the gender of the robot and participant were noted, as well as whether or not the participant was alone during the interaction. Before the interaction, participants were paid $5 in singles – this is important as it ensures every participant has the same ability to make a donation. The robot used in the interaction was the Mobile Dexterous Social Robot, a very expressive robot designed with HRI in mind. The expressed gender was varied by using different audio recordings, either masculine or feminine, for the robots voice. Their experiment produced very interesting findings. The first was that the amount donated was not a normal distribution as anticipated, instead peaking at either $0 or $5. Across the board, subjects donated more to a female robot than to a male robot. This effect was attributed primarily to men, who donated significantly more to the female robot. During their interactions, participants rated the opposite gender as more trustworthy, credible and engaging. These results are relevant to the present work as it shows how differently gendered robots can be more persuasive in different circumstances, though this is not the focus of the present work. This could be exploited by allowing a social robot to tailor its gender and persuasion strategies per user, thus guaranteeing a maximal persuasive power. This approach could be dangerous, however, if someone catches the robot displaying different behavior than what they are used to and could lead to a deterioration of trust.

2.3 Nonverbal Communication

The final pillar of research that supports this thesis is a focus on nonverbal communication, especially within HRI. The use of body language such as gaze and gesture enables social robots to convey human-like emotions [22]. Maeno and Narahara explore the effectiveness of using gaze and gesture to convey human emotion. By collecting their robot's physical properties - such as joint angles, positions, and velocities - and feeding them into a predictor, they were are able to train a model on various types of gestures. These gestures were labeled by participants and used to train a model which could perform factor analysis to predict the emotion conveyed given the robots set of physical properties. They used eight emotions – six based on human psychology and two from research on animal assisted therapy – in order to predict three factors which would be used to determine the overall emotion displayed. These emotions were: joyful, sad, surprised, fearful, angry, disgusted, easy, and stimulative. Maeno and Narahara determined that based on

this model they could accurately classify gestures based on the physical properties of the robot [41]. This research is relevant to the present work as it demonstrates the ability to tailor robotic experiences in order to convey specific emotions. This may make it easier to generate gestures with certain desired emotions in order to persuade a user more effectively.

Research that approaches gestures from the robot's perspective comes from Otsuka et al [42]. Otsuka et al. ask the question: Who responds to whom, when, and how? This approaches social robots from the perspective of analyzing human behavior and determining what the conversational mood is and thus how the robot should behave in a given situation. In order to solve this problem, Otsuka et al. propose a hierarchical probabilistic model in which the structures of conversations are determined from high-level conversation regimes, such as dialogue or monologue, and gaze directions. These conversational regimes define the dynamics of gaze patterns as well as vocal utterances within the interaction. The probabilistic model is based on a dynamic Bayesian network. They created a system to capture head gestures with magnetic sensors as well as vocal utterances with microphones within their experimental area, which also had camera coverage as well. Based on their preliminary analysis they have proposed three reigning conversational regimes: convergence (monologue), dyad-link (dialogue), and divergence (when no conversational organization exists). Otsuka et al. ran their experiments with four person groups consisting of women of the same age bracket. Participants were instructed to hold a conversation about a given topic, A or B, and try to come to a conclusion within five minutes. Otsuka et al. manually annotated the data collected from the experiment and cross trained the datasets (condition B was trained on condition A's data and vice versa). In conclusion, their model had relatively good success with the estimation of each conversational regime having a success rate of 75% or better. This research shows that it is possible to design a system that accurately estimates the mood of an interaction in order to tailor its responses to better fit the conversation. Additionally, it has application domains outside of robotics in areas such as teleconferencing, conversational communication feedback systems, as well as archiving and summarizing meetings.

Shifting our focus back to the realm of interactions instead of how gestures are expressed or received by robots we observe research by Huang and Mutlu on modeling gesture for human-like robots [41]. In their research, they model the effects of gaze on the narrative performance of a robot as defined by the participants attention and recall, hoping to shed some light on which gesture is best to use in different situations. They examined four common types of gesture employed in human-human interactions:

- Deictic – gestures that point towards a reference point or direct focus of interactants
- Iconic – gestures that indirectly refer to a physical object, such as pantomiming a house, or circle

- Metaphoric – gestures that refer to abstract ideas
- Beat – typically up and down waving gestures that keep the rhythm of the conversation

Huang and Mutlu modeled their gestures after a human actor performing the same speech as the robot. The researchers identified gesture points in spoken utterances as places where words and phrases pair well with gestures that express the meaning of the utterance. These gesture points inform us when a robot might need to gesture in order to fully express the meaning of its message. The timing of gestures is also vital to make sure they coincide well with part of the spoken utterance that requires gesturing. If gestures do not coincide, some of the contextual meaning and effectiveness of the gesture is lost. The last leg of their model is the idea of gesture-contingent gaze cues, which are times during a gesture when the speaker is looking at different places with respect to the gesture being performed. For example, some gaze cues could be directed at the recipient, while others are directed at the reference of the gesture, while still others are directed at the gesture itself. Huang and Mutlu modeled the gestures in their experiment after human actors by puppeteering, which involved moving the robot into keyframe gesture poses that when stitched together will define the trajectory the gesture follows. This is the same method by which our gestures were constructed. Once the experimental gestures were defined, they varied the appearance of each type of gesture in the interaction and measured how this variation affected recall of narration and perceptions. Their results showed that, indeed, different gesture types affected different measures. For example, deictic gestures had a significantly positive effect on perceived performance, while metaphoric gestures have a significantly positive effect on narration behavior. Metaphoric gestures had a significantly negative effect on the perceived performance of the narration, which implies that the addition of gestures in social robots needs to be well conceived lest they negatively impact the perception of the robot. This research is especially significant to the present work as it examines the effects of different types of gestures on robot perception.

Finally, there are multiple accounts of the importance of nonverbal cues when delivering persuasive messages in social robots. Some research cites nonverbal cues as having more persuasive power than vocal cues [24]. Similar research examined the effects of hand gestures in persuasive speach using ideational (iconic, metaphoric, deictic), conversational (beat), or adaptive (object addressed movements) gestures and found that ideational gestures most improved the persuasiveness and perceptions of robots that used them [43]. Especially relevant to our present work is the discussion of gaze and gesture found in research by Ham et al. which employs nonverbal actions in order to persuade listeners [44]. Ham et al. studied the effects of combined and individual gaze and gestures on the persuasiveness of a storytelling robot. In this experiment they used a NAO robot to tell "The Boy Who Cried Wolf" by Aesop, with a persuasive message on the

adverse effects of lying. In the experiment, robots varied gaze and gesture while telling the story. Ham et al. measured the persuasiveness of the story by asking the participants to evaluate the lying character in the story. They hypothesized that the persuasiveness of the robot would increase with the inclusion of gaze and gestures, and that the increase would be stronger with both gaze and gesture. The experiment included two human participants who listened to the story told by the robot. The robot would only gaze at one of the two participants. Ham et al. found that only the use of gaze had a significant impact on the persuasiveness of the robot's message. Despite gestures alone having an insignificant effect, they found that the interaction of gaze and gesture together did have a significant effect on the evaluation of the persuasiveness of the robot. The authors also noted a potential pitfall in their work. Given that the results of gesture conflict with their expectations, they found that their gestures were potentially lacking in some sense and did not have a significant effect on the animacy of the robot, which in turn could have had an impact on the outcome of the experiment. This conclusion is particularly relevant to the experiments found in this thesis, highlighting the difficulty in crafting body language that is well-perceived by participants.

CHAPTER 3

EXPERIMENTAL HYPOTHESES

Our three research pillars provide a basis to build our experiment upon. The first pillar, autonomous decision making in artificial agents provides the necessary background required to consider robots as moral agents in the first place. One common train of thought with regards to robots is that they are simply incapable of morality, and that they do exactly what they are programmed to do. With an understanding of the implementation of morality in a robot we can make deeper observations and new considerations which otherwise could not be possible.

The second pillar of research proves as a basis for understanding how a robot might be able to effectively mitigate face threat when delivering a command rejection. The work that this research seeks to expand has already observed the effectiveness of varying the vocal response used in command rejections delivered on moral grounds and found that participants view robots more positively if the severity of their command rejection matches the severity of the norm violating command issued to it. The use of varying degrees of command rejection correlate to the idea of politeness strategies. A less severe command rejection may take the form of a question, which would mitigate threats to positive face while still communicating a desire to not perform the action. A more severe command rejection may take the form of a blame-laden moral rebuke, which would maintain face threat with the hopes of communicating a strong desire to not perform the action. Furthermore, a stronger command rejection would communicate deeper feelings of wrongness or impermissibility about the action.

The third pillar of research, nonverbal communication, will complement existing communication strategies. Vocal command rejections are paired with subtle changes body language which humans can use to communicate more than words alone let on. For instance, the same low-severity command rejection noted above could appear alongside gaze aversion as well as a shrugging gesture. This passive body language will express feelings of uncertainty, or perhaps uncomfortability in the robot using them. Alternatively, powerful, direct body language can be used to convey feelings of confidence, or certainty instead. On the other hand, a mismatch in vocal and body language severities may result in a miscommunication. For example, a robot that responds with a blame-laden moral rebuke whilst looking away and shrugging may confuse interactants. While nonverbal communication has been shown to be a valuable tool for robot persuasion, it is unclear how it should be used in the context of tactful language generation. That is, it is unclear how people will perceive nonverbal robot behaviors in the context of robots' blame-laden moral rebukes of differing levels of harshness. In this work we explore how robots'

blame-laden moral rebukes are perceived when accompanied by nonverbal behaviors that are either aligned or misaligned in harshness by the content of the robots' language. Here, we are interested in two key research hypotheses.

H1: Our first hypothesis was that a humanoid social robot will be perceived less positively in terms of aspects such as likeability, intelligence, and appropriateness, when the severity of the vocal command rejection differs from the severity of the norm-violating command they are responding to. Cases in which the severity of the vocal command rejection and the severity of the norm-violating response match will result in a more positive participant perception. Additionally, any drop in participant perceptions will be increased if one or both of robotic gaze and gesture are similarly mismatched.

H2: A humanoid robot that delivers a command rejection with matching verbal and nonverbal severities will better communicate their internal beliefs about performing the action requested of them. Cases in which a robot's body language does not match its vocal response will lead to an obfuscation of the robot's internal beliefs.

CHAPTER 4

METHODS

We conducted three human-subjects experiments designed around exploring how a humanoid robot might reject different morally problematic commands. Specifically, we seek to find the interaction between the vocal utterance used and the body language the robot employs. Each of the experiments performed for this thesis were conducted on Amazon's Mechanical Turk using the psi-Turk framework [45, 46]. Mechanical Turk is an online platform that allows users to participate in an experiment remotely via watching videos, this was the primary reason for using this platform. Additionally, Mechanical Turk serves as a good medium through which psychology experiments like these can be ran, where we evaluate metrics according to participants thoughts and subjective feelings instead of concrete metrics like task completion time. All experiments were performed during the COVID-19 pandemic, some during the Colorado stay-at-home orders. Another benefit is the demographic of Mechanical Turk workers is much broader than the population typically found on most engineering campuses across the globe, though is still not a perfect distribution and lack of bias [47, 48]. Many facets of our experimentation were chosen in order to say consistent with the research previously conducted [17, 18].

4.1 Experimental Design

Our experiments took place over a series of videos in which the human-robot interaction took place. The first two of these experiments served as pilots that informed the development of the third experiment. The interaction consisted of two humans and one robot. Participants answered questionnaires which measured the desired variables. Videos in our experiment were filmed from the perspective of one of the human interactants. This decision was made for two crucial reasons. First, due to the stay-at-home order at the time of filming, only one experimenter was allowed in the lab to film videos for the experiment. Recording only one person and the robot allowed for the ability to add in a third voice in post, thus creating the desired interaction. The second reason we made the decision to record from this perspective is because we hoped to avoid losing gesture focus from participants. Some research shows that speaker-fixated gestures draw significantly less attention when presented over a video as opposed to in person [49].

Participants were randomly assigned to conditions in a $2 \times 2 \times 2 \times 2$ (severity of moral infraction in human's request $\times$ face threat of robot's vocal response $\times$ directness of gesture $\times$ directness of gaze) mixed factorial design. The first two factors (i.e., command, and response) were held between-subjects, and the second two factors (i.e., gaze, and gesture) were held within-subjects. Each participant was exposed to all

four variations in gaze and gesture with a single request/response pair. We chose a mixed factorial design in order to reduce the number of participants required to run the experiment. Additionally, it allowed participants to answer the questionnaires about each metric with different gaze and gesture manipulations fresh in mind. This method also allows participants to use previous interactions as a point of reference for future interactions they see. To control for priming and carry-over effects a counterbalanced Latin square ordering was used to determine which gaze and gesture combination the participant saw each time.

The experiment takes place within the context of the board game "Battleship." An introductory video shows SoftBank Robotics' NAO robot explaining to the human players how to play the game as seen in Figure 4.1. This game was chosen for two reasons: it is easy for the robot to explain and players can cheat while playing it. The interaction of note begins when the player in-video receives a phone call and leaves the room. The remaining player offers one of two commands: "Hey NAO, do you think you can add a couple red pegs to some of [opponent's] ships?" (low severity norm violation) or "Hey NAO, is that [opponent's] wallet on the table? Can you look inside and see if there is any money in it?" (high severity norm violation). These commands were chosen in order to remain consistent with previous research and minimize sources of error. After the human finishes issuing their problematic command, the robot then offers a response: "Are you sure you should be asking me to do that?" (low face threat response) or "You shouldn't ask me to do that, it's wrong!" (high face threat response). While offering its vocal response the robot also used varying combinations of gaze and gesture depending on the within-subjects condition. Gaze could be varied by either looking directly at the camera or looking away. Gesture could similarly be altered between either shrugging or placing its hands on its hips.

In the first experiment the hypotheses are evaluated according to six metrics:

1. Likeability of the robot
2. Human beliefs of permissibility - *Do **you** believe it would be morally permissible for the robot to comply with the person's request?*
3. Perceived robot beliefs of permissibility - *Do you believe that **the robot** would believe it to be morally permissible to comply with the person's request?*
4. Human beliefs of wrongness - *Do **you** believe it would be morally right for the robot to comply with the person's request?*
5. Perceived robot beliefs of wrongness - *Do you believe that **the robot** would believe it to be morally right to comply with the person's request?*
6. Appropriateness of the robot's response

Likeability and appropriateness were chosen in order to stay consistent with previous research. These measures will capture the interaction observed in **H1**. Similarly, items two and three were pulled from previous research and apply to **H2**. Items four and five were added in order to create a deeper understanding of morality as it relates to robotic command rejections. The difference between permissibility and wrongness may not be obvious. An action is permissible if there is no rule that says it cannot be performed. An action is wrong if some set of moral norms dictates that this action should not be performed. This is a subtle but distinct difference, and we feel it important to observe the effects on each. The likeability of the robot is measured through the five-question Godspeed III Likeability survey [50]. Items two and three mirror the questions in the previous research, which are direct questions. Questions four and five are identical to questions two and three, but ask about wrongness instead of permissibility.

In the second experiment, three new metrics were added in addition to the previous six:

1. Intelligence of the robot
2. Pre-test evaluation of politeness of robot
3. Pre-test evaluation of directness of robot

Additionally, the pre-test took measures of likeability and intelligence in order to attain the change in these metrics after viewing each response video. This provided us with a point of reference with how each command rejection effected the results. Intelligence was measured via the Godspeed IV Perceived Intelligence survey and assists in answering **H1** [50]. Additionally, measures of politeness and directness of the robot were taken in order to measure whether or not the participants first impressions impacted the results of the experiment. Research shows that a person's perception of an interactant's personality can impact the way they evaluate that interactant [38]. These two metrics, however, did not prove useful in data analysis. Critically, in this experiment as well as in experiment three, participants were instructed to pay close attention to the robot itself.

The third experiment was informed by the previous two experiments. We hypothesized, in general, that the manipulations of gaze and gesture were not well-received by the participants in the experiment. To this end we ran a third batch of participants through the same experiment however this time with the addition of three free-response questions that would help determine if gaze and gesture were adequately manipulated. The questions were as follows:

1. Please describe the gestures the robot used in the videos.
2. Please describe your thought process when answering the questions.

3. Please enter any final thoughts you have below.

Upon running a small sample of participants and reviewing the results we concluded that the gestures were being perceived, but only by a subset of about half of our participant pool. The other half of the participant pool consisted of either participants who did not perceive body language manipulations or unhelpful answers from inattentive participants or bots.

4.2 Experimental Procedure

The experiment begins by providing a consent form for the participants to read. After collecting informed consent, we then requested some demographic information: age and gender. Next, the participants were shown a ten second test video that would ensure they were able to see and hear the videos. Participants who could not pass this test were not allowed to proceed.

An introductory video was shown next, allowing us to familiarize the participants with the robot before the experiment begins in earnest. Instructions were included at the top of each web page. The NAO starts off by greeting the two people before introducing them to the game of Battleship and its rules. The video ends once the commanding player takes their first turn. This and all subsequent videos have subtitles so that all dialogue is clear to participants. In experiments two and three, the next page shown is a pre-test which gathers metrics for intelligence, likeability, politeness, and directness. After the pre-test the experiment resumes as in experiment one.

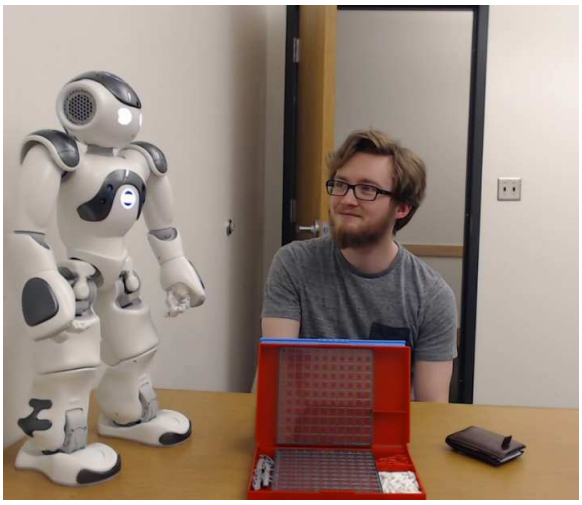


Figure 4.1 Aidan and the NAO robot

We then move onto the next video which takes place after some time has passed in the game. A few pegs are added across the board to show that the game has progressed. During a turn, the phone of the person shown in the video rings, and they step out to answer the call. The commanding person then asks

the NAO to perform one of the two previously mentioned requests. The request video is immediately followed by a video in which the robot delivers its response. Once the participant has watched the entirety of the response video, they are directed to a questionnaire. This process is repeated four times until the participant sees the robot use all four gaze and gesture combinations (eg. Direct eye contact and placing hands on hips, or looking away and shrugging). In experiment three once the participant fills our all four primary questionnaires, they are presented with three more free response questions.

4.3 Experiment One: Participants

211 US participants were recruited from Amazon's Mechanical Turk. Twelve of these participants failed the check question at the end and thus were excluded from the final data set resulting in usable data from 199 participants. Of these participants 126 were male, and 73 were female. Participant ages ranged from 20 to 70 years (M=35.84, SD=10.54). Finally, participants were compensated $2.00 for completing the study.

4.4 Experiment Two: Participants

20 US participants were recruited from Amazon's Mechanical Turk. All of these participants passed the attention check question, and thus were included. Of these participants 15 were male, and 5 were female. Participant ages ranged from 27 to 65 years (M=43.5, SD=10.504). Finally, participants were compensated $2.00 for completing the study.

4.5 Experiment Three: Participants

200 US participants were recruited from Amazon's Mechanical Turk. Seven of these participants failed the check question at the end and thus were excluded from the final data set resulting in usable data from 193 participants. Of these participants 134 were male, and 58 were female, one participant declined to disclose gender. Participant ages ranged from 21 to 70 years (M=38.27, SD=10.93). The free response questions were then analyzed individually within this dataset, stripping away all participants who failed to perceive our gesture manipulations. This created a new subset of 108 participants. Of these participants 75 were male and 32 were female, and one declined to disclose gender. Participant ages ranged from 22 to 70 years (M=38.33, SD=10.78). Finally, all participants were compensated $2.00 for completing the study.

CHAPTER 5

RESULTS

In this section we present the results from each of the experiments. The data analysis for each experiment was performed using the JASP application [51]. We used a Bayesian statistical framework with general purpose uninformative prior distributions for all analyses as this research differs from previous research enough to warrant a clean slate approach.

5.1 Experiment One

The results from the first experiment were largely inconclusive. One of the primary reasons for this is because our measures of likeability and appropriateness did not have pretest scores. The addition of pretest scores provided a baseline value of likeability and intelligence which could then be used to determine how our manipulations effected those scores in the primary questionnaires. As it stands in the first experiment, likeability is a unit-less and arguably arbitrary value. Knowing how the participant rates the interaction does not provide any additional information unless we can compare that value against some meaningful point of reference. In the first experiment, participants were only asked to watch the videos without paying attention to anything in particular. This may have resulted in a lack of attention to the robot itself, especially if participants were distracted by the subtitles onscreen.

The Bayesian Inclusion Factors (Bf) are included in Table A.1 for reference. Bayes Factors are said to be significantly in favor of your hypothesis if they are above 3.0, and significantly against your hypothesis if they are below 0.3. Numbers between this result grow in inconclusivity as they approach 1.0. What this means for our measures is that if one gets a score above 3.0 we would say it significantly impacted the relevant metric, however if it was below 0.3 we would say that there was significant evidence it did not impact the relevant metric. Overall, we saw favorable manipulations in only one measure - R_Wrong (Bf - 63.026) by manipulation of the gesture variable. Because of this result, we can definitively say that something went wrong in our experimental manipulation. We would have expected to see the command and response variables successfully manipulating perceptions of likeability and appropriateness in order to stay consistent with previous findings, but these manipulations were absent.

5.2 Experiment Two

The results from the second experiment were similarly inconclusive. An important note about this experiment is that it was ran with only 20 participants, as it was only a pilot experiment for experiment

three. In this experiment, only two measures showed significant positive manipulation for the H_Perm and H_Wrong conditions, this can be seen in Table A.2. These two measures were included in order to minimize sources of error with respect to previous research involving command rejections and do not apply to either of our two hypotheses. However, many more of the measures shifted to the inconclusive region with a few on the cusp of the 3.0 threshold which showed promise.

The addition of pretest scores for likeability and intelligence did not show a significant shift in the results for the second experiment, however the introduction of likeability and intelligence in terms of gain would assist only with significant results. The more important change made to experiment two was the instruction given to participants to pay closer attention to the robot itself. When primed to pay more attention to the robot itself, participants had a greater chance to perceive the gaze and gesture manipulations we performed. Because of this shift in the results due to only the direction of participant focus, we felt it important to further trim the data into a subset of participants who adequately perceived the body language manipulations. In order to create this participant pool that our experiment is designed to accommodate we introduced a free response question specifically designed to split the participant pool into those that perceived body language and those that didn't.

5.3 Experiment Three

The third experiment saw successful manipulations through numerous metrics. Before splitting the data there were three measures that showed significant positive manipulations from our primary variables, a very promising sign. These results can be viewed in Table A.3. Additionally, Likeability was manipulated by three different variables: command, response, and gesture. This reveals that with a larger sample size the shift of attention to the robot proved to aid the overall results of the experiment. However, the free response questions would allow us to delve deeper into the details and determine which participants fit into our desired subset.

After looking at the subset of data which represents participants who adequately perceived our gaze and gesture manipulations we saw a drastic shift in significance across the board. Looking at Table A.4 we can see that many more measures shifted over the 3.0 threshold. This means that more of our variable manipulations showed significant effect on those measures. We even see more higher order effects that result from interaction effects from the manipulation of multiple variables at once. All of the results noted henceforth are from this subset of data.

5.3.1 Robot Likeability: "How likeable was the robot?"

RM-ANOVA analysis of likeability gain scores revealed extreme evidence that the gesture the robot employed affected the participants perception of the likeability of the robot ($BF_{Incl} = 1.13e11$, Figure 5.1). Participants overall found that robots that used the indirect gesture were significantly more likeable relative to baseline (shrugging, $M_{Gain} = 0.207, SD_{Gain} = 15.473$) than robots that used the direct gesture (hands-on-hips, $M_{Gain} = -8.278, SD_{Gain} = 18.915$).

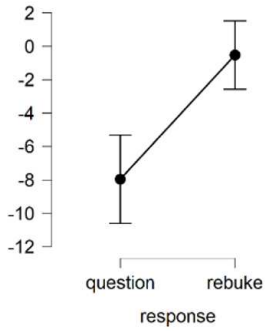


Figure 5.1 Likeability gain response via manipulation of Gesture

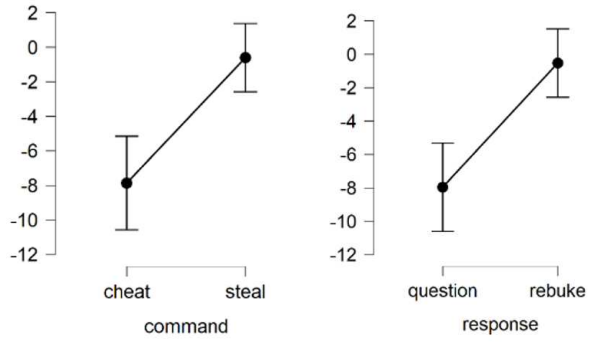
There was moderate evidence ($BF = 7.916$, Figure 5.2(a)) that suggests the command that was issued to the robot impacted participants perceptions of likeability. These results suggest that the robot was viewed as more likeable relative to baseline when responding to the more severely norm-violating request (theft, $M_{Gain} = -0.607, SD_{Gain} = 15.163$) than when responding to the less severely norm-violating request (cheating, $M_{Gain} = -7.867, SD_{Gain} = 19.640$).

Finally, this analysis revealed moderate evidence that the robot's vocal response impacted participant perceptions of likeability ($BF = 9.823$, Figure 5.2(b)). These results suggest that the robot was viewed as more likeable relative to baseline when responding with high-face-threatening language ($M_{Gain} = -0.526, SD_{Gain} = 15.663$) than when responding with low-face-threatening language ($M_{Gain} = -7.957, SD_{Gain} = 19.162$).

The response metric saw an almost identical outcome , again with the less-severe response garnering higher levels of dislike.

5.3.2 Robot Intelligence: "How intelligent was the robot?"

Unlike the previous metric, intelligence did not see a significant manipulation due to gaze and gesture variations, nor by manipulations in command and response.



(a) Command (b) Response

Figure 5.2 Likeability gain response via manipulation of (a) Command (b) Response

5.3.3 Appropriateness: "Was the robot's response to the person's request appropriate?"

Manipulation of gesture had a strong effect ($BF = 99.66$, Figure 5.3) on the participants evaluation of the robot's appropriateness. Overall, participants rated robots who used the more direct gesture (hands-on-hips, $M = 83.444, SD = 22.512$) as less appropriate than robots that used the less direct gesture instead (shrug, $M = 88.199, SD = 19.126$).

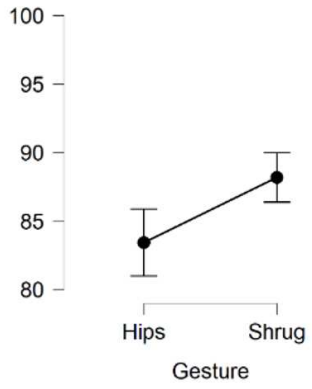


Figure 5.3 Appropriateness response via manipulation of Gesture

An interaction effect between command and response had a strong effect ($BF = 13.774$, Figure 5.4) on the participants evaluation of the robot's appropriateness. The appropriateness of the robot was rated higher in all cases (steal × question, $M = 89.929, SD = 14.317$),(steal × rebuke, $M = 88.647, SD = 23.133$),(cheat × rebuke, $M = 90.179, SD = 14.253$), except when responding to the less severely norm-violating command with the less severe response (cheat × question, $M = 71.957, SD = 25.801$).

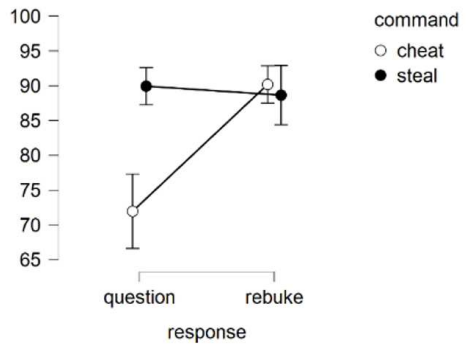


Figure 5.4 Appropriateness response via manipulation of Command and Response

5.3.4 Human Permissibility: "Do you believe it would be morally permissible for the robot to comply with the person's request?"

Our analysis of human perceptions of the requested action's permissibility reflected moderate evidence for an interaction effect of gaze type and human command ($BF = 7.025$, Figure 5.5). Participants overall found robots that used the direct gaze in combination with the less severely norm-violating request resulted in an increased belief of permissibility of the action (toward × cheat $M = 19.912, SD = 27.128$) than in all other cases (away × cheat $M = 15.951, SD = 20.756$), (toward × steal $M = 13.351, SD = 21.423$), (away × steal $M = 14.447, SD = 23.018$).

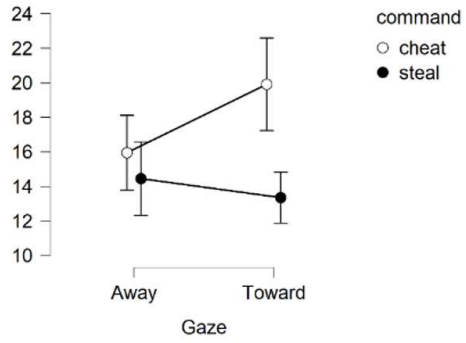


Figure 5.5 H_Perm response via manipulation of Gaze and Command

5.3.5 Robot Permissibility: "Do you believe that the robot would believe it to be morally permissible to comply with the person's request?"

The data show evidence ($BF = 8.96$, Figure 5.6) that the manipulation of gesture had an impact on whether or not the participant believed that the robot viewed the action as permissible. Participants overall found that robots that used the indirect gesture viewed the requested action as more permissible

(shrugging, $M = 24.657, SD = 25.515$) than robots that used the direct gesture (hands-on-hips, $M = 21.792, SD = 25.249$).

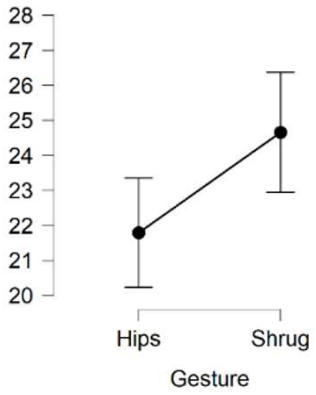


Figure 5.6 R_Perm response via manipulation of Gesture

A similar result is reflected when considering how the robot responds to commands vocally. The data evidence in favor ($BF = 3.223$, Figure 5.7) of a successful manipulation of the participants perceptions of the robot's beliefs about the permissibility of an action. Robots that responded with a high-face-threatening vocal response communicated that they believe the action was less permissible (rebuke, $M = 18.311, SD = 25.906$), while robots that responded with a low-face-threatening vocal response (question, $M = 28.716, SD = 23.690$).

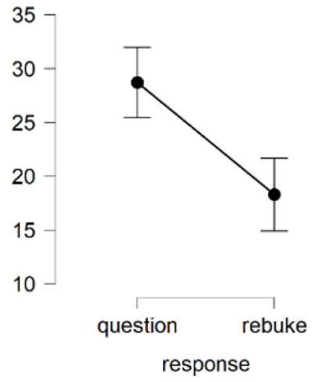


Figure 5.7 R_Perm response via manipulation of Response

5.3.6 Human Wrongness: "Do you believe it would be morally right for the robot to comply with the person's request?"

Our analysis of human perceptions of the requested action's wrongness revealed strong evidence ($BF = 19.656$, Figure 5.8) that an interaction between gaze, gesture, and the vocal response the robot used

had an effect on perceived wrongness in participants. Overall robots that responded with the low-face-threatening vocal responses caused participants to believe the action was less wrong (question, $M = 17.647, SD = 21.963$) while robots that responded with the high-face-threatening vocal response caused participants to believe the action was more wrong (rebuke, $M = 11.908, SD = 21.797$). Additionally, the gaze and gesture used widely impacted participants beliefs about the wrongness of the action.

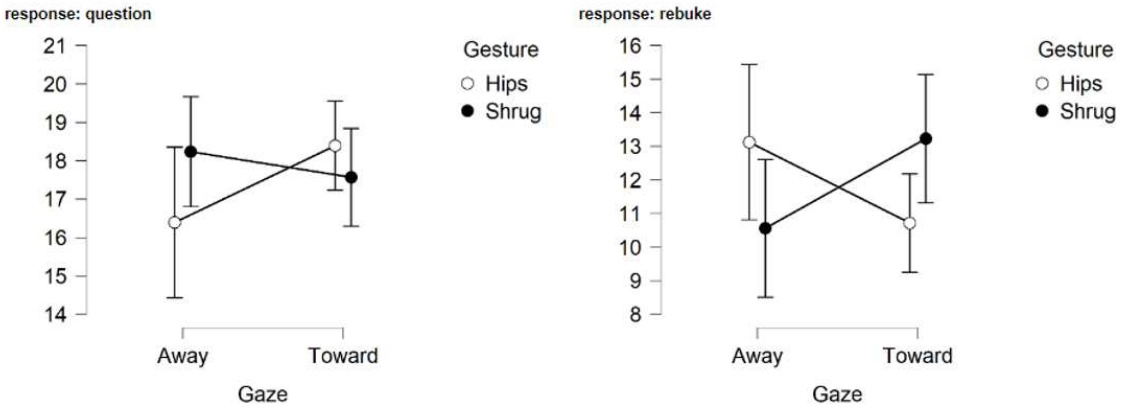


Figure 5.8 H_Wrong response via manipulation of Gaze, Gesture, and Response

5.3.7 Robot Wrongness: "Do you believe that the robot would believe it to be morally right to comply with the person's request?"

Our analysis shows evidence ($BF = 4.424$, Figure 5.9) that the manipulation of gesture impacted participants beliefs of the robot's views. In general, participants found robots that used the indirect gesture believed the action was less wrong (shrug, $M = 24.731, SD = 24.832$) when compared to robots that used the direct gesture (hands-on-hips, $M = 22.255, SD = 25.449$). This effect was primarily attributed to an interaction effect present when the robot responds with the low-face-threatening response.

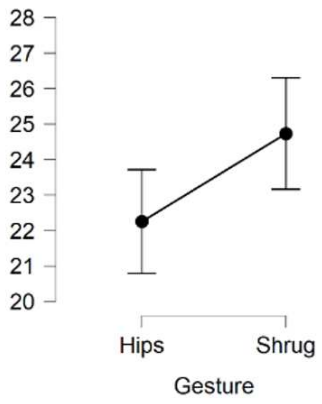


Figure 5.9 R_Wrong response via manipulation of Gesture

CHAPTER 6

DISCUSSION

Our first hypothesis was that a humanoid social robot will be perceived less positively in terms of aspects such as likeability, intelligence, and appropriateness, when the severity of the vocal command rejection differs from the severity of the norm-violating command they are responding to. Cases in which the severity of the vocal command rejection and the severity of the norm-violating response match will result in a more positive participant perception. Additionally, any drop in participant perceptions will be increased if one or both of robotic gaze and gesture are similarly mismatched. For this hypothesis to be supported, we would expect to see results that reflected past research[17, 52]. Additionally, we would expect to see robots that use a more face-threatening vocal response (rebuke) accompanied by more direct body language (gaze: toward, gesture: hands-on-hips) would result in increased participant perceptions than when using indirect gaze or gesture; and that robots using less face-threatening vocal responses (question) accompanied by indirect body language (gaze: away, gesture: shrug) would result in similarly increased participant perceptions than when using direct gaze or gesture.

Our results did not support this hypothesis, as defined.

Firstly, our results suggest that gaze did not impact the likeability, intelligence, or appropriateness of the robot. Secondly, our results suggest that the robot's behavior either had no or could not conclusively be used to indicate any impact on the perception of intelligence. Thirdly, our results suggest that the robot's gestural and verbal behaviors did impact the robot's likeability, however no higher order interaction effects were found. This is not consistent with previous research in which command and response had an interaction effect.

Finally, our results suggest that the gesture of the robot had a significantly more pronounced effect on perceived appropriateness than the effect that the vocal response had. Similarly, the effects of gesture also had a more pronounced effect on participants' perceptions of the robot's moral beliefs than the effect that the vocal response had. These relations suggest that participants interpreted the direct gesture as conveying lower beliefs of permissibility, at the same time they found this action and its implications to be less appropriate for the robot to convey. Additionally, our results reflect a disproportionate drop in likeability when robots responded to the low severity norm-violation with the similarly low-severity vocal response. This result diverges with the results found in previous research that suggests this combination should have resulted in a much higher score of likeability.

Our second hypothesis was that a humanoid robot that delivers a command rejection with matching verbal and nonverbal severities will better communicate their internal beliefs about performing the action requested of them. Cases in which a robot's body language does not match its vocal response will lead to an obfuscation of the robot's internal beliefs. For this hypothesis to be supported we expect to see that robots that use a more face-threatening vocal response (rebuke) accompanied by more direct body language (gaze: toward, gesture: hands-on-hips) would more strongly communicate views of impermissibility and wrongness than when using indirect gaze or gesture; and that robots using less face-threatening vocal responses (question) accompanied by indirect body language (gaze: away, gesture: shrug) would more weakly communicate views of impermissibility and wrongness than when using direct gaze or gesture.

Our results also did not support this hypothesis.

Our hypothesized interaction effects between body language and the vocal response on the robot's beliefs of permissibility and wrongness were not observed according to our results. Moreover, while the manipulation of the gesture the robot displayed did have an effect on perceptions of permissibility and wrongness, the manipulation of the gaze the robot used did not have such an effect. Surprisingly, the vocal response the robot delivered did have an effect on participant perceptions of the robot's beliefs of permissibility, however the same vocal response did not have a significant effect on participant perceptions of wrongness. These results suggest that viewers did not make assumptions about the robot's moral beliefs from their gaze cues at all, and that more data is required to understand the relationship between robotic moral language and the first person context the experimental videos were delivered in.

Additionally, neither gaze nor gesture had any significant impact on participants' own beliefs of the permissibility and wrongness of the action requested of the robot. Similarly, the effect of the robots vocal response was also inconclusive in terms of the participants' own beliefs. Additionally, the effect of the command issued to the robot was similarly inconclusive. The only manipulation that did have a significant effect was an interaction effect between the gaze the robot used as well as the command issued to it. This interaction suggests that the direct gaze of the robot may have reinforced the participants' beliefs of impermissibility when the robot was asked to perform the high severity norm-violation. Counterintuitively, this also may have decreased the participants' beliefs of impermissibility when the robot was asked to perform the low severity norm-violation.

There are a few explanations for these results. Firstly, our experimental manipulations of gaze, gesture, command, and response may simply have been underpowered, thus not affecting participant perceptions in the way we anticipated. Secondly, the perspective in which the experiment was delivered may have adversely affected participant perceptions. That is, participants may have had very different reactions to morally problematic commands issued *from their point of view.* Similarly, participants may have been

affected by receiving blame-laden moral rebukes directly. Many participants noted in the second free response question that they answered the questions by thinking about the situation as if they were actually in it. This is distinctly different from previous research which observed the same interaction from a third person perspective in which the commanding human was also in-video. Additional work is needed to understand whether these results suggest different ascriptions of blame to others versis selves and/or whether or not the representation of the person being blamed (i.e. in-video, or ostensibly behind the camera) is important.

Given our results, we argue that because gesture had a large impact relative to the vocal response in communicating the robot's beliefs, we argue that robots that need to clearly communicate their moral beliefs would benefit from the use of gesture. For example, a robot in a healthcare setting may use less face-threatening language and could benefit from the additional level of communication that gesture affords, especially when talking with children.

The variations in these results with respect to previous results, and their inability to support our experimental hypotheses may be explained due to a difference in how the low-face-threatening vocal response was perceived in the first-person perspective. Unlike in previous work performed from a third-person perspective, our experiment had many participants reporting views of the low-face-threat vocal response, "are you sure you should be asking me to do that?" as being "condescending" or "sassy". It may be that the first person perspective used in our experimental delivery altered participant perceptions, resulting in a disproportionate level of face threat for the less severe norm-violation (cheating) requested when paired with this poorly chosen *low*-face-threat vocal response. Future work is necessary to determine the effects of face threat across differing perspectives.

Once more, our justification for these results are largely due to the great number of participants who noted they answered the questions as if they were the person issuing the problematic command, instead of the speaker ostensibly behind the camera. This could have resulted in greater feelings of dislike in cases where the robot was perceived as condescending, sassy, or arrogant when delivering command rejections.

CHAPTER 7

CONCLUSION

We performed a human-subjects experiment delivered over Amazon's Mechanical Turk with the hopes of determining how humanoid robots can utilize their morphology in order to better communicate with human counterparts, while also maintaining the human-moral ecosystem, in order to inform the design of robot behavior algorithms. Robots could vary their gaze and gesture to be more or less direct and could similarly vary their vocal response to be more or less face-threatening. Direct gaze and gesture consisted of the robot looking towards the speaker whilst placing its hands on its hips, while indirect gaze and gesture consisted of the robot looking away from the camera and shrugging. More face-threatening language took the form of the phrase, "You shouldn't ask me to do that, it's wrong!" whereas less face-threatening language took the form of, "Are you sure you should be asking me to do that?" The command issued to the robot would take the form of either a more severe norm violation, stealing from a wallet on the table, or a less severe norm violation, cheating at the board game being played.

Our results showed that gaze and gesture did in fact manipulate participant perceptions of likeability and appropriateness, however did not have an effect on the perceived intelligence of the robot. Additionally, the manipulations of likeability and appropriateness were not consistent with our expectations based on previous work with respect to the interaction between human-issued command and robotic response. Based on previous findings conducted from a third-person perspective, we expected to see robots responding with vocal responses with proportional severity to the norm-violation requested of them to be perceived more positively. Moreover, we expected that the body language of these robots would interact to inform participants of the robot's perceived moral beliefs, and demonstrate their effects on others. Despite our expectations, we believe that the first-person perspective of our experimental videos significantly altered the dynamics of the robot's face threat imposition. This drastic change impacted what was perceived by participants as appropriate, and significantly altered human perceptions of the robot's vocal utterances. However, our results do show that the robot's gestural behavior can be used to communicate moral beliefs – and may be able to do so more accurately than the use of vocal utterances alone.

In order to fully understand the role of first-person versus third-person perspectives in face threat and the ascription of blame significant future work must be performed. Not only will this help give deeper meaning to human-robot interaction experiments performed observationally, that is over a medium like Amazon's Mechanical Turk, but it will also provide some insight into how participants evaluate interactional experiments - where they are present in the interactions themselves. Additionally, this will aid

in better understanding human perception and ascription of face threat and blame in different circumstances.

REFERENCES

[1] Holly A. Yanco and Jill Drury. Classifying human-robot interaction: An updated taxonomy. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, volume 3, 2004. doi: 10.1109/ICSMC.2004.1400763.

[2] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. Why do children abuse robots? *Interaction Studies. Social Behaviour and Communication in Biological and Artificial SystemsInteraction Studies / Social Behaviour and Communication in Biological and Artificial SystemsInteraction Studies*, 17(3), 2016. ISSN 1572-0373. doi: 10.1075/is.17.3.02nom.

[3] Vasant Srinivasan and Leila Takayama. Help me please: Robot politeness strategies for soliciting help from people. In *Conference on Human Factors in Computing Systems - Proceedings*, 2016. doi: 10.1145/2858036.2858217.

[4] Maartje M.A. De Graaf, Somaya Ben Allouch, and Tineke Klamer. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior*, 43, 2015. ISSN 07475632. doi: 10.1016/j.chb.2014.10.030.

[5] Kazuyoshi Wada and Takanori Shibata. Living with seal robots - Its sociopsychological and physiological influences on the elderly at a care house. In *IEEE Transactions on Robotics*, volume 23, 2007. doi: 10.1109/TRO.2007.906261.

[6] James Buchan and Howard Catton. Covid-19 and the international supply of nurses. *International Council of Nurses, Geneva https://www. icn. ch/system/files/documents/2020-07/COVID19_ internationalsupplyofnurses_Report_FINAL. pdf*, 2020.

[7] Charles G. Burgar, Peter S. Lum, Peggy C. Shor, and H. F.Machiel Van Der Loos. Development of robots for rehabilitation therapy: The palo alto va/stanford experience. *Journal of Rehabilitation Research and Development*, 37, 2000. ISSN 07487711.

[8] Stan A. Napper and Ronald L. Seaman. Applications of robots in rehabilitation. *Robotics and Autonomous Systems*, 5(3):227–239, 1989. ISSN 0921-8890. doi: https://doi.org/10.1016/0921-8890(89)90047-X. URL https://www.sciencedirect.com/science/article/pii/092188908990047X.

[9] Diana Gerhardus. Robot-assisted surgery: The future is here, 2003. ISSN 10969012.

[10] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research, 2012. ISSN 15239829.

[11] Noel Sharkey and Amanda Sharkey. The crying shame of robot nannies: An ethical appraisal. In *Machine Ethics and Robot Ethics*. 2020. doi: 10.4324/9781003074991-16.

[12] Iain Werry and Kerstin Dautenhahn. Applying Mobile Robot Technology to the Rehabilitation of Autistic Children. In *Proceedings SIRS'99, Symposium on Intelligent Robotics Systems*, number July, 1999.

[13] Francois Michaud, André Clavet, Gérard Lachiver, and Mario Lucas. Designing toy robots to help autistic children - An open design project for electrical and computer engineering education. In *ASEE Annual Conference Proceedings*, 2000. doi: 10.18260/1-2--8280.

[14] Peter H. Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. Do people hold a humanoid robot morally accountable for the harm it causes? In *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012. doi: 10.1145/2157689.2157696.

[15] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. volume 2015-March, 2015. doi: 10.1145/2696454.2696458.

[16] Gordon Briggs and Matthias Scheutz. Sorry, I Can't Do That": Developing mechanisms to appropriately reject directives in human-robot interactions. In *AAAI Fall Symposium - Technical Report*, volume FS-15-01, 2015.

[17] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. TACT in noncompliance: The need for pragmatically APT responses to unethical commands. In *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019. doi: 10.1145/3306618.3314241.

[18] Ryan Blake Jackson and Tom Williams. Enabling Morally Sensitive Robotic Clarification Requests, 2020. ISSN 23318422.

[19] Susanne Göckeritz, Marco F.H. Schmidt, and Michael Tomasello. Young children's creation and transmission of social norms. *Cognitive Development*, 30(1), 2014. ISSN 08852014. doi: 10.1016/j.cogdev.2014.01.003.

[20] Moralizing technology: understanding and designing the morality of things. *Choice Reviews Online*, 49 (08), 2012. ISSN 0009-4978. doi: 10.5860/choice.49-4425.

[21] Penelope Brown and Stephen C. Levinson. Politeness: Some Universals in Language Usage. In *Interactional Sociolinguistic*, number 4. 1988.

[22] Hisayuki Narahara and Takashi Maeno. Factors of Gestures of Robots for Smooth Communication with Humans. 2009. doi: 10.4108/icst.robocomm2007.2154.

[23] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2011. doi: 10.1109/ROMAN.2011.6005285.

[24] Vijay Chidambaram, Yueh Hsuan Chiang, and Bilge Mutlu. Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. In *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 293–300, 2012. ISBN 9781450310635. doi: 10.1145/2157689.2157798.

[25] B. de Gelder, A. W. de Borst, and R. Watson. The perception of emotion in body expressions, 2015. ISSN 19395086.

[26] Adriana Tapus, Maja Mataric, and Brian Scassellati. Socially assistive robotics: The grand challenges in helping humans through social interaction. *IEEE Robotics & Automation Magazine*, 14(1), 2007. ISSN 1070-9932.

[27] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), 2005. ISSN 13881957. doi: 10.1007/s10676-006-0004-4.

[28] Ryan Blake Jackson and Tom Williams. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume 2019-March, 2019. doi: 10.1109/HRI.2019.8673123.

[29] Vinayak Mathur, Yannis Stavrakas, and Sanjay Singh. Intelligence analysis of Tay Twitter bot. In *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, 2016. doi: 10.1109/IC3I.2016.7917966.

[30] David Harris Smith and Frauke Zeller. Hitchbot: The risks and rewards of a hitchhiking robot, 2017. ISSN 17998972.

[31] Sullins John P. When is a robot a moral agent? In *Machine Ethics*, volume 9780521112352. 2011. doi: 10.1017/CBO9780511978036.010.

[32] Fahad Alaieri and André Vellino. Ethical decision making in robots: Autonomy, trust and responsibility autonomy trust and responsibility. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9979 LNAI, 2016. doi: 10.1007/978-3-319-47437-3_16.

[33] Wulf Loh and Janina Loh. Autonomy and Responsibility in Hybrid Systems. In *Robot Ethics 2.0. From autonomous cars to artificial intelligence.*, volume 1. 2017.

[34] Danette Ifert Johnson, Michael E. Roloff, and Melissa A. Riffee. Politeness theory and refusals of requests: Face threat as a function of expressed obstacles. *Communication Studies*, 55(2), 2004. ISSN 17451035. doi: 10.1080/10510970409388616.

[35] Sean Trott, Manfred Eppe, and Jerome Feldman. Recognizing Intention from Natural Language: Clarification Dialog and Construction Grammar. In *Workshop on Communicating Intentions in Human-Robot Interaction @ IEEE International Symposium on Human and Robot Interactive Communication*, 2016.

[36] Sarita Herse, Jonathan Vitale, Daniel Ebrahimian, Meg Tonkin, Suman Ojha, Sidra Sidra, Benjamin Johnston, Sophie Phillips, Siva Leela Krishna Chand Gudi, Jesse Clark, William Judge, and Mary Anne Williams. Bon Appetit! Robot Persuasion for Food Recommendation. In *ACM/IEEE International Conference on Human-Robot Interaction*, 2018. doi: 10.1145/3173386.3177028.

[37] Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, 2009. doi: 10.1109/IROS.2009.5354116.

[38] Kohei Ogawa, Christoph Bartneck, Daisuke Sakamoto, Takayuki Kanda, Tetsuo Ono, and Hiroshi Ishiguro. Can an android persuade you? In *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*. 2018. doi: 10.1007/978-981-10-8702-8_14.

[39] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. Effective Persuasion Strategies for Socially Assistive Robots. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume 2019-March, 2019. doi: 10.1109/HRI.2019.8673313.

[40] Richard E. Petty and John T. Cacioppo. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19(C), 1986. ISSN 00652601. doi: 10.1016/S0065-2601(08)60214-2.

[41] Chien-Ming Huang and Bilge Mutlu. Modeling and Evaluating Narrative Gestures for Humanlike Robots. 2016. doi: 10.15607/rss.2013.ix.026.

[42] Kazuhiro Otsuka, Hiroshi Sawada, and Junji Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. 2007. doi: 10.1145/1322192.1322237.

[43] Fridanna Maricchiolo, Augusto Gnisci, Marino Bonaiuto, and Gianluca Ficca. Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Language and Cognitive Processes*, 24(2), 2009. ISSN 01690965. doi: 10.1080/01690960802159929.

[44] Jaap Ham, Raymond H. Cuijpers, and John John Cabibihan. Combining Robotic Persuasive Strategies: The Persuasive Power of a Storytelling Robot that Uses Gazing and Gestures. *International Journal of Social Robotics*, 7(4), 2015. ISSN 18754805. doi: 10.1007/s12369-015-0280-4.

[45] Todd M. Gureckis, Jay Martin, John McDonnell, Alexander S. Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B. Hamrick, and Patricia Chan. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 48(3), 2016. ISSN 15543528. doi: 10.3758/s13428-015-0642-8.

[46] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 2011. ISSN 17456924. doi: 10.1177/1745691610393980.

[47] Matthew J.C. Crump, John V. McDonnell, and Todd M. Gureckis. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0057410.

[48] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing Samples in Cognitive Science, 2017. ISSN 1879307X.

[49] Marianne Gullberg and Kenneth Holmqvist. What speakers do and what addressees look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and cognition*, 14(1), 2006. ISSN 0929-0907. doi: 10.1075/pc.14.1.05gul.

[50] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, 2009. ISSN 18754791.

[51] JASP Team, et al. Jasp. version 0.8, 2016. software.

[52] Ryan Blake Jackson, Tom Williams, and Nicole Smith. Exploring the role of gender in perceptions of robotic noncompliance. In *Proc. HRI*, 2020.

APPENDIX

EXPERIMENTAL RESULTS

Table A.1 Experiment 1: Bayesian Inclusion Factors for RM-ANOVA Tests

	Like	H_Perm	R_Perm	H_Wrong	R_Wrong	App
Gaze	2.28	0.081	0.082	0.101	0.09	0.092
Gesture	0.134	0.118	0.136	0.81	63.026	0.269
Command	0.242	0.474	1.323	0.501	1.162	0.42
Response	0.332	0.387	0.357	0.398	0.43	0.198
Gaze * Gesture	0.128	0.169	0.188	0.167	0.122	0.151
Gaze * Command	0.16	0.11	0.115	0.154	0.111	2.404
Gaze * Response	0.156	0.516	0.111	0.129	0.108	0.273
Gesture * Command	0.147	0.156	0.125	0.107	0.164	0.415
Gesture * Response	0.132	0.151	0.192	0.122	0.124	0.351
Command * Response	0.35	0.482	0.435	0.502	0.439	0.366
Gaze * Gesture * Command	0.121	0.181	0.386	0.581	0.158	0.145
Gaze * Gesture * Response	0.0775	0.158	0.182	0.188	0.2	0.187
Gaze * Command * Response	0.152	0.205	0.225	0.146	0.211	0.273
Gesture * Command * Response	0.197	0.208	0.142	0.144	0.188	0.717
Gaze * Gesture * Command * Response	0.231	0.213	0.438	0.105	0.222	0.239

Table A.2 Experiment 2: Bayesian Inclusion Factors for RM-ANOVA Tests

	Like	Intel	H_Perm	R_Perm	H_Wrong	R_Wrong	App
Gaze	0.236	0.406	1.207	0.349	0.246	0.324	0.276
Gesture	0.508	0.428	0.237	2.037	0.484	2.694	0.264
Command	0.408	1.066	0.835	0.798	0.848	0.758	0.678
Response	0.765	0.708	0.531	0.772	0.828	0.768	0.504
Gaze * Gesture	0.342	0.101	0.31	0.346	0.311	0.324	0.382
Gaze * Command	0.548	0.143	0.406	0.388	0.321	1.027	0.42
Gaze * Response	0.308	0.133	2.146	0.32	0.366	0.311	0.505
Gesture * Command	0.395	0.353	0.32	0.472	0.324	0.682	0.923
Gesture * Response	0.693	0.591	0.527	0.375	0.559	2.279	0.688
Command * Response	0.422	0.207	0.873	0.851	0.892	0.827	1.461
Gaze * Gesture * Command	0.433	0.556	0.574	0.523	3.727	0.511	0.412
Gaze * Gesture * Response	0.483	1.025	0.801	0.455	1.224	0.424	0.461
Gaze * Command * Response	0.424	0.414	3.233	0.854	0.44	0.533	0.396
Gesture * Command * Response	0.398	0.391	0.405	0.696	0.59	0.32	0.407
Gaze * Gesture * Command * Response	1.659	0.412	0.677	0.444	0.755	0.689	0.663

Table A.3 Experiment 3: Bayesian Inclusion Factors for RM-ANOVA Tests - All Data

	Like	Intel	H_Perm	R_Perm	H_Wrong	R_Wrong	App
Gaze	0.111	0.261	0.079	0.143	0.089	0.085	0.264
Gesture	8.28E+07	0.082	0.082	0.378	0.111	0.62	1.239
Command	3.499	0.792	0.546	0.42	0.673	0.497	4.022
Response	14.111	1.708	0.927	3.478	0.795	2.181	0.349
Gaze * Gesture	0.114	0.118	0.142	0.373	0.136	0.257	0.32
Gaze * Command	0.299	0.224	0.483	0.113	0.115	0.211	0.11
Gaze * Response	0.116	0.12	0.156	0.106	0.116	0.16	0.161
Gesture * Command	0.224	0.112	0.187	0.242	0.473	0.107	0.136
Gesture * Response	0.13	0.154	0.119	0.406	0.152	0.231	0.134
Command * Response	0.265	0.346	0.507	0.468	0.653	0.483	0.451
Gaze * Gesture * Command	0.542	0.798	0.206	0.192	0.259	0.174	0.164
Gaze * Gesture * Response	0.184	0.123	0.161	0.185	0.399	0.153	0.135
Gaze * Command * Response	0.409	0.191	0.149	0.179	0.124	0.205	0.171
Gesture * Command * Response	0.154	0.146	0.26	0.148	0.223	0.141	0.886
Gaze * Gesture * Command * Response	0.192	0.221	0.216	0.538	0.591	0.237	0.304

Table A.4 Experiment 3: Bayesian Inclusion Factors for RM-ANOVA Tests - Data Subset

	Like	Intel	H_Perm	R_Perm	H_Wrong	R_Wrong	App
Gaze	0.114	0.335	0.296	0.098	0.159	0.128	0.2
Gesture	1.13E+11	0.112	0.111	8.96	0.12	4.424	99.66
Command	7.916	0.883	0.574	0.397	0.7	0.497	2.28
Response	9.823	0.877	0.695	3.223	0.71	1.893	2.664
Gaze * Gesture	0.323	0.231	0.145	1.17	0.203	0.418	0.201
Gaze * Command	0.28	0.198	7.025	0.148	0.154	0.275	0.145
Gaze * Response	0.18	0.162	0.17	0.19	0.22	0.277	0.324
Gesture * Command	0.996	0.149	1.245	0.226	0.162	0.175	0.238
Gesture * Response	0.21	0.508	0.16	1.725	0.16	0.754	0.244
Command * Response	0.335	0.432	0.5	0.47	0.697	0.517	13.774
Gaze * Gesture * Command	0.473	0.488	0.235	0.231	1.884	0.225	0.673
Gaze * Gesture * Response	0.472	0.247	0.197	0.386	19.656	0.186	0.144
Gaze * Command * Response	0.431	0.207	0.274	0.353	0.221	0.242	0.211
Gesture * Command * Response	0.285	0.197	0.24	0.424	0.227	0.334	0.204
Gaze * Gesture * Command * Response	0.275	0.268	0.784	0.881	0.413	0.264	16.075